UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 10-29
Statistic and Econometric Series 15
June 2010

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34-91) 6249849

Simplicial similarity and its application to hierarchical clustering*

Ángel López and Juan Romo

Abstract

In the present document, an extension of the statistical depth notion is introduced with the aim to allow for measuring proximities between pairs of points. In particular, we will extend the simplicial depth function, which measures how central is a point by using random simplices (triangles in the two-dimensional space). The paper is structured as follows: In first place, there is a brief introduction to statistical depth functions. Next, the simplicial similarity function will be defined and its properties studied. Finally, we will present a few graphical examples in order to show its behavior with symmetric and asymmetric distributions, and apply the function to hierarchical clustering.

Keywords: Statistical depth; Similarity measures; Hierarchical clustering

Ángel López is PhD student, Department of Statistics, Universidad Carlos III de Madrid. Av. Universidad, 30, 28911 Leganés (Madrid) Spain (e-mail: alopez.stat@gmail.com). Juan Romo is Professor, Department of Statistics, Universidad Carlos III de Madrid. C/ Madrid, 126, 28903 Getafe (Madrid), Spain (e-mail: juan.romo@uc3m.es). **Acknowledgements:** The research of Ángel López and Juan Romo was supported by Comunidad de Madrid, project 06/HSE/0181/2004, Ministerio de Educación y Ciencia and Ministerio de Ciencia y Tecnología, projects ECO2008-05080, SEJ2005-06454 and BEC2002-03769.

Simplicial similarity and its application to hierarchical clustering

Ángel López and Juan Romo

Statistics Department, Universidad Carlos III de Madrid.

28911 - Leganés, Madrid, Spain.

Abstract

In the present document, an extension of the statistical depth notion is introduced with the aim to allow for measuring proximities between pairs of points. In particular, we will extend the simplicial depth function, which measures how central is a point by using random simplices (triangles in the two-dimensional space). The paper is structured as follows: In first place, there is a brief introduction to statistical depth functions. Next, the simplicial similarity function will be defined and its properties studied. Finally, we will present a few graphical examples in order to show its behavior with symmetric and asymmetric distributions, and apply the function to hierarchical clustering.

Keywords: Statistical depth, Similarity functions, Hierarchical clustering.

Ángel López is PhD student, Department of Statistics, Universidad Carlos III de Madrid. Av. Universidad, 30, 28911 Leganés (Madrid) Spain (e-mail: alopez.stat@gmail.com). Juan Romo is Professor, Department of Statistics, Universidad Carlos III de Madrid. C/ Madrid, 126, 28903 Getafe (Madrid), Spain (e-mail: romo@est-econ.uc3m.es).

1 INTRODUCTION

A recent field of research in Statistics is the ordering of points in spaces with more than one dimension. Multivariate statistics has been widely developed –both methodologically and theoretically, since the end of the 19th century, successfully adapting univariate methodology. Compared to univariate statistics, multivariate statistics has opened a wider application field due to the possibility of analyzing the relationship between variables, which allows us to use a wide range of models. However, there is a hard problem to solve, a problem which has been studied during the last decades: The ordering of data in high-dimensional spaces.

The possibility to order points in the multivariate space allows, for example, the extension of univariate robust estimators. The simplest example is the estimation of the center of a symmetric distribution. Apart from average and median, we have, among others, trimmed means, for which both estimators are limit cases. Robust estimators of the center are based on data ordering with the aim to control the weights of extreme observations. So, for multivariate robust estimation, obtaining a multivariate order (at least, partially) is essential.

In Barnett (1976) several multivariate orderings are defined: marginal order, reduced order, partial order and conditional order. In the marginal order, observations are ordered according to a univariate ordering of marginal distributions. In the reduced order, space dimension is reduced to a scalar (by using, for example, a generalized distance measure), and univariate ordering is performed for the reduced values. Partial ordering consists in splitting the sample in groups and giving the same value to all the members of the group. The convex hull peeling method belongs to this type of ordering. In this method, the first group is the convex hull of all observations (value 1, for example). The elements in this hull are removed. The elements of the second group are the points of the convex hull of the rest of observations (value 2, for example), and so on. In conditional ordering, the space is split into blocks until ordering all the blocks is possible.

Another concept of order is statistical depth, it consists in ordering with respect to a distribution function, and it is based on a center-outward ordering, that is, the deepest points of a data cloud are those closest to the center of the data cloud, whereas the less deep are the points far from the center. Statistical depth functions quantify this notion of depth. These functions assign real and non-negative values.

The paper is organized as follows: The second section contains a brief introduction to statistical depth functions. The most relevant functions and their most important properties are listed. In third section, the motivation of extending depth functions to the comparison of two or more points is introduced. Particularly, a definition is provided for simplicial depth similarity. In the fourth section, a set of desirable properties is defined and its fulfillment analyzed. Also, continuity and asymptotic properties will be examined. The fifth section includes the results of applying simplicial similarity to clustering. Finally, in the last section, the main conclusions and remarks are summarized.

2 Statistical depth

Statistical depth functions quantify how central is a point with respect to a distribution function or, in other words, how close is a point to the center of a distribution function. Given a point $x \in \mathbb{R}^d$ and a multivariate distribution function F, the population statistical depth function is denoted by $D(x; F)$. To obtain an estimation through a sample, F has to be replaced with a reasonable estimation F_n. In this case, the empirical (or sample) depth function is denoted by $D(x; F_n)$ or $D_n(x)$. Given a random sample of F, $x_1, x_2, \ldots, x_n$, and the depth values of the observations, $D(x_1; F), D(x_2; F), \ldots, D(x_n; F)$. If these depth values are ordered from higher to lower, observations will be ordered from the most internal to the most external.

A great deal of depth functions have been defined according to several geometric aspects, out-

lyingness measures or distances. Next, we will see some of the most relevant ones:

1. Mahalanobis depth: It is based on the Mahalanobis distance (Mahalanobis (1936)) between each point and the mean vector. It is defined as

$$MD(x;F) = \left[1 + (x-\mu_F)' \, \Sigma_F^{-1} \, (x-\mu_F)\right]^{-1},$$

where μ_F is the expected vector of F, and Σ_F its covariance matrix.

2. Halfspace depth: Introduced in Tukey (1975). For each point x, the depth value is the minimum probability of the halfspaces containing the point x. Given the d-dimensional distribution function F, the halfspace depth is defined as

$$HD(x;F) = \inf\left\{Pr(H) : H \text{ closed halfspace}, \, x \in H\right\}.$$

3. Oja depth: It was introduced in Oja (1983). It is based on convex hulls of $d+1$ points, that is, simplices whose volume is calculated. The information about point x is introduced because it is one of the vertices of all the simplices. Given the point x and $X_1, X_2, ..., X_d$ (d independent random variables with distribution F), the simplex $S[x, X_1, X_2, ..., X_d]$ is formed and its volume computed. If the simplex volume is high, this means that the point x is far from the other vertices, whereas if it is small, then the point x is close to the hyperplane defined by them. The Oja depth is defined as

$$OD(x;F) = \left[1 + E_F\left(Vol(S[x, X_1, X_2, ..., X_d])\right)\right]^{-1}.$$

4. Simplicial depth: Depth function defined in Liu (1990). The depth value of x is defined as the probability that the point belongs to a random simplex whose vertices are a $d+1$ independent random variables with distribution F. In other words, it verifies whether the point x is a convex linear combination of the vertices. So, it is expected that a point far from the center

of the distribution belongs to a small proportion of simplices. It is defined as

$$SD(x;F) = Pr\left\{x \in S\left[X_1, X_2, ..., X_{d+1}\right]\right\}.$$

The sample version of this depth function consists in computing the proportion of simplices containing the point x; i.e., given a random sample of F, $x_1, x_2, \ldots, x_n$, the depth value of x is estimated by

$$SD_n(x) = \binom{n}{d+1}^{-1} \sum_{1 \leq i_1 < \ldots < i_{d+1} \leq n} I(x \in S[x, x_{i_1}, x_{i_2}, ..., x_{i_{d+1}}]).$$

5. Depth based on L_1 median: Defined in Vardi and Zhang (2000). Given a definition of multivariate median θ and a d-dimensional distribution function F, in order to obtain the depth value for point x in $\mathbb{R}^d$, the minimum probability mass ω needed to convert x into the multivariate median of the mixture $(\omega\delta_x + F)/(\omega + 1)$ is calculated, where δ_x is the distribution function of a random variable degenerate at point x. Formally, it is defined as

$$L_1 D = 1 - \inf\left\{\omega \geq 0 : \theta\left(\frac{\omega\delta_x + F}{\omega + 1}\right) = x\right\}.$$

Computationally speaking, this depth function is less demanding than many of the above mentioned.

These are the most relevant depth functions. Others are, for example, likelihood depth (Fraiman and Meloche (1999)), projection depth, angular depth (Liu and Singh (1992)), zonoid depth (Koshevoy and Mosler (1997)), and band and modified band depths (López-Pintado and Romo (2009)).

Liu (1990) and Zuo and Serfling (2000) introduced several desirable properties for depth functions, and both affine invariance of simplicial depth and its fulfillment of those properties for distributions with angular symmetry are stated. The worst performance of this depth occurs for discrete distributions, since neither maximality nor monotonicity properties are ensured.

Liu (1990) also, proved that, for absolutely continuous distributions with bounded density, $SD_n(\cdot)$ is uniformly consistent:

$$\sup_{x\in\mathbb{R}^d} |SD_n(x) - SD(x;F)| \overset{c.s}{\underset{n\to\infty}{\to}} 0,$$

and, that the deepest point $\hat{\mu}$ converges almost sure to μ, if this is the only maximum of $SD(\cdot;F)$.

In Dümbgen (1992), limit theorems are studied for simplicial depth, and it is concluded that, under certain assumptions about the distribution function F, if succession F_n converges weakly to F, then $\|SD_n(\cdot) - SD(\cdot;F)\|_\infty \to 0$. Also the asymptotic normality of the L-statistics with appropriate weight functions is assessed.

3 Simplicial similarity

In this section, a function is defined to measure the proximity or similarity between points from a statistical point of view, i.e., in the same way that depth functions measure point centrality (or center proximity): taking into account the shape of the data set or the shape of the distribution function. Although for simplicial depth extending the depth function to compare simultaneously more than two points is straightforward, this work focuses in the comparison of pairs of points.

The simplicial depth function of a point $x \in \mathbb{R}^d$ is based on the membership of this point to random simplices with $d+1$ vertices, where d is the dimension of the space of x. This function measures the probability that a random simplex contains the point for which depth is calculated. The extension proposed in this document consists in demanding the membership to the random simplices of both points x and y to be compared. The two examples in Figure 1 show that the proposed idea is a right way to measure proximities. It can be observed that, when x and y (red circles) are close, two of the five random sample triangles drawn contain both points at the same time, whereas if the points are far, none of these five triangles contain both of them at the same

time.

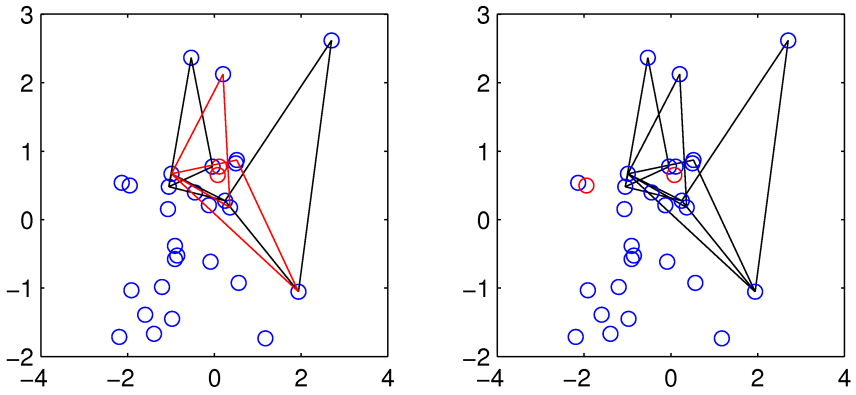


Figure 1: *Examples of random simplices for close and distant points.*

Definition 1 *Given a d-dimensional distribution function F and the points x and y in $\mathbb{R}^d$, the simplicial similarity between x and y with respect to F is defined as*

$$SS(x, y; F) = P\left(x, y \in S\left[X_1, X_2, \ldots, X_{d+1}\right]\right),$$

where $X_1, X_2, \ldots, X_{d+1}$ are independent random variables with distribution function F.

Remark 2 *Simplicial similarity can be rewritten as*

$$SS(x, y; F) = E_F\left[h\left(x, y; X_1, X_2, \ldots, X_{d+1}\right)\right],$$

where $h\left(x, y; X_1, X_2, \ldots, X_{d+1}\right)$ is the indicator function of membership of x and y to the simplex, $I\left(x, y \in S\left[X_1, X_2, \ldots, X_{d+1}\right]\right)$

A statistical concept used in some depth functions and in simplicial similarity is the U-statistic notion, which is introduced below.

Let $X_1, X_2, \ldots, X_n$ be a random sample in $\mathbb{R}^d$ with a distribution function F. Given a m-dimensional function $h(x_1, x_2, \ldots, x_m)$ with $m \leq n$ called kernel, the parameter $\theta(F) = E[h(X_1, X_2, \ldots, X_m)]$ is estimated by its appropriate U-statistic, which is obtained as

$$U_n = U(X_1, X_2, \ldots, X_n) = \binom{n}{m}^{-1} \sum_{1 \leq i_1 \leq \ldots \leq i_m \leq n} h(X_{i_1}, \ldots, X_{i_m}).$$

Given the random sample $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ of F, the sample counterpart of the simplicial similarity is a U-statistic with kernel function $h(X_{i_1}, \ldots, X_{i_m}) = I(x, y \in S[X_1, X_2, \ldots, X_{d+1}])$, i. e.,

$$SS_n(x, y) = \binom{n}{d+1}^{-1} \sum_{1 \leq i_1 < \ldots < i_{d+1} \leq n} I(x, y \in S[x_{i_1}, x_{i_2}, \ldots, x_{i_{d+1}}]).$$

This similarity can be extended to obtain the global proximity of sets with more than two points, demanding the membership of all of them to the simplices. However, this has a practical limitation: If the number of points to be compared represents a high percentage of the sample, probably there will be only a few simplices that contain all of them (especially if any of them is far from the rest), so the majority of comparisons will have small or even null similarity values.

3.1 Practical examples

To demonstrate the usefulness and the behavior of simplicial similarity, two data sets in the two-dimensional space with sample size 40 have been simulated. The first data set was drawn from the bivariate standard normal, and the second one from a bidimensional vector with exponentially distributed and independent coordinates.

Several figures are presented. Figure 2 is a three-dimensional surface showing the similarities between any point in the plane and a point previously fixed with respect to the Normal distribution. Figures 3 and 4 contain two graphs. Both represent the level curves of similarity between the points in the plane and two fixed points: a central and an external one. Regarding the behavior over a sample, the level curves present picks and triangular (in dimension two) shapes. So, these curves

are not convex. This is due to the angular shape of the simplices. Similarity has the ability to adapt to the shape of the distribution and to the depth value of the points compared. If the fixed point is central, for Normal sample (Figure 3) the contours are approximately symmetric and, for exponential sample, it can be observed that similarity outside the first quadrant is equal to zero and the contours are not symmetric.

One of the most important features of simplicial similarity (and hence, simplicial depth) is that outside of the convex hull is equal to zero. That is, similarity between two points outside the convex hull is equal to zero. The same happens for the similarity between one point inside the convex hull and another one outside it. This can be seen both as an advantage and a disadvantage. As a disadvantage, because the sample similarity will have an infinite number of points with zero similarity, even if the real distribution of the sample is continuous. On the other hand, it could be an advantage when the distribution used to compute similarity takes values in bounded regions, since similarity outside this region will be zero.

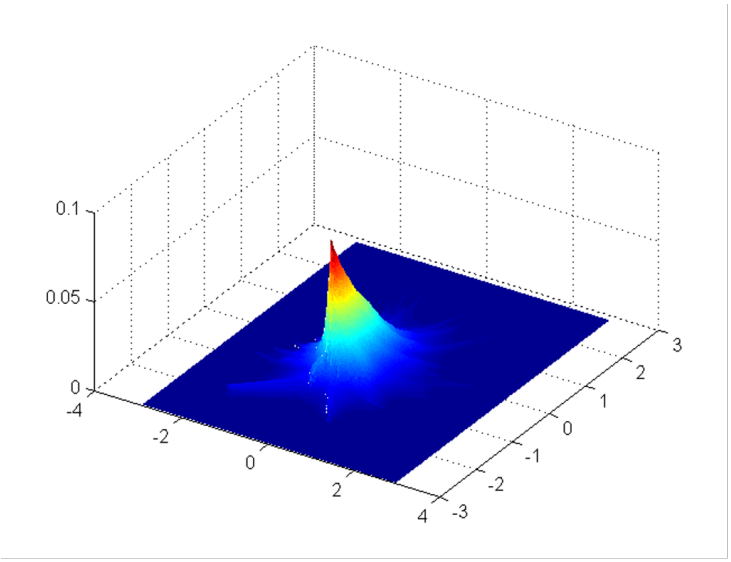
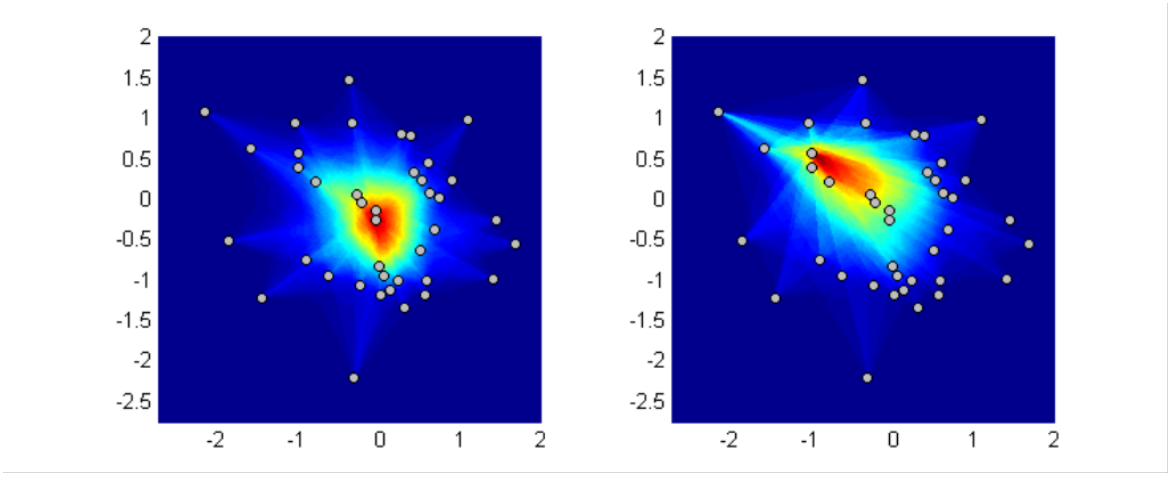
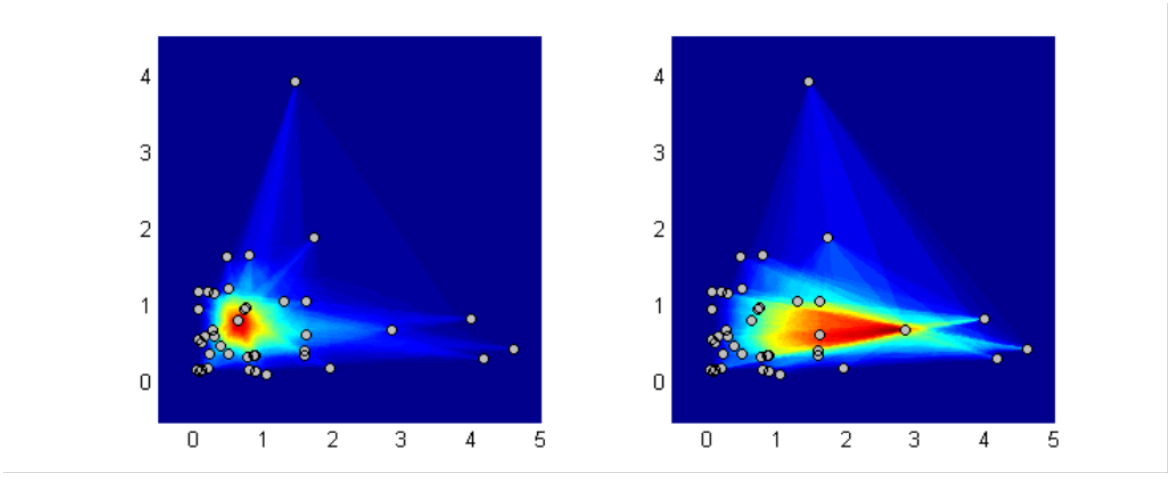


Figure 2: *Simplicial similarity.*



(a) Central fixed point. (b) External fixed point

Figure 3: *Simplicial similarity for a normal sample.*



(a) Central fixed point. (b) External fixed point

Figure 4: *Simplicial similarity for an exponential sample.*

4 PROPERTIES

4.1 Desirable properties

In Liu (1990) and Zuo and Serfling (2000) several desirable properties are proposed and analyzed for some depth functions. These properties are as follows: the maximum value of the function

has to be reached in the center of the distribution (if it has any); the function has to decrease monotonically over any segment which starts at the center of the distribution; the limit for points far from the center has to be zero, and the function has to be affine invariant. Below are listed these properties, adapted to a depth-based similarity. We include the symmetry property:

1. Given a point x, the maximum similarity value between this point and any other y is equal to the similarity between point x and itself.

2. Given any points x and y, the similarity between x and any point in the segment from x to y is greater than or equal to the similarity between points x and y.

3. Similarity between points x and y tends to zero when point y is far from x.

4. Given two points x and y, the distribution function F and an affine transformation $T(z)$, the similarity between points $T(x)$ and $T(y)$ with respect to the transformed distribution function $F_{T(z)}$ is equal to the similarity between points x and y with respect to F.

5. Similarity has to be symmetric.

When similarity fulfills these five properties it is called depth-based similarity.

Definition 3 *Let F_X be a d-dimensional distribution function. The bounded and non negative function $S(x, y; F_X)$ is called depth-based similarity if it verifies that:*

(i) $S(y, y; F_X) = \sup_{x \in \mathbb{R}^d} S(x, y; F_X)$*, for any* $y \in \mathbb{R}^d$

(ii) For any $x, y \in \mathbb{R}^d$ *and for all* $\alpha \in [0, 1]$*, then* $S(x, y; F_X) \leq S(y + \alpha(x - y), y; F_X)$

(iii) For any $y \in \mathbb{R}^d$*, then* $S(x, y; F_X) \to 0$ *when* $\|x\| \to \infty$

(iv) $S(x, y; F_X) = S(Ax + b, Ay + b; F_{AX+b})$ *for any pair of vectors* x *and* y *in* $\mathbb{R}^d$*, any nonsingular matrix* A *of size* $d \times d$ *and any vector* $b \in \mathbb{R}^d$

(v) For any pair of points x and y in $\mathbb{R}^d$, then $S(x, y; F_X) = S(y, x; F_X)$

Given a point x in $\mathbb{R}^d$, if similarity is computed over a set of points and values are ordered from higher to lower, points become ordered from least to greatest distance from x.

Proposition 4 *Simplicial similarity is a similarity based on depth in the sense of Definition 3 for absolutely continuous functions.*

Proof. The proof involves both properties of the simplicial depth function and of the sets defined above. Given two points x and y in $\mathbb{R}^d$, the events A, B, C and A_α are defined as sets of random simplices that verify certain conditions of membership for points x and y, as well as convex linear combinations of them. They are defined as

$$\begin{aligned} A &= \{X_1, X_2, ..., X_{d+1} : x, y \in S[X_1, X_2, ..., X_{d+1}]\}, \\ B &= \{X_1, X_2, ..., X_{d+1} : x \in S[X_1, X_2, ..., X_{d+1}]\}, \\ C &= \{X_1, X_2, ..., X_{d+1} : y \in S[X_1, X_2, ..., X_{d+1}]\} \text{ and} \\ A_\alpha &= \{X_1, X_2, ..., X_{d+1} : \alpha x + (1-\alpha) y, y \in S[X_1, X_2, ..., X_{d+1}]\}, \alpha \geq 0. \end{aligned}$$

These events verify that $A \subseteq B$ and $A \subseteq C$, and, due to the convexity of the simplices, if $\alpha_1 \geq \alpha_2$, then $A_{\alpha_1} \subseteq A_{\alpha_2}$. Proof of each property follows:

(i) Given a point y, the maximum value of $SS(x, y; F)$ is reached for $x = y$, since $SS(x, y; F) = Pr(A) \leq Pr(C) = SS(y, y; F)$

(ii) Monotonic decrease is ensured also due to the convexity of the set $S[X_1, X_2, ..., X_{d+1}]$, because all the simplices containing both points x and y contain also all the convex linear combinations of these points, $\alpha x + (1-\alpha) y$, and hence, the sets A and A_α previously defined verify that $A \subseteq A_\alpha$. Then $SS(x, y; F) = Pr(A) \leq Pr(A_\alpha) = SS(\alpha x + (1-\alpha) y, y; F)$.

(iii) The vanishing at infinite is true because it is true for simplicial depth. We have that

$$SS(x, y; F) = Pr(A) \leq Pr(B) = SS(x, x; F) = SD(x; F).$$

So, as simplicial similarity between two points is bounded by the simplicial depth of them, it is possible to use the result of vanishing at infinite of simplicial depth (see Theorem 1 in Liu (1990)):

$$0 \leq \lim_{\|x\| \to \infty} SS(x, y; F) \leq \lim_{\|x\| \to \infty} PS(x; F) = 0.$$

(iv) The affine invariance property is true because of the convexity of the simplices, that is, given the nonsingular matrix A and vector b, then $x \in S[x_1, x_2, ..., x_{d+1}]$ is equivalent to $Ax + b \in S[Ax_1 + b, Ax_2 + b, ..., Ax_{d+1} + b]$. So, the probability that this happens the same for all the simplices.

(v) The similarity is symmetric by definition. ■

4.2 Continuity and asymptotic properties

Next, the continuity property of the simplicial similarity will be proven, and some asymptotic results stated. In order to get these results, the next Lemma (Lemma 3 in Liu (1990)) will be used.

Lemma 5 (Lemma 3 in Liu (1990)) *Let F be a distribution function in $\mathbb{R}^d$, and let $x_1, x_2, \ldots, x_n$ be a random sample of F. Given the U-statistic $U_n = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \ldots < i_m \leq n} h(x_{i_1}, x_{i_2}, \ldots, x_{i_m})$ with kernel $h(\cdot)$ of degree m. If h is bounded, say by c, then for any $r \geq 2$,*

$$E[(U_n - E(U_n))^r] \leq \frac{K}{n^{r/2}},$$

where K is a constant that depends on c.

We will begin with the result of continuity of the simplicial similarity.

Theorem 6 *Given $y \in \mathbb{R}^d$, if F is an absolutely continuous distribution function, then $SS(\cdot, y; F)$ is continuous.*

Proof. Function $SS(\cdot, y; F)$ is continuous if, given a succession $x_n \in \mathbb{R}^d$ converging to x, then the $\lim_{n\to\infty} |SS(x, y; F) - SS(x_n, y, F)| = 0$. To prove it, the following events are used:

$$A_x = \{X_1, X_2, ..., X_{d+1} : x \in S[X_1, X_2, ..., X_{d+1}]\},$$

$$A_{x_n} = \{X_1, X_2, ..., X_{d+1} : x_n \in S[X_1, X_2, ..., X_{d+1}]\}, \text{ and}$$

$$A_y = \{X_1, X_2, ..., X_{d+1} : y \in S[X_1, X_2, ..., X_{d+1}]\}.$$

Simplicial similarity $SS(x, y; F)$ is equal to $Pr(A_x \cap A_y)$; therefore, the difference $SS(x, y; F) - SS(x_n, y, F)$ is equal to $Pr(A_x \cap A_y) - Pr(A_{x_n} \cap A_y)$. This quantity may be bounded, since

$$A_x \cap A_y \subseteq (A_{x_n} \cap A_y) \cup \left(A_x \cap \overline{A_{x_n}} \cap A_y\right),$$

then

$$Pr(A_x \cap A_y) \leq Pr(A_{x_n} \cap A_y) + Pr\left(A_x \cap \overline{A_{x_n}} \cap A_y\right),$$

and then

$$Pr(A_x \cap A_y) - Pr(A_{x_n} \cap A_y) \leq Pr\left(A_x \cap \overline{A_{x_n}} \cap A_y\right).$$

Applying the same reasoning to $A_{x_n} \cap A_y$, it is obtained that

$$Pr(A_{x_n} \cap A_y) - Pr(A_x \cap A_y) \leq Pr\left(\overline{A_x} \cap A_{x_n} \cap A_y\right).$$

And, therefore,

$$\begin{aligned} |Pr(A_x \cap A_y) - Pr(A_{x_n} \cap A_y)| &\leq Pr\left(A_x \cap \overline{A_{x_n}} \cap A_y\right) + Pr\left(\overline{A_x} \cap A_{x_n} \cap A_y\right) \\ &\leq Pr\left(A_x \cap \overline{A_{x_n}}\right) + Pr\left(\overline{A_x} \cap A_{x_n}\right) \\ &\leq (d+1)\, Pr(B_n), \end{aligned}$$

where $B_n = \{X_1, X_2, ..., X_d : H(X_1, X_2, ..., X_d)$ intersects the segment from x_n to $x\}$, and $H(X_1, X_2, \ldots, X_d)$ is the hyperplane containing points $X_1, X_2, ..., X_d$, since the event $\left(A_x \cap \overline{A_{x_n}}\right) \cup \left(\overline{A_x} \cap A_{x_n}\right)$ is included in the event defined as the set of simplices with a side intersecting the segment $\overline{xx_n}$. Notice that the $\limsup_{n\to\infty} B_n = \{X_1, X_2, ..., X_d : x \in H(X_1, X_2, ..., X_d)\}$ is the hyperplanes beam containing x. If F is continuous, this is a measure null set. And therefore, due to the continuity of F, then

$$\begin{aligned}
\lim_{n\to\infty} |SS(x, y; F) - SS(x_n, y, F)| &= \limsup_{n\to\infty} |Pr(A_x \cap A_y) - Pr(A_{x_n} \cap A_y)| \\
&\leq (d+1)\limsup_{n\to\infty} Pr(B_n) \\
&\leq (d+1) Pr\left(\limsup_{n\to\infty} B_n\right) \\
&= 0. \blacksquare
\end{aligned}$$

Theorem 7 *Let $D \subseteq \mathbb{R}^d$ be an open set, and $F : D \longrightarrow \mathbb{R}^+$ an absolutely continuous distribution function. Then, simplicial similarity satisfies that*

$$SS(x, y; F) = SS(y, y; F) \text{ if, and only if, } x = y.$$

Proof. If $x = y$, the statement is true by definition. To prove it, checking the inverse implication is enough, that is, $x \neq y$ implies the inequality $SS(x, y; F) \neq SS(y, y; F)$ (or, more exactly, $SS(x, y; F) < SS(y, y; F)$).

Let be the sets A_{xy} and A_y defined in the same way that in the previous proof, then $Pr(A_{xy}) \leq Pr(A_y)$ $\forall x, y \in D$; therefore, in order to finish the proof, it must be proven that the probability of the difference of those sets, $A_y \backslash A_{xy}$, is greater than zero. To prove this statement, for simplicity and without loss of generality, d is taken equal to 2. The aim of the proof is, through the use of graphics, finding a set that belongs to the difference $A_y \backslash A_{xy}$ such that it has no null probability. In the two-dimensional space, simplices are triangles. Triangles useful for the proof are those containing the point y but not x.

The three vertices are selected as follows: the first one is any point inside the set D minus the halfline with origin at x and direction $x - y$, that is, any point in $D \backslash SL_{x,y}$, where $SL_{x,y} = \{x + \alpha (x - y) : \alpha \geq 0\}$. However, since F is continuous, the set $SL_{x,y}$ has zero probability; therefore, the region for which the first vertex can be selected has a probability equal to 1. Given the point $X_1 \in D \backslash SL_{x,y}$, the set S_{X_1} is defined as $\left\{ z \in D : \overrightarrow{X_1 z} \text{ intersects } \overrightarrow{xy} \right\}$, where $\overrightarrow{ab}$ is the segment between a and b. As can be seen in Figure 5, the region $S_{X_1} \neq \phi$, so its probability is greater than zero. More formally, this is true because F is continuous and strictly positive, and D is an open set, with which $\exists \varepsilon > 0 \backslash Ball (y, \varepsilon) \subset D$.

Given the two first vertices of the triangle, $X_1 \in D \backslash SL_{x,y}$ and $X_2 \in S_{X_1}$, the region for the last

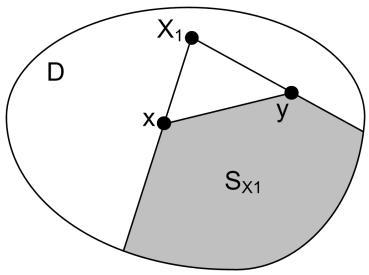


Figure 5: *Region for the second vertex of the triangle.*

vertex (in order to construct a triangle excluding the point x, but containing the point y) can be defined as $S_{X_1,X_2} = \{z : y \in S [X_1, X_2, z]\}$. This region is represented in Figure 6 where a triangle is shown containing y, but not x. Again, since D is open and F is continuous and strictly positive, it is concluded that $Pr (S_{X_1,X_2}) > 0$ and, finally, taking expectations over this quantity we have that $Pr (A_y \backslash A_{xy}) > 0$. ■

Finally, the asymptotic results of the sample counterpart of the similarity are analyzed.

Theorem 8 *Sample simplicial similarity is unbiased and consistent,*

$$SS_n (x, y) \xrightarrow[n \to \infty]{p} SS (x, y; F), \forall x, y \in \mathbb{R}^d.$$

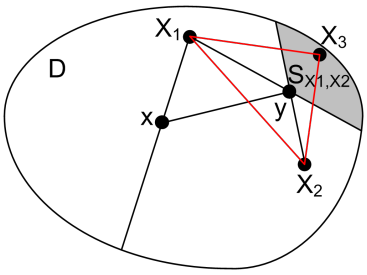


Figure 6: *Region S_{X_1,X_2} and triangle belonging to $A_y \backslash A_{xy}$.*

Proof. Given any pair of points x and $y \in \mathbb{R}^d$, since $SS_n(x,y)$ is a U-statistic, then

$$\begin{aligned} E[SS_n(x,y)] &= E\left[\binom{n}{d+1}^{-1} \sum I\left(x,y \in S\left[x_{i_1}, x_{i_2}, ..., x_{i_{d+1}}\right]\right)\right] = \\ &= \binom{n}{d+1}^{-1} \sum E\left[I\left(x,y \in S\left[x_{i_1}, x_{i_2}, ..., x_{i_{d+1}}\right]\right)\right] = \\ &= \binom{n}{d+1}^{-1} \binom{n}{d+1} Pr\left[I\left(x,y \in S\left[X_1, X_2, ..., X_{d+1}\right]\right)\right] = \\ &= Pr\left[I\left(x,y \in S\left[X_1, X_2, ..., X_{d+1}\right]\right)\right] = SS(x,y;F). \end{aligned}$$

Sample similarity is consistent if its variance converges to zero when the sample size increases. This is true because sample similarity is a U-statistic whose kernel is bounded by 1, and therefore, making use of Lemma 5 for $r = 2$ and a constant K (only dependent on the bound of the kernel), we have that,

$$E\left[\left(SS_n(x,y) - E\left(SS_n(x,y)\right)\right)^2\right] \leq \frac{K}{n} \overset{n\to\infty}{\longrightarrow} 0. \quad \blacksquare$$

Theorem 9 *Sample simplicial similarity is strongly consistent,*

$$SS_n(x,y) \xrightarrow[n\to\infty]{a.s.} SS(x,y;F), \forall x,y \in \mathbb{R}^d.$$

Proof. The proof is similar to that of Theorem 8, because Lemma 5 is also used, showing that, for $r \geq 2$, sample similarity converges in r-th mean. It is known that if convergence is sufficiently fast, that is, if

$$\sum_{n=1}^{\infty} E\left[|SS_n(x,y) - SS(x,y;F)|^r\right] < \infty,$$

then it converges almost surely to $SS(x,y;F)$. So, for $r=4$, U-statistic satisfies

$$\sum_{n=1}^{\infty} E\left[|SS_n(x,y) - SS(x,y;F)|^4\right] \leq \sum_{n=1}^{\infty} \frac{K}{n^2} < \infty,$$

and then it converges almost surely. ■

Lemma 10 *For any distribution function F in $\mathbb{R}^d$ and any point $y \in \mathbb{R}^d$ considered as fixed, then*

$$\sup_{\|x\|\geq M} SS_n(x,y) \xrightarrow{a.s.} 0, \text{ when } M \to \infty.$$

Proof. Simplicial similarity satisfies this convergence because $SS(x,y;F) \leq SD(x;F)$ and simplicial depth verifies the convergence:

$$\sup_{\|x\|\geq M} SS_n(x,y) \leq \sup_{\|x\|\geq M} SD_n(x,y) \xrightarrow{a.s.} 0, \text{ when } M \to \infty. \blacksquare$$

Lemma 11 *Let F be an absolutely continuous distribution function; given any point $y \in \mathbb{R}^d$ considered as fixed, then for all $c > 0$*

$$\sup_{\{x_1,x_2 \in Ball(y,c):\|x_1-x_2\|<\varepsilon\}} |SS_n(x_1,y) - SS_n(x_2,y)| \underset{\varepsilon\to 0, n\to\infty}{\longrightarrow} \gamma(\varepsilon) + R_n,$$

where $Ball(y,c) = \left\{x \in \mathbb{R}^d : \|x-y\| \leq c\right\}$, $\gamma(\varepsilon)$ is not random, its limit is zero, and R_n converges almost surely to zero.

Proof. First, population terms are introduced and triangular inequality is used for decomposing the absolute value into three addends:

$$\begin{aligned} & |SS_n(x_1,y) - SS_n(x_2,y)| \\ = \; & |SS_n(x_1,y) - SS(x_1,y;F) + SS(x_1,y;F) - SS_n(x_2,y) \\ & + SS(x_2,y;F) - SS(x_2,y;F)| \\ \leq \; & |SS_n(x_1,y) - SS(x_1,y;F)| + |SS_n(x_2,y) - SS(x_2,y;F)| \\ & + |SS(x_1,y;F) - SS(x_2,y;F)| \end{aligned}$$

and, since the supremum of this quantity is less than or equal to the sum of the supremums, the original supremum will be bounded by the sum of the other three supremums. These supremums are analyzed separately.

$$\begin{aligned} & \sup_{\{x_1,x_2 \in Ball(y,c):\|x_1-x_2\|<\varepsilon\}} |SS_n(x_1,y) - SS_n(x_2,y)| \\ \leq \; & \sup_{\{x_1,x_2 \in Ball(y,c):\|x_1-x_2\|<\varepsilon\}} |SS_n(x_1,y) - SS(x_1,y;F)| \\ & + \sup_{\{x_1,x_2 \in Ball(y,c):\|x_1-x_2\|<\varepsilon\}} |SS_n(x_2,y) - SS(x_2,y;F)| \\ & + \sup_{\{x_1,x_2 \in Ball(y,c):\|x_1-x_2\|<\varepsilon\}} |SS(x_1,y;F) - SS(x_2,y;F)| . \end{aligned}$$

On one hand, it is obtained that $\gamma(\epsilon) = \sup_{\{x_1,x_2 \in Ball(y,c):\|x_1-x_2\|<\varepsilon\}} |SS(x_1,y;F) - SS(x_2,y;F)|$ tends to zero when ε tends to zero, because similarity is continuous and the supremum is taken over the set $Ball(y,c)$, which is bounded and closed. On the other hand, it is obtained that

$$\sup_{\{x_1,x_2 \in Ball(y,c):\|x_1-x_2\|<\varepsilon\}} |SS_n(x_2,y) - SS(x_2,y;F)|$$

is equal to

$$\sup_{A_{x_1,x_2}(y,c,\varepsilon)} |Pr_{F_n}(A_{x_1,x_2}(y,c,\varepsilon)) - Pr_F(A_{x_1,x_2}(y,c,\varepsilon))| ,$$

where $A_{x_1,x_2}(y,c,\varepsilon)$ is the set of all the simplices containing points x_1 and y, for x_1 such that $\{x_1 \in Ball(y,c) : \|x_1 - x_2\| < \varepsilon\}$. And, as the simplices in set $A_{x_1,x_2}(y,c,\varepsilon)$ belong to the set of

simplices containing x_1 and y ($A_{x_1,y}$), then

$$\sup_{A_{x_1,x_2}(y,c,\varepsilon)} \left| Pr_{F_n}\left(A_{x_1,x_2}\left(y,c,\varepsilon\right)\right) - Pr_F\left(A_{x_1,x_2}\left(y,c,\varepsilon\right)\right)\right|$$

is equal to

$$\sup_{A_{x_1,x_2}} \left| Pr_{F_n}\left(A_{x_1,x_2}\right) - Pr_F\left(A_{x_1,x_2}\right)\right|$$

and, since the class of all convex Borel-measurable sets in $\mathbb{R}^d$ forms a Glivenko-Cantelli class if F is a density with respect to the Lebesgue measure, that is,

$$\sup_{A\in C} \left|F_n\left(A\right) - F\left(A\right)\right| \xrightarrow{c.s.} 0,$$

where C is the set of all convex Borel-measurable sets, then the supremum converges almost surely to zero. The same reasoning can be used for point x_2. ■

Theorem 12 *Let F be an absolutely continuous distribution function. Then simplicial similarity $SS\left(x,y;F\right)$ is uniformly consistent:*

$$\sup_{x\in\mathbb{R}^d} \left|SS_n\left(x,y\right) - SS\left(x,y;F\right)\right| \xrightarrow[n\to\infty]{a.s.} 0.$$

Proof. The proof is analogous to that of Theorem 5 in Liu (1990). Given a point x such that $\|x\|$ is sufficiently large, the property (iv) in Proposition 4 and Lemma 10 ensure that $\left|SS_n\left(x,y\right) - SS\left(x,y;F\right)\right| \xrightarrow[n\to\infty]{a.s.} 0$. Then, the statement has to be proven for *small* points, that is, given $M > 0$, it will be proven that

$$\sup_{x\in Q(y,M)} \left|SS_n\left(x,y\right) - SS\left(x,y;F\right)\right| \xrightarrow[n\to\infty]{a.s.} 0,$$

where $Q\left(y,M\right)$ is a hypercube with center y and side size equal to M.

Splitting each side of the hypercube into N pieces, the hypercube is split into N^d subhypercubes. Making N sufficiently large and applying Lemma 11 and Theorem 6, the conclusion is reached that it is enough to prove

$$\max_{x\in C(y,M)} \left|SS_n\left(x,y\right) - SS\left(x,y;F\right)\right| \xrightarrow[n\to\infty]{a.s.} 0,$$

where $C(y, M)$ is the set of all the corners of the hypercubes. Applying Lemma 5 with $r = 4$ and $c = 1$, it is obtained that

$$
\begin{aligned}
& Pr\left(\max_{x \in C(y,M)} |SS_n(x,y) - SS(x,y;F)| > \varepsilon\right) \\
& \leq N^d \max_{x \in C(y,M)} Pr\left(|SS_n(x,y) - SS(x,y;F)| > \varepsilon\right) \\
& = N^d \max_{x \in C(y,M)} Pr\left(|SS_n(x,y) - SS(x,y;F)|^4 > \varepsilon^4\right) \\
& \leq N^d \max_{x \in C(y,M)} E\left[\varepsilon^{-4}\left(|SS_n(x,y) - SS(x,y;F)|^4\right)\right] = O\left(n^{-2}\right),
\end{aligned}
$$

and, due to the Borel-Cantelli Lemma, then the sum over n of

$$
Pr\left(\max_{x \in C(y,M)} |SS_n(x,y) - SS(x,y;F)| > \varepsilon\right)
$$

is finite, with which the probability of the upper limit of this event is equal to zero, thus leading to conclude that simplicial similarity is strongly consistent. ■

5 APPLICATION TO CLUSTERING

Clustering analysis is categorized within non-supervised classification methods, whose aim is to classify and group the observations of a sample when membership to groups is unknown. This is an exploratory technique which tries to describe how observations are distributed in the space. Some clustering methods need some prior information about the number of groups into which the sample is divided before grouping. The most used methods are divided into two categories: hierarchical and partitioning methods.

Partitioning methods make a partition into K groups, where K is usually defined by the user, although the choice can be automatic. A clustering is considered as a partition when each group contains at least one observation and each one of them belongs to a single group. Under these

conditions, at most there would be as many groups as observations. The purpose is to obtain a partition in which the elements of both groups are heterogeneous between them, and the elements of the same groups are homogeneous. Some partition methods are, for instance, k-means (Hartigan (1975)), PAM (Partitioning Around Medoids, Kaufman and Rousseeuw (1987)), CLARA (Clustering Large Applications, Kaufman and Rousseeuw (1986)), FANNY (Fuzzy Analysis, Kaufman and Rousseeuw (1990)) and, based on statistical depth, that proposed in Jornsten (2004).

The other type of clustering techniques includes hierarchical methods. These, unlike the partition ones, do not provide for a single grouping of the observations because they are iterative methods, and in each step a grouping is obtained. There are two ways to obtain hierarchical clusterings: the agglomerative and the divisive. The first type starts with as many groups as observations, and in each step groups are joined until a single group including all observations is obtained. The divisive ones work in the opposite direction: they start with one group composed by all the observations of the sample and, iteratively, it splits groups into subgroups until there are as many groups as observations. All the agglomerative or divisive steps can be represented by trees called dendrograms. Such trees make possible to obtain a clustering for a specific number of groups.

There are several criteria for determining the distances between groups. The most important are:

1. Single linkage or nearest neighbor: The distance between two groups is taken as the minimum distance between the elements of both groups.

2. Complete linkage or maximum distance: The distance between two groups is taken as the maximum distance between the elements of both groups.

3. Average linkage: The distance between two groups is taken as the average distance between

the elements of both groups.

4. Ward method: It starts with as many groups as observations. Each step consists in clustering the two groups that make the minimum increase of the sum over all the groups of the squared distances between the elements and the centroid of their groups.

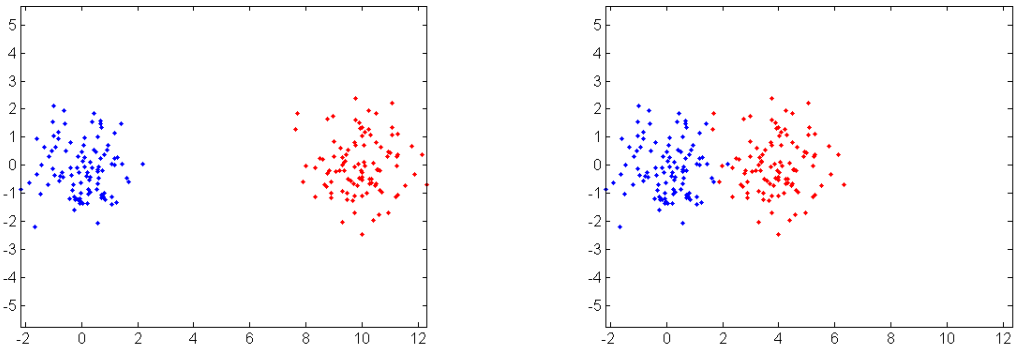
In this section, simplicial similarity will be applied to hierarchical clustering. Several two-dimensional data sets are analyzed. Such data sets will be simulated from symmetric and asymmetric distributions. The results will be compared to those obtained for the Euclidean distance.

5.1 Calculating the similarity matrix

Simplicial similarity is appropriate for measuring proximities between pairs of points according to the shape of the data set or the distribution function of the points. Its performance is good for the proximities in a descriptive sense. However, when similarity is applied to search for groups or clusters, the results can be inadequate. This is due to the fact that there are no observations in the gaps separating the clusters, and then it is possible that similarity does not take into account these gaps, with which similarity values between members of the different groups can be small, even though the distance between groups is large.

Figure 7 contains the scatter plot of two samples drawn from a mixture of normals. In the first one, the sample of a mixture with separate groups is represented. The second represents the sample of a mixture with close groups. Table 1 presents, on one hand, the average similarity between members of different groups, and, on the other, the average similarity between members of the same group. The ratios of the average between groups and within groups for the sample with separated groups and closed groups are, respectively, 5.08 and 2.65. This ratio is almost two-fold when groups are separated.

With the purpose to obtain a more effective measure that differentiates better between groups,



(a) Separated groups

(b) Close groups

Figure 7: *Scatter plot for samples of mixtures of normals.*

	Between groups	Within groups
Separated groups	0.0417	0.2117
Close groups	0.0723	0.1913

Table 1: *Average similarity with respect to the empirical distribution.*

it is proposed to complete the space with a continuous distribution function. Such completion would introduce probability mass in gaps and allow for better distinguishing between groups. In practice, this completion of the space consists in computing similarities not with respect to the empirical distribution, but to another distribution. However, given that one of the advantages of the depth-based similarity with respect to the empirical distribution is that it adapts itself to the shape of the data set, completely changing the reference distribution seems inappropriate. So, it is proposed using a mixture of the empirical distribution and the completion function.

This completion function should not distort the empirical distribution at a great extent, because in such case undesired results could be obtained. So, a desirable condition is that the parameters of the completion function be estimated with the sample. This would get a consistent completion. For

instance, it could be possible taking the multivariate Normal distribution with the sample mean and covariance matrix, or a Student's t if more heavy tails are sought. Another choice would be filling out a hypercube (or ellipse) with a uniform distribution containing all the observations.

In this section, the multivariate normal is used as the completion distribution, and is weighted equal to the empirical distribution, i.e., the simplicial similarities are computed with respect to the mixture $F = \alpha \cdot Normal\left(\hat{\mu}, \hat{\Sigma}\right) + (1-\alpha) \cdot F_n$, with $\alpha = 0.5$, and where $\hat{\mu}$ and $\hat{\Sigma}$ are, respectively, the vector of means and the sample covariance matrix.

Table 2 contains the average similarities of Table 1 taking this mixture as the computation distribution. Now the ratio for the sample with separated groups is equal to 5.97, and for the sample with close groups equal to 2.98. By completing the space, ratios have been increased by 17.5% (separated groups) and 12.5% (close groups). This means that using the mixture allows for better distinction between groups because similarity between elements of the same group is higher, and similarity between elements of different groups is lower.

	Between groups	Within groups
Separated groups	0.0391	0.2335
Close groups	0.0654	0.1952

Table 2: *Average similarities with respect to the mixture.*

Finally, another advantage of completion is that, for a continuous completion distribution, similarities for points outside of the convex hull of the sample are not equal to zero, avoiding the drawbacks of the simplicial similarity when it is applied to classification problems.

The main disadvantage is that completion makes impossible to actually calculate similarity, and, therefore, it can only be estimated. This estimators can only be done in a reasonable time in the two-dimensional case.

Given a sample of n observations $x_1, x_2, \ldots, x_n$ in $\mathbb{R}^d$, the similarity matrix is obtained

$$S = \begin{pmatrix} S_{1,1} & S_{1,2} & \ldots & S_{1,n} \\ S_{1,2} & S_{2,2} & \ldots & S_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n,1} & S_{n,2} & \ldots & S_{n,n} \end{pmatrix},$$

where $S_{i,j}$ represents the simplicial similarity between points x_i and x_j with respect to the mixture $\alpha \cdot Normal\left(\hat{\mu}, \hat{\Sigma}\right) + (1-\alpha) \cdot F_n$, where $\hat{\mu}$ and $\hat{\Sigma}$ are the mean vector and the sample covariance matrix, respectively, and $\alpha \in [0, 1]$.

Once the matrix has been computed, it should be transformed into a dissimilarity matrix, because hierarchical clustering works with this kind of matrices.

Given that, in simplicial similarity, the range of values is bounded by the depth value of the two points compared, a scaling is carried out for the similarity matrix. With this one, the similarity between any two points can take values in the same interval ($[0, 1]$), which implies that the scaled similarities for any pair of points are comparable.

Let D be a diagonal matrix with diagonal values equal to those in the diagonal of S

$$D = \begin{pmatrix} S_{1,1} & 0 & \ldots & 0 \\ 0 & S_{2,2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & S_{n,n} \end{pmatrix},$$

the scaled matrix is obtained as $S^{esc} = D^{-1/2} S D^{-1/2}$.

Finally, given the scaled similarity matrix with ones in the main diagonal, S^{esc}, logarithm is

applied in order to obtain dissimilarities,

$$\delta = \begin{pmatrix} -log(S^{esc}_{1,1}) & -log(S^{esc}_{1,2}) & \dots & -log(S^{esc}_{1,n}) \\ -log(S^{esc}_{1,2}) & -log(S^{esc}_{2,2}) & \dots & -log(S^{esc}_{2,n}) \\ \vdots & \vdots & \ddots & \vdots \\ -log(S^{esc}_{n,1}) & -log(S^{esc}_{n,2}) & \dots & -log(S^{esc}_{n,n}) \end{pmatrix}.$$

5.2 Examples of application

This section presents the results of applying simplicial similarity to hierarchical clustering. Fourteen samples have been simulated from different models. Due to the problem with simplicial similarity in spaces with dimension greater than two, all the models have been drawn from bivariate distributions. The results for similarity have been compared to those obtained with the Euclidean distance. In all of the examples, the distribution for similarity computation is the 50% mixture with the bivariate normal.

Regarding the criterion of the distances between groups in the clustering algorithm, our experience suggests using the Ward method for simplicial similarity. As far as Euclidean distance is concerned, the single linkage and the Ward method were used.

There are 14 data sets made up of two, three or four groups. These have been simulated from symmetric and asymmetric distributions. In the last examples, groups show a non-linear relationship. Due to the sample size of the examples, dendrograms are not showed, but they have been used to obtain the clustering for the number of groups in each example. For each example, four scatter plots are shown: the real grouping, the grouping with simplicial similarity and both groupings with Euclidean distance. First, the results for the data set with symmetric groups are presented.

5.2.1 Groups with symmetric distributions

The first group of data sets is made up of six samples with symmetric groups. Examples 1 to 4 have groups from the normal distribution. All of them have two groups, except one, which has three. Examples 5 and 6 have groups simulated from the uniform distribution in rectangles. The examples are as follows:

Example 1: Two groups of sample size $n_1 = n_2 = 100$, drawn from distributions $Normal\,(\mathbf{0}, \mathbf{I})$ and $Normal\,(\mu, \mathbf{I})$, where $\mathbf{0}$ is the zero vector of dimension two, $\mathbf{I}$ is the identity matrix of size 2×2, and $\mu = (5, 5)'$.

Example 2: Two groups with opposite correlations, of sizes $n_1 = n_2 = 100$, simulated from distributions $Normal\,(\mathbf{0}, \Sigma_1)$, and $Normal\,(\mu,\ \Sigma_2)$, where $\mathbf{0}$ is the zero vector, $\mu = (3.5,\ 3.5)'$,

$$\Sigma_1 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \text{ and } \Sigma_2 = \begin{pmatrix} 1 & \text{-}0.8 \\ \text{-}0.8 & 1 \end{pmatrix}.$$

Example 3: Two groups of variables with different variances of sizes $n_1 = n_2 = 100$, simulated from distributions $Normal\,(\mathbf{0}, \Sigma)$ and $Normal\,(\mu, \Sigma)$, where $\mathbf{0}$ is the zero vector, $\mu = (5, 0)'$, and

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix}.$$

Example 4: Three groups with variables of different variances and sizes $n_1 = n_2 = n_3 = 100$. Two groups are simulated from the distributions of Example 3. However, the third group has been drawn from $Normal\,(\mu, \mathbf{I})$, where $\mu = (10, 20)'$ and $\mathbf{I}$ is the identity matrix of size 2×2.

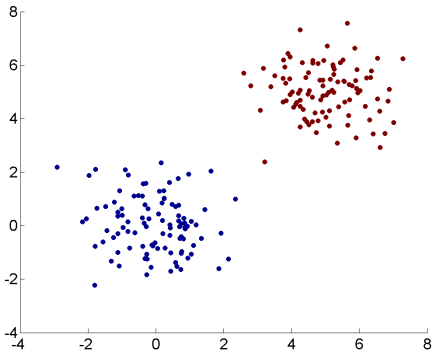
Example 5: Two rectangular groups with uniformly distributed coordinates of sizes $n_1 = n_2 = 100$. The first group has the first coordinate distributed as a U (-0.5, 0.5), whereas the second one is distributed as a U (0.5, 10.5). The second group has the first and the second coordinates distributed as U (0.5, 10.5) and U (-0.5, 0.5), respectively.

Example 6: Four rectangular groups with uniformly distributed coordinates and sample sizes $n_1 = n_2 = n_3 = n_4 = 100$. Two groups are simulated from the models of Example 5. The third group has the first coordinate distributed as a U (-0.5, 0.5), and the second one as a U (-10.5, -0.5). The fourth group has the first coordinate with distribution U (-10.5, -0.5), and the second one with distribution U (-0.5, 0.5).

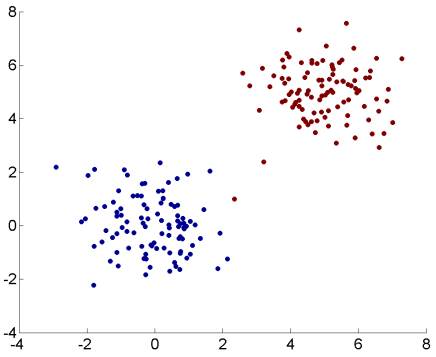
Figures from 8 to 13 contain the scatter plots of the real groups and the clusterings for the simplicial similarity and the Euclidean distance of the six examples. These clusterings are the result of cutting the dendrogram such that the number of resulting groups is the number of groups of each example.

In the easiest example, in Figure 8, the Euclidean distance does not make a mistake, whereas simplicial similarity erroneously categorizes one observation of the group with zero mean in the other group. In the case of example 2 (Figure 9), the one with correlated variable groups, neither Euclidean distance nor similarity make any clustering mistake. If one of the coordinates has much more variability than the other (Figure 10), Euclidean distance split both groups into halves and cluster such halves, obtaining around a 50 percent error. In this example, simplicial similarity performs much better than distance, with an error equal to 2.5%. If another group with covariance matrix identity and different mean is added, as in Figure 11, Euclidean distance with single linkage will obtain an error of 66% (with Ward, the rate is 29.7%). Again, simplicial similarity obtains the lowest value (a 3% error).

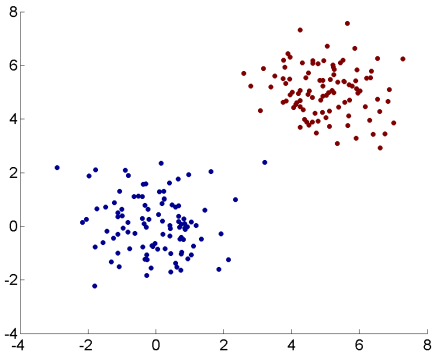
Regarding the sample with uniformly distributed groups (Figures 12 and 13), the results for Euclidean distance are very heterogeneous, because in the two-group example its maximum error is 9% (Ward method), whereas in the four-group example the percentage error is 64% (single linkage). Similarity shows errors equal to 8.5% and 0.2%, respectively.



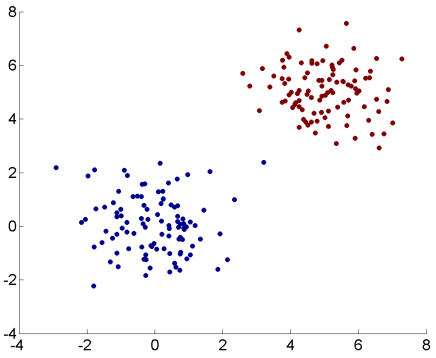
(a) Real clusters



(b) Simplicial similarity

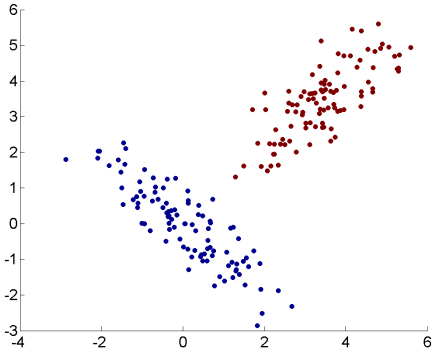


(c) Euclidean distance (Ward)

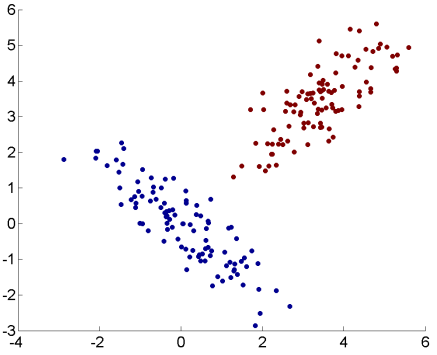


(d) Euclidean distance (single linkage)

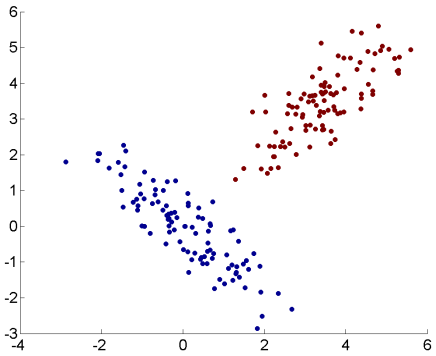
Figure 8: *Symmetric groups: Example 1.*



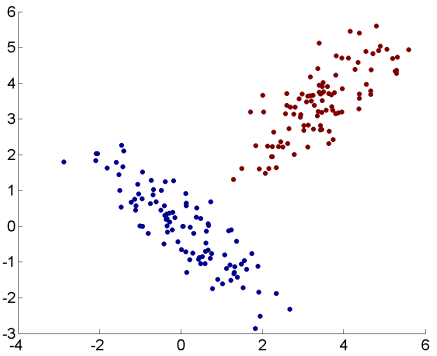
(a) Real clusters



(b) Simplicial similarity

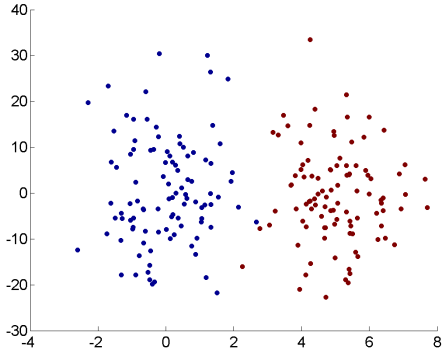


(c) Euclidean distance (Ward)

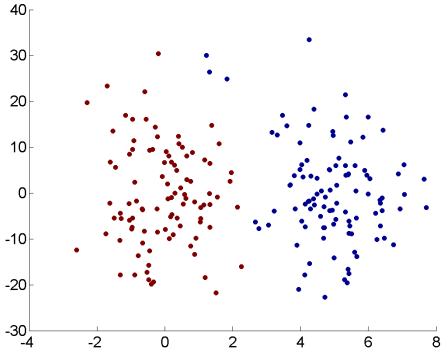


(d) Euclidean distance (single linkage)

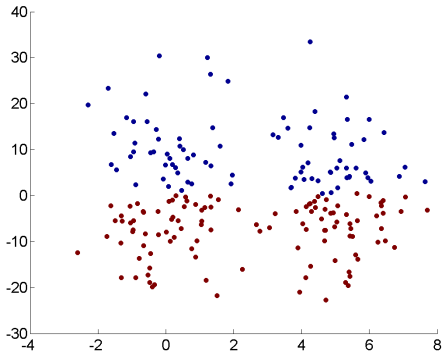
Figure 9: *Symmetric groups: Example 2.*



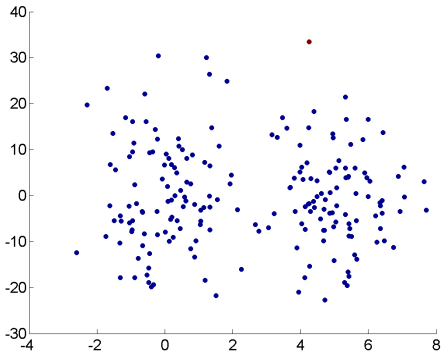
(a) Real clusters



(b) Simplicial similarity

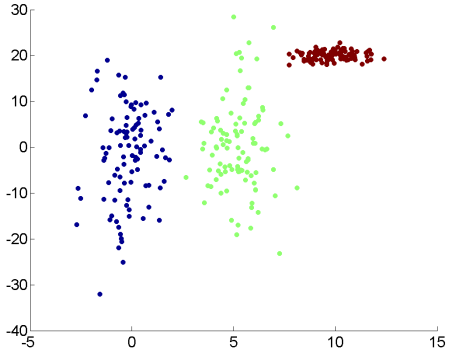


(c) Euclidean distance (Ward)

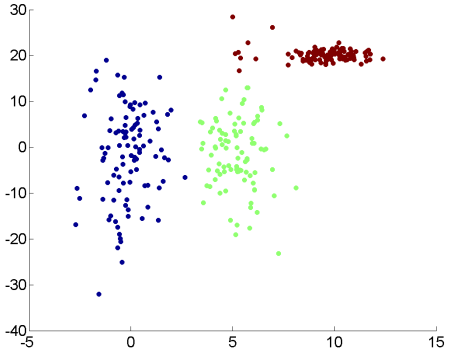


(d) Euclidean distance (single linkage)

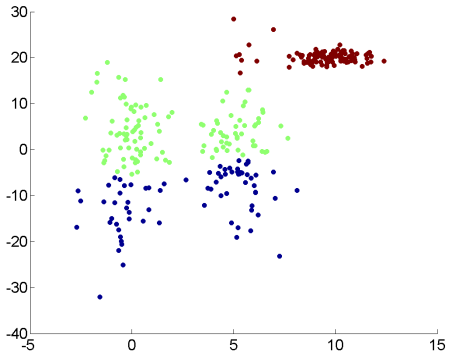
Figure 10: *Symmetric groups: Example 3.*



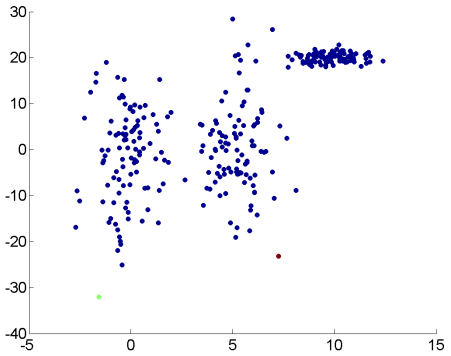
(a) Real clusters



(b) Simplicial similarity

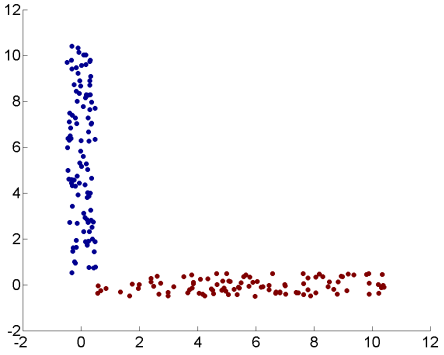


(c) Euclidean distance (Ward)

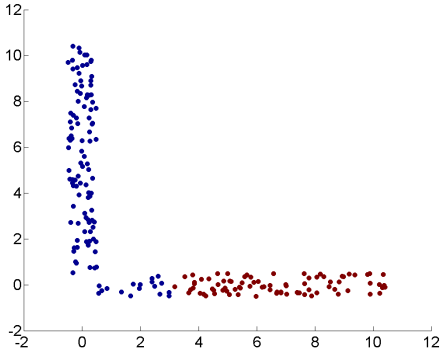


(d) Euclidean distance (single linkage)

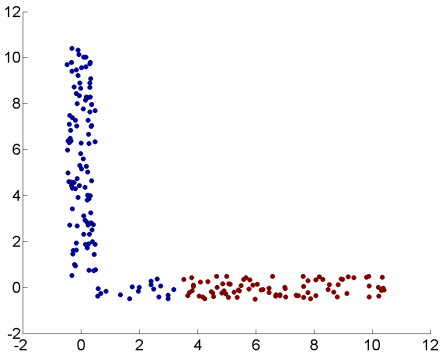
Figure 11: *Symmetric groups: Example 4.*



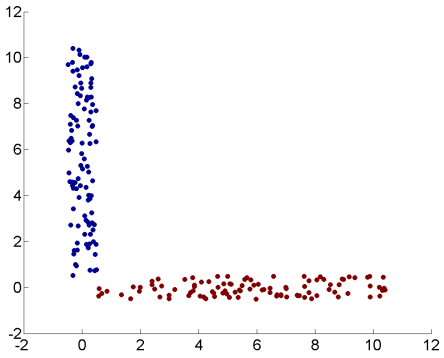
(a) Real clusters



(b) Simplicial similarity

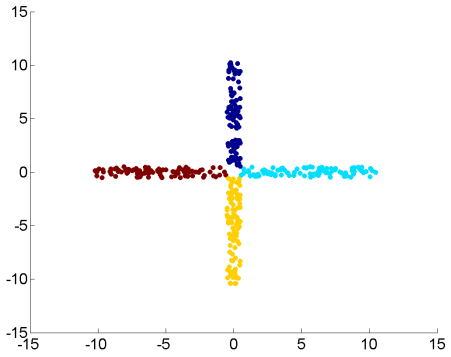


(c) Euclidean distance (Ward)

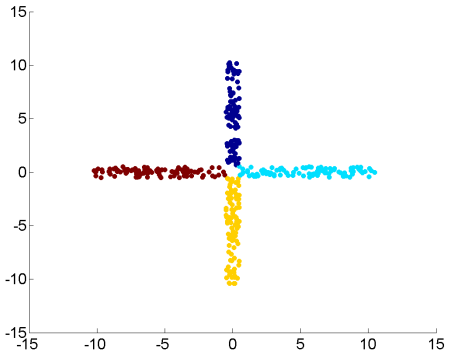


(d) Euclidean distance (single linkage)

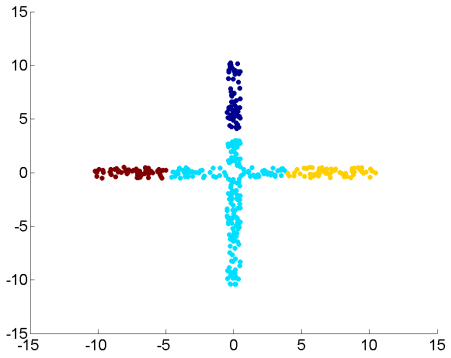
Figure 12: *Symmetric groups: Example 5.*



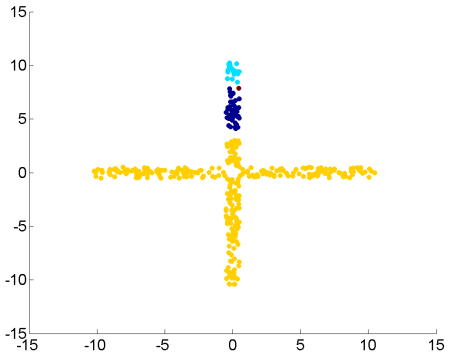
(a) Real clusters



(b) Simplicial similarity



(c) Euclidean distance (Ward)



(d) Euclidean distance (single linkage)

Figure 13: *Symmetric groups: Example 6.*

5.2.2 Assymmetric groups

The second group of samples is formed by data sets with some coordinates asymmetrically distributed. This asymmetry is introduced by means of the exponential distribution. Coordinate asymmetry increases the possibility to have values far from the coordinate median, and in such a case, it is possible that some distances be affected by outliers.

Example 1: Two groups of sizes $n_1 = n_2 = 100$. The first group has independent coordinates distributed according to an exponential random variable of mean 1. The coordinates of the second group have the same distribution, but with the opposite sign.

Example 2: Four groups of sizes $n_1 = n_2 = n_3 = n_4 = 100$. Two groups are simulated in the same way that those in the previous example. The third group is generated as the first one, but it is translated down to 4 units in the second coordinate. Also, the fourth group is generated as the second one, but translated up to 4 units.

Example 3: Two groups with an approximately rectangular shape, with an exponentially-distributed coordinate and another one uniformly distributed. Group sizes are $n_1 = n_2 = 100$. The first group has its first coordinate with distribution U (-0.5, 0.5), and the second one with exponential distribution of mean equal to 3 and origin 0.5. The second group has the first coordinate distributed as an exponential random variable with a mean equal to 3 and origin 0.5, and the second coordinate distributed as U (-0.5, 0.5).

Example 4: Four groups with an approximately rectangular shape, a uniform coordinate and an exponential coordinate. Group sizes are $n_1 = n_2 = n_3 = n_4 = 100$. Two of the groups have been drawn from the distributions of the previous example. The third group has the first coordinate distributed as a U (-0.5, 0.5), and the second one distributed as an exponential random variable of mean 3 with changed sign and origin -0.5. The first coordinate of the fourth group is distributed as an exponential of mean 3 with changed sign and origin -0.5, whereas the second one is distributed

as a U (-0.5, 0.5).

Figure 14 displays the scatter plots for the first example. Both simplicial similarity and Euclidean distance for the Ward method have an error percentage equal to zero. Euclidean distance for single linkage produces a cluster with one element; therefore, the error is close to 50%. When there are four exponentially-distributed groups (Figure 15), performance of the Euclidean distance for single linkage is the worst, with an error of 75%. Simplicial similarity is slightly better than the distance for the Ward method.

Figures 16 and 17 show the clusterings for the examples with rectangular groups. For both samples, the best clustering is that obtained with simplicial similarity, for which differences between examples are small. Again, the Euclidean distance for single linkage is the worst. For the Ward method, the error in the example with two groups is a ten percent lower than in the example with four groups.

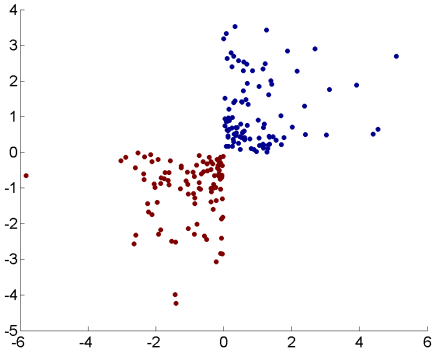
5.2.3 Nonlinear groups

The last group of examples corresponds to those in which at least one of the group presents non linear shapes. The shapes of the groups are circumferences, rings and ring halves. Coordinates are uniformly distributed in the region of each group. Next, the four examples are introduced.

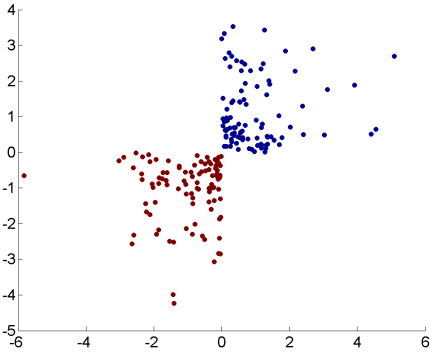
Example 1: Two groups of sizes $n_1 = n_2 = 200$. The first group is uniformly distributed in a circumference of radius 1 and center $(0, 0)'$, whereas the second one is uniformly distributed in a ring with the same center and radius 1.5 and 2.5.

Example 2: Two groups of sizes $n_1 = n_2 = 200$. The first group is uniformly distributed in a circumference of radius 1 and center $(0, 0)'$, whereas the second one is uniformly distributed in a half of a ring with the same center and radius 1.5 and 2.5.

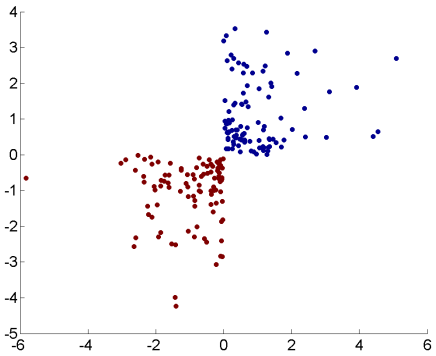
Example 3: Two groups of sizes $n_1 = n_2 = 200$. The first group is uniformly distributed in



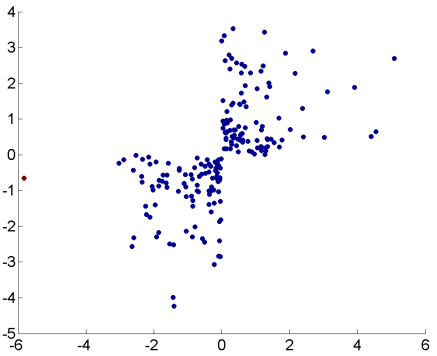
(a) Real clusters



(b) Simplicial similarity

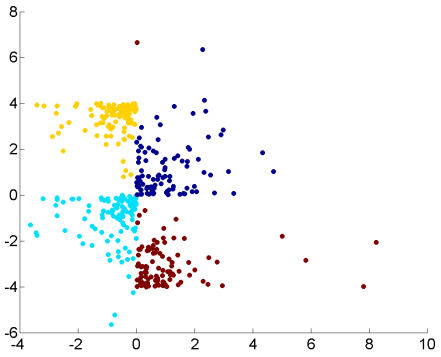


(c) Euclidean distance (Ward)

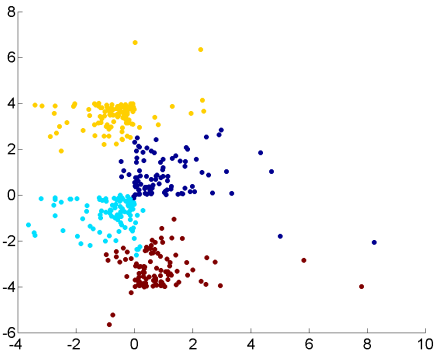


(d) Euclidean distance (single linkage)

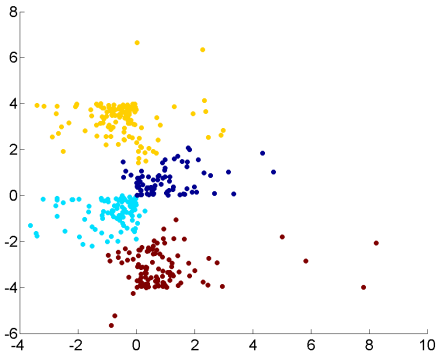
Figure 14: *Asymmetric groups: Example 1.*



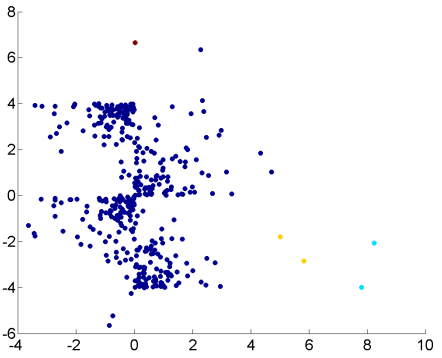
(a) Real clusters



(b) Simplicial similarity

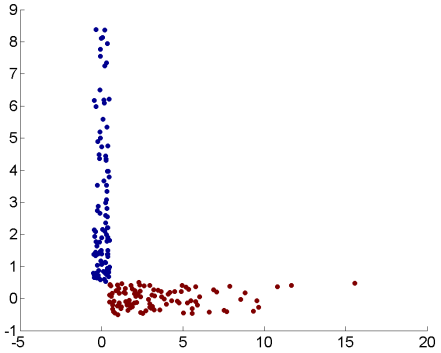


(c) Euclidean distance (Ward)

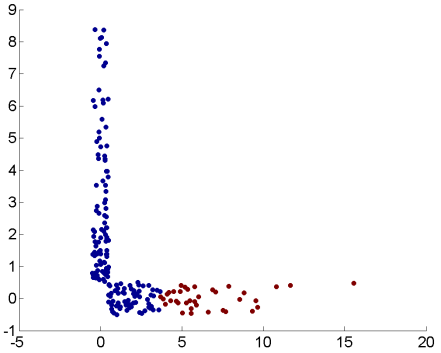


(d) Euclidean distance (single linkage)

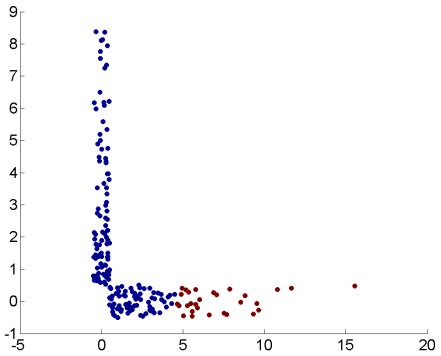
Figure 15: *Asymmetric groups: Example 2.*



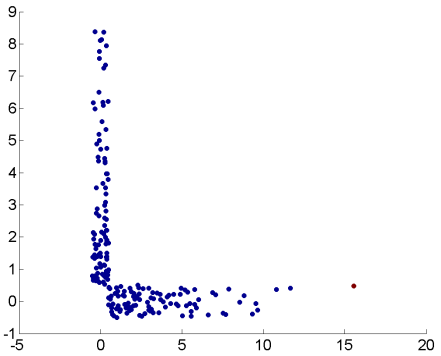
(a) Real clusters



(b) Simplicial similarity

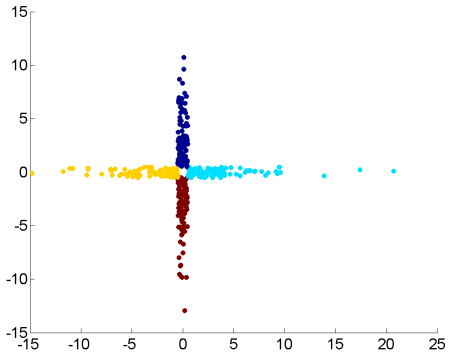


(c) Euclidean distance (Ward)

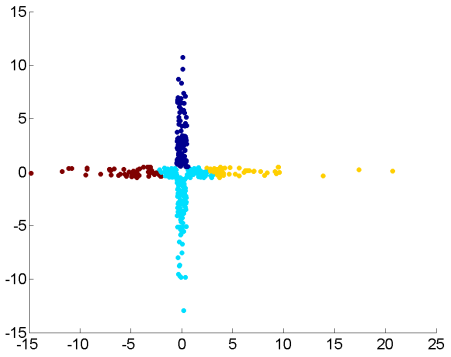


(d) Euclidean distance (single linkage)

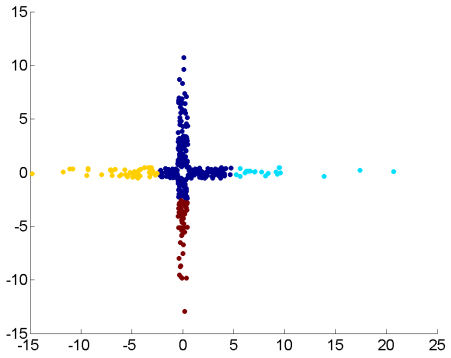
Figure 16: *Asymmetric groups: Example 3.*



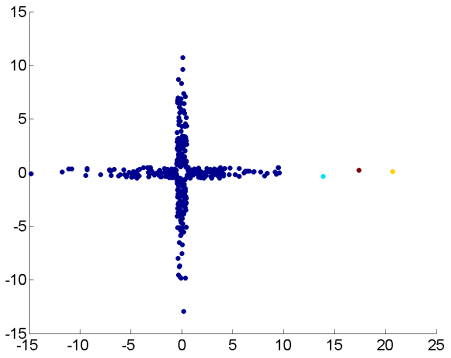
(a) Real clusters



(b) Simplicial similarity



(c) Euclidean distance (Ward)



(d) Euclidean distance (single linkage)

Figure 17: *Asymmetric groups: Example 4.*

the left half of a ring centered in the origin and radius 1.5 and 2.5, whereas the second group is uniformly distributed in the right half of a ring with the same radius, but centered in the point $(0, 2)'$.

Example 4: Two groups of sizes $n_1 = n_2 = 200$. The first group is uniformly distributed in the left half of a ring centered in the origin and radius 1.5 and 2.5, whereas the second group is uniformly distributed in the right half of a ring with the same radius, but centered in the point $(\text{-}0.75, 2)'$.

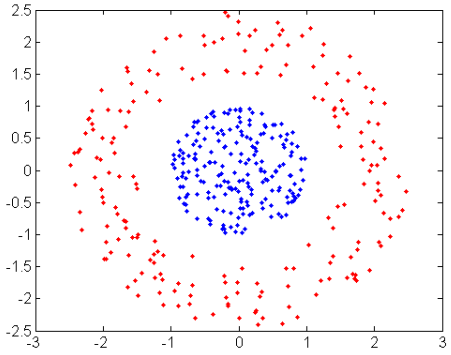
Figure 18 shows the scatter plots of the four samples. In three of them, a nonlinear classification rule is needed in order to distinguish the groups. Only in the third example, groups can be separated with a straight line.

The clustering results for these examples are represented in Figures 19 to 22. In the example with the circumference and the ring (Figure 19), the Euclidean distance with single linkage does not make mistakes, being simplicial similarity the worst measure (error of 40%). When only the half ring is considered (20), the error percentage decreases. In this case, simplicial similarity works better (16%) than Euclidean distance with the Ward method (23.5%), but far from the Euclidean distance with single linkage.

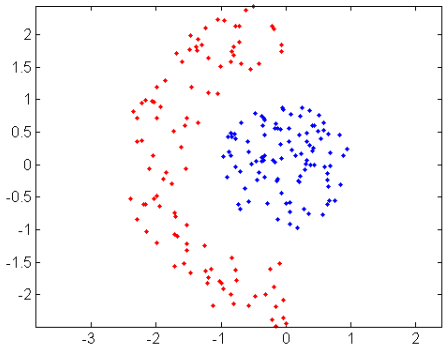
In the Example 3, Figure 21, similarity has an error close to zero. Again, with single linkage, the Euclidean distance gets the best result. Finally, in the last example, Figure 22, simplicial similarity performs worse than in the third example, but again it is better than the distance with the Ward method.

5.2.4 Comparison of results

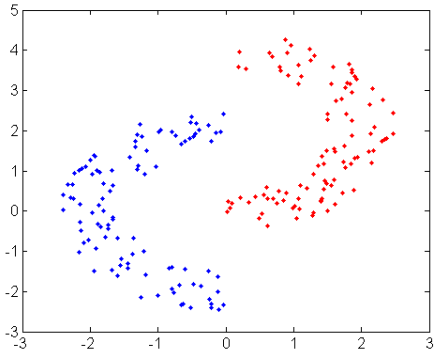
This section analyzes the results over the 14 examples. Table 3 contains the error percentages for each example divided by group types, and the global mean.



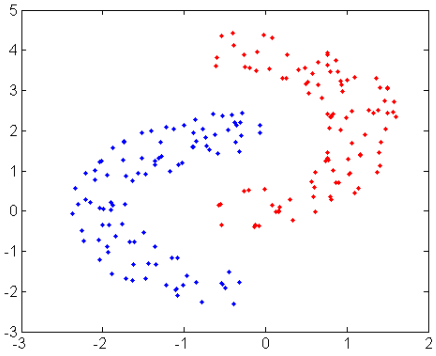
(a) Example 1



(b) Example 2

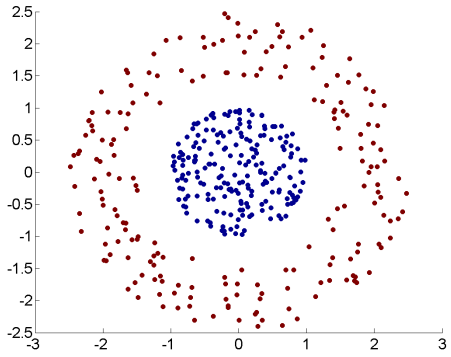


(c) Example 3

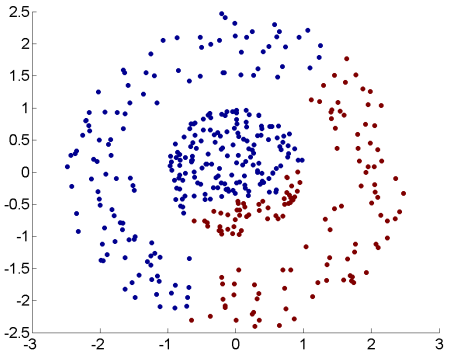


(d) Example 4

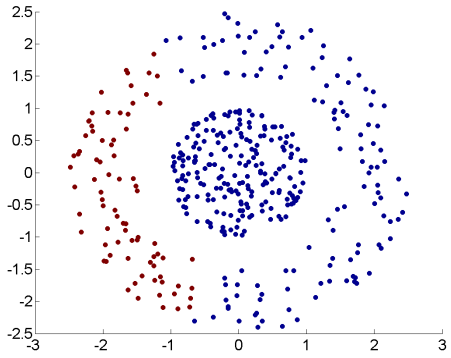
Figure 18: *Scatter plot for the examples with non-linear groups.*



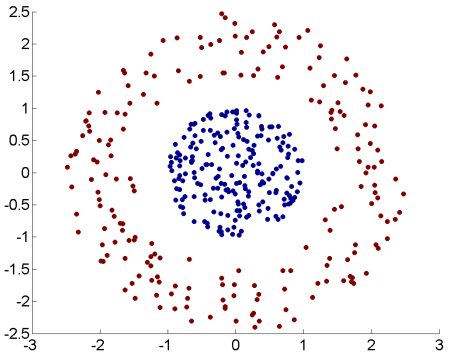
(a) Real clusters



(b) Simplicial similarity

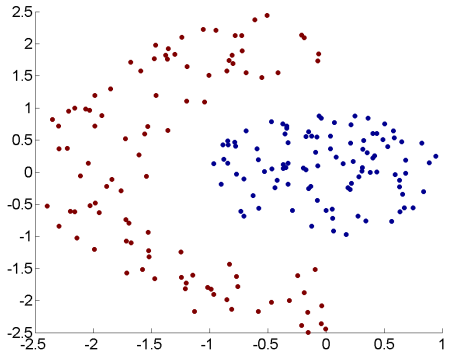


(c) Euclidean distance (Ward)

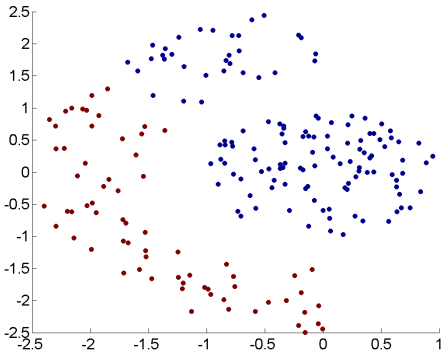


(d) Euclidean distance (single linkage)

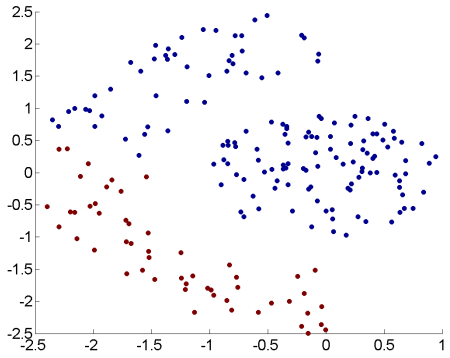
Figure 19: *Nonlinear groups: Example 1.*



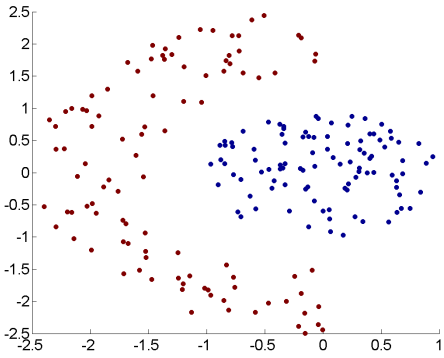
(a) Real clusters



(b) Simplicial similarity

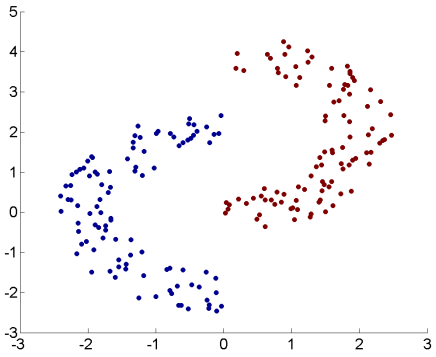


(c) Euclidean distance (Ward)

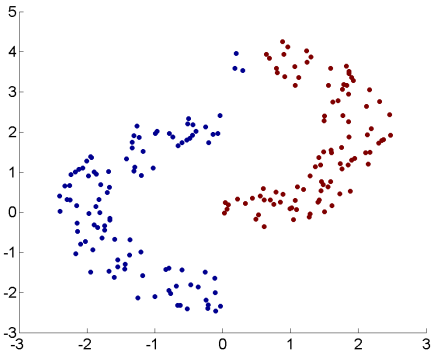


(d) Euclidean distance (single linkage)

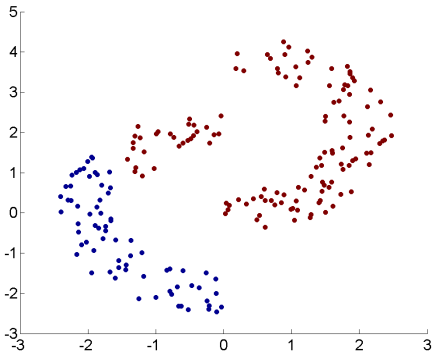
Figure 20: *Nonlinear groups: Example 2.*



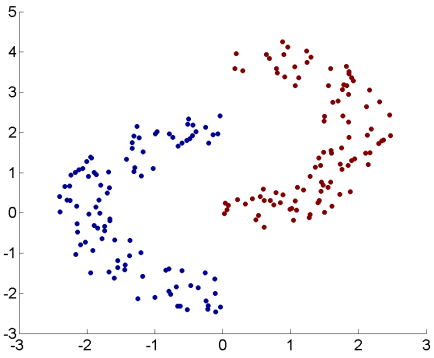
(a) Real clusters



(b) Simplicial similarity

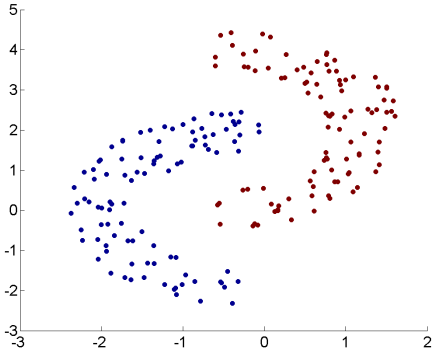


(c) Euclidean distance (Ward)

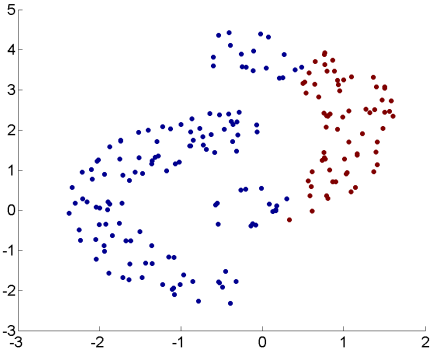


(d) Euclidean distance (single linkage)

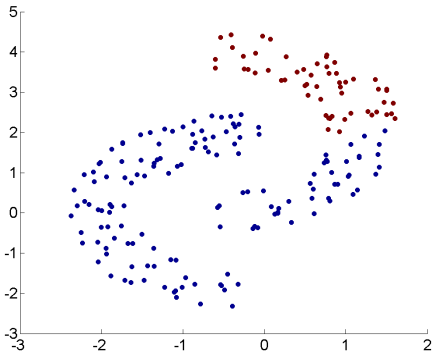
Figure 21: *Nonlinear groups: Example 3.*



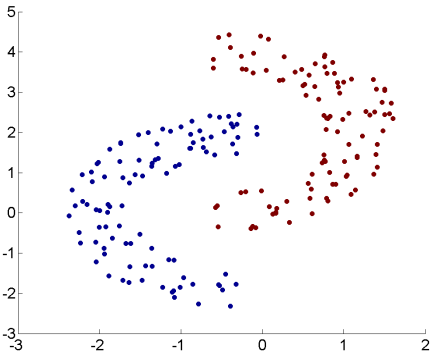
(a) Real clusters



(b) Simplicial similarity



(c) Euclidean distance (Ward)



(d) Euclidean distance (single linkage)

Figure 22: *Nonlinear groups: Example 4.*

		Euclidean distance		Simplicial
Type	Example	Ward	Single	Similarity
Symmetric groups	1	0,5	0,5	0,5
	2	0	0	0
	3	50	49,5	2,5
	4	29,7	66	3
	5	9	0	8,5
	6	26,5	64	0,2
Assymetric groups	1	0	0,5	0
	2	11,5	74,8	9,8
	3	35	49,5	31,5
	4	45,8	74,8	25,3
Non-linear groups	1	30,3	0	40,5
	2	23,5	0	16
	3	14,5	0	1,5
	4	22	0	16,5
	Mean	21,3	27,1	11,1

Table 3: *Percentages of the classification errors.*

According to the mean over all the examples, the function offering the best results is simplicial similarity, with a mean percentage error of 11.1%. Its maximum error (40.5%) is for the non-linear example with the circumference and the ring. In 7 of the 14 examples errors are below 5%.

Regarding Euclidean distance, in some examples single linkage performs better, and in others it does worse than the Ward method. For example, in samples with non-linear groups, the error for

single linkage is zero, whereas for Ward method the error is in all cases greater than 15%. Samples with asymmetric groups show the opposite situation: larger percentages for single linkage. As an average, Ward methods perform better than single linkage, 21.3% versus 27.1%.

The comparison between the two measures is, for most examples, favorable to simplicial similarity. It is not affected by the variability of the components, at least in normal groups. Distance is more affected by single linkage than by the Ward method. Exactly the same happens to asymmetry in the coordinates of the groups, since simplicial similarity is the less affected function by this asymmetry and, again, the Ward method obtains less errors than single linkage. In the last group of examples, Euclidean distance is clearly the best for single linkage (errors equal to zero in all cases). Similarity in these examples works better than distance with the Ward method. Globally, simplicial similarity has the lowest error percentages: on average, its errors are a 10% lower than the best choice for the Euclidean distance.

6 CONCLUSION

The main conclusion is that it is possible to use the statistical depth notion for defining similarity or proximity measures. This extension preserves the depth function ability to adapt measures to the shape of the data set or the reference distribution. The simplicial similarity with respect to F defined in this work is continuous if the distribution F is absolutely continuous. Several types of convergence of the sample version have also been proven. Its usefulness has been demonstrated through its application to hierarchical clustering. The results achieved over several examples are better than those obtained with Euclidean distance, since it is insensitive to different variability and to asymmetry of the variables, unlike Euclidean distance, which produces less robust clusterings in these situations.

REFERENCES

Barnett, V. (1976). Ordering of multivariate data. *Journal of the Royal Statistical Society Series A-Statistics in Society*, 139:318–354.

Dümbgen, L. (1992). Limit-theorems for the simplicial depth. *Statistics & Probability Letters*, 14(2):119–128.

Fraiman, R. and Meloche, J. (1999). Multivariate L-estimation. *Test*, 8(2):255–289.

Hartigan, J. (1975). *Clustering Algorithms*. John Wiley and Sons, New York.

Jornsten, R. (2004). Clustering and classification based on the L_1 data depth. *Journal of Multivariate Analysis*, 90(1):67–89.

Kaufman, L. and Rousseeuw, P. J. (1986). *Pattern Recognition in Practice II*, pages 425–437. Elsevier/North-Holland, Amsterdam.

Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of medoids. *Statistical Data Analysis based on the L_1-Norm*, pages 405–416.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data*. Wiley Series in Probability and Mathematical Statistics.

Koshevoy, G. and Mosler, K. (1997). Zonoid trimming for multivariate distributions. *Annals of Statistics*, 25(5):1998–2017.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *Annals of Statistics*, 18(1):405–414.

Liu, R. Y. and Singh, K. (1992). Ordering directional-data - concepts of data depth on circles and spheres. *Annals of Statistics*, 20(3):1468–1484.

López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. In *Proceedings National Institute of Science, India*, volume 2, pages 49–55.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1:327–332.

Tukey, J. W. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians, Vancouver*, pages 523–531.

Vardi, Y. and Zhang, C. H. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences of the United States of America*, 97(4):1423–1426.

Zuo, Y. J. and Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics*, 28(2):461–482.