Topological Principal Component Analysis

Rafik Abdesselam

Topological Principal Component Analysis

Rafik Abdesselam

ERIC-COACTIS Laboratories, University of Lyon, Lumière Lyon 2
Campus Berges du Rhône, 69635 Lyon Cedex 07, France
(E-mail: rafik.abdesselam@univ-lyon2.fr)
(http://perso.univ-lyon2.fr/~rabdesse/fr/)

Abstract. Topological Principal Component Analysis (TPCA) is a multidimensional descriptive method witch studies a homogeneous set of continuous variables defined on the same set of individuals. Its a topological method of data analysis that consists of comparing and classifying proximity measures from among some of the most widely used measures for continuous data. Its proposes an adjacency matrix associated to a proximity measure according to the data under consideration, then analyzes and visualizes, with graphic representations, the relationship structure of the variables relating to, the known problem of Principal Component Analysis (PCA). Based on the notion of neighborhood graphs, some of these proximity measures are more-or-less equivalent. A topological equivalence index between two measures is defined and statistically tested according to the topological correlation between the variables. The principle of the proposed TPCA is illustrated using a real data set.

Keywords: Proximity measure, neighborhood graph, adjacency matrix, topological equivalence, correlation matrix, MDS graphical representations.

1 Introduction

Similarity measures play an important role in many areas of data analysis. The results of any operation involving structuring, clustering or classifying objects are strongly dependent on the proximity measure chosen. The user has to select one measure among many existing ones. Yet, according to the notion of topological equivalence chosen, some measures are more-or-less equivalent. The concept of topological equivalence uses the basic notion of local neighborhood. We define the topological equivalence between two proximity measures, in the context of correlation observed between the continuous variables considered, through the topological structure induced by each measure.

Principal Component Analysis (PCA) [16,10,5,18] is an important methodology among factorial techniques due to the extent of its field of application. It allows us, among others things, to describe continuous data tables.

This method concerns the relations between or within a set of quantitative variables simultaneously observed on a sample of individuals. Generally the variables are homogeneous in the sense that they revolve around a particular theme.

6^{th} *SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain*

PCA is statistically considered as a widely used multivariate method for dimension reduction and as a technique of representing data. It aims is to find common factors, the socalled principal components, in form of linear combinations of the variables under investigation. It allows to have an idea of the correlations structure of the set of variables, as well as possible similarities of behavior between individuals.

In order to understand and act on situations that are represented by a set of objects, very often we are required to compare them. Humans perform this comparison subconsciously using the brain. In the context of artificial intelligence, however, we should be able to describe how the machine might perform this comparison. In this context, one of the basic elements that must be specified is the proximity measure between objects.

Certainly, application context, prior knowledge, data type and many other factors can help in identifying the appropriate measure. For instance, if the objects to be compared are described by Boolean vectors, we can restrict our comparisons to a class of measures specifically devoted to this type of data. However, the number of candidate measures may still remain quite large. Can we consider that all those measures remaining are equivalent and just pick one of them at random? Or are there some that are equivalent and, if so, to what extent? This information might interest a user when seeking a specific measure. For instance, in information retrieval, choosing a given proximity measure is an important issue. We effectively know that the result of a query depends on the measure used. For this reason, users may wonder which one is more useful? Very often, users try many of them, randomly or sequentially, seeking a "suitable" measure. If we could provide a framework that allows the user to compare proximity measures in order to identify those that are similar, they would no longer need to try out all measures.

The present study proposes a new framework for comparing proximity measures in order to choose the best one in the context of association between a set of quantitative variables. The aim is to establish a PCA.

We deliberately ignore the issue of the appropriateness of the proximity measure, as it is still an open and challenging question currently being studied. The comparison of proximity measures can be analyzed from various angles.

The comparison of objects, situations or ideas is an essential task in order to assess a situation, to rank preferences, to structure a set of tangible or abstract elements, and so on. In a word, to understand and act, we have to compare. These comparisons that the brain naturally performs, however, must be clarified if we want them to be done by a machine. For this purpose, we use proximity measures. A proximity measure is a function which measures the similarity or dissimilarity between two objects within a set. These proximity measures have mathematical properties and specific axioms. But are such measures equivalent? Can they be used in practice in an undifferentiated way? Do they produce the same learning database that will serve to find the membership class of a new object? If we know that the answer is negative, then how do we decide which one to use? Of course, the context of the study and the type of data being considered can help in selecting a few possible proximity measures,

but which one should we choose from this selection as the best measure for summarizing the correlation structure of the variables?

The topological correlation structure of the variables partly depends on the data being used. The results of TPCA are different according to the selected proximity measure.

Several studies on the topological equivalence of proximity measures have been proposed, [4,17,13,24], also in contexts of discrimination [3] and correspondences [2,1], but none of these propositions has an objective of the correlations synthesis of a set of quantitative variables.

Therefore, this article focuses on how to construct the best adjacency matrix induced by a proximity measure, taking into account the association between all the modalities of the qualitative variables.

In this paper we compare different proximity measures in an aim to synthesize the relationships of a set of continuous variables in the topological context. Comparison of these measures show that the results are different and depending on the proximity measure chosen. The rest of the paper is organized as follows. In section 2, we discuss topological equivalence between two proximity measures and show how to build an adjacency matrix associated with a proximity measure, how to compare and statistically test the degree of topological equivalence between proximity measures and how to select the best measure to describe topologically the structure of the correlations of the variables. Section 3 presents an illustrative example and surveys existing proximity measures on continuous data and presents a comparison between them. This comparison helps the researchers to take quick decision about which measure to use for considered data. A conclusion of this work is given in section 4.

Table 7 in Appendix summarizes some classic proximity measures used for continuous data [23], we give on $\mathbb{R}^n$ the definition of 15 of them.

We assume that we have at our disposal $\{x^k; k = 1, .., p\}$ a set of p homogeneous quantitative variables measured on n individuals. The interest is to analyze the topological structure of all these variables.

2 Topological Correlation

Topological equivalence is based on the concept of the topological graph also referred to as the neighborhood graph. The basic idea is actually quite simple: two proximity measures are equivalent if the corresponding topological graphs induced on the set of objects remain identical. Measuring the similarity between proximity measures involves comparing the neighborhood graphs and measuring their similarity. We will first define more precisely what a topological graph is and how to build it. Then, we propose a measure of proximity between topological graphs that will subsequently be used to compare the proximity measures.

Consider a set $E = \{x^1, x^2, \ldots, x^k, \ldots, x^p\}$ of p objects in $\mathbb{R}^n$, associated with the p variables. We can, by means of a proximity measure u, define a

neighborhood relationship V_u to be a binary relationship on $E \times E$. There are many possibilities for building this neighborhood binary relationship.

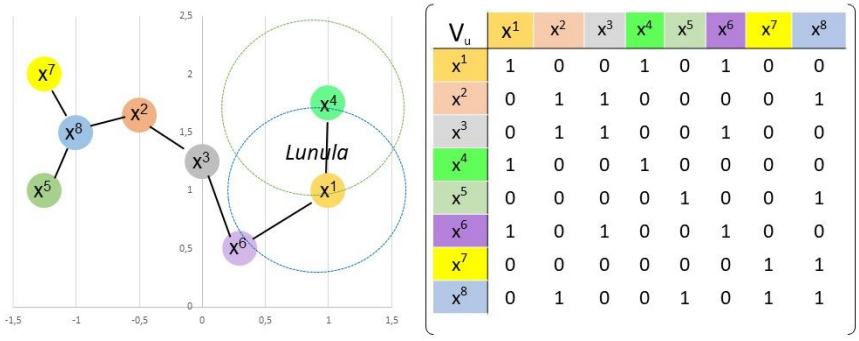
Thus, for a given proximity measure u, we can build a neighborhood graph on a set of objects-variables, where the vertices are the variables and the edges are defined by a property of the neighborhood relationship.

Many definitions are possible to build this binary neighborhood relationship. One can choose the Minimal Spanning Tree (MST) [11], the Gabriel Graph (GG) [15] or, as is the case here, the Relative Neighborhood Graph (RNG) [21].

For any given proximity measure u, we construct the associated adjacency binary symmetric matrix V_u of order p, where, all pairs of neighboring variables (x^k, x^l), where $k, l = 1, p$, satisfy the following RNG definition.

Definition 1: Relative Neighborhood Graph (RNG)

$$\begin{cases} V_u(x^k, x^l) = 1 \quad if \quad u(x^k, x^l) \leq \max[u(x^k, x^r), u(x^r, x^l)] \;\; ; \;\; \forall x^k, \; x^l, \; x^r \in E, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad x^r \neq x^k \;\; and \;\; x^r \neq x^l \\ V_u(x^k, x^l) = 0 \quad otherwise \end{cases}$$



V_u	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
x^1	1	0	0	1	0	1	0	0
x^2	0	1	1	0	0	0	0	1
x^3	0	1	1	0	0	1	0	0
x^4	1	0	0	1	0	0	0	0
x^5	0	0	0	0	1	0	0	1
x^6	1	0	1	0	0	1	0	0
x^7	0	0	0	0	0	0	1	1
x^8	0	1	0	0	1	0	1	1

Fig. 1. RNG example with eight variables - Associated adjacency matrix

This means that if two variables x^k and x^l which verify the RNG property are connected by an edge, the vertices x^k and x^l are neighbors.

Thus, for any proximity measure given, u, we can associate an adjacency matrix V_u, of binary and symmetrical order p. Figure 1 illustrates an example of RNG in $\mathbb{R}^2$ of a set of $p = 8$ objects-variables.

For example, for the first and four variables, $V_u(x^1, x^4) = 1$, it means that on the geometrical plane, the hyper-Lunula (intersection between the two hyperspheres centered on the two variables x^1 and x^4) is empty.

For a given neighborhood property (MST, GG or RNG), each measure u generates a topological structure on the objects in E which are totally described by the adjacency binary matrix V_u. In this paper, we chose to use the Relative Neighbors Graph (GNR).

2.1 Comparison and selection of proximity measures

First we compare different proximity measures according to their topological similarity in order to regroup them and to better visualize their resemblances.

To measure the topological equivalence between two proximity measures u_i and u_j, we propose to test if the associated adjacency matrices V_{u_i} and V_{u_j} are different or not. The degree of topological equivalence between two proximity measures is measured by the following definition of concordance.

Definition 2: Topological equivalence between two adjacency matrices

$$S(V_{u_i}, V_{u_j}) = \frac{1}{p^2} \sum_{k=1}^{p} \sum_{l=1}^{p} \delta_{k\,l}(x^k,\, x^l)$$

$$\text{with} \quad \delta_{k\,l}(x^k,\, x^l) = \begin{cases} 1 & \text{if } \ \mathrm{V}_{u_i}(x^k,\, x^l) = V_{u_j}(x^k, x^l) \\ 0 & \text{otherwise.} \end{cases}$$

Then, in our case, we want to compare these different proximity measures according to their topological equivalence in a context of correlation. So we define a criterion for measuring the deviation from the position of independence.

The data can arise from several different sampling frameworks, and the interpretation of the hypothesis of no association depends on the framework. The question of interest is whether there is a correlation between the two variables.

We construct the adjacency matrix denoted by $V_{u_\star}$, which corresponds to the correlation matrix. Thus, to examine the correlation structure between the variables, we examine the significance of their linear correlation coefficient. This adjacency matrix can be written as follows using the t-test of the linear correlation coefficient ρ of Bravais-Pearson:

Definition 3: Adjacency matrix $V_{u_\star}$ associated to reference measure $u_\star$

$$\begin{cases} \mathrm{V}_{u_\star}(x^k, x^l) = 1 & \text{if} \quad \text{p-value} = \mathrm{P}[\ |\ T_{n-2}\ | > \text{t-value}] \leq \alpha\ ;\ \forall k, l = 1, p \\ \mathrm{V}_{u_\star}(x^k, x^l) = 0 & \quad \text{otherwise} \end{cases}$$

Where p-value is the significance test of the correlation coefficient for the two-sided test of the null and alternative hypotheses, $H_0 : \rho(x^k\,,\, x^l) = 0$ vs. $H_1 : \rho(x^k\,,\, x^l) \neq 0$.

The p-value is the evidence against a null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis which means that there is no correlation between x^k and x^l variables in the population.

Formula for the Student t-test for significance of correlation: $t = r\sqrt{\frac{n-2}{1-r^2}}$ with $\nu = n - 2$ degrees of freedom (d.f.) and $r = r(x^k\,,\, x^l)$ is the linear correlation coefficient observed between the variables x^k and x^l.

Let T_{n-2} be a t-distributed random variable of Student with $\nu = n - 2$ d.f. In this case, the null hypothesis is rejected with a p-value less or equal a chosen α significance level, for example $\alpha = 5\%$. Using linear correlation test, if the p-value be very small, it means that there is very small opportunity that null hypothesis is correct, and consequently we can reject it. Statistical significance in statistics is achieved when a p-value is less than the significance level of α.

The p-value is the probability of obtaining results which acknowledge that the null hypothesis is true.

The robustness according to the α error risk chosen for the null hypothesis: no linear correlation, can be studied by setting a minimum threshold in order to analyze the sensitivity of the results. Certainly the numerical results will change, but probably not their interpretation.

The binary and symmetric adjacency matrix build $V_{u_\star}$, is associated with an unknown proximity measure denoted $u_\star$ and called a reference measure. Thus, with this reference proximity measure we can establish $S(V_{u_i}, V_{u_\star})$, the topological equivalence between the two proximity measures u_i and $u_\star$, by measuring the percentage of similarity between the adjacency matrix V_{u_i} and the reference adjacency matrix $V_{u_\star}$.

In order to graphically describe the similarities between proximity measures, we can for example apply the notion of themascope [12], which is a methodological sequence of a clustering method on the results of a factorial method. In this case, a Principal Component Analysis (PCA) followed by a Hierarchical Ascendant Classification (HAC) were performed upon the 15 component dissimilarity matrix defined by: $[D]_{ij} = D(V_{u_i}, V_{u_j}) = 1 - S(V_{u_i}, V_{u_j})$ to partition them into homogeneous groups and to view their similarities in order to see which measures are close to one another.

We can use any classic visualization techniques to achieve this. For example, we can build a dendrogram of hierarchical clustering of the proximity measures. We can also use multidimensional scaling or any other technique, such as Laplacian projection, to map the 15 proximity measures into a two dimensional space.

Finally, in order to evaluate and determine the closest class of proximity measures to the reference measure $u_\star$, we project the latter as a supplementary element into the two data analysis methods, positioned by the dissimilarity vector with 15 components $[D]_{*i} = 1 - S(V_{u_\star}, V_{u_i})$.

2.2 Statistical comparisons between two proximity measures

In this section, we use the Fisher's Exact Test [9] which is an alternative to the Chi-square test when the samples are small. The principle of this test is to determine if the configuration observed in the contingency table is an extreme situation compared to the possible situations taking into account the marginal distributions. Fisher's exact test is an exact statistical test used for the analysis of contingency tables. It is a test qualified as exact because the probabilities can be calculated exactly rather than relying on an approximation which becomes correct only asymptotically as for the chi-square test used in the contingency tables. It is not based on a test statistic whose law is known when n is large enough but it calculates, as its name suggests, the exact p-value directly. To test statistically the topological equivalence between two proximity measures. This non parametric test compares these measures based on their associated adjacency matrices. Two proximity measures are statistically in topological equivalence if the null hypothesis H_0 of independence is rejected.

The comparison between indices of proximity measures has also been studied by [19], [20] and [7] from a statistical perspective. The authors proposed an approach that compares similarity matrices obtained by each proximity measure, using Mantel's test [14], in a pairwise manner.

Fisher's exact test is the statistical test best suited to compare matched binary data, the Cohen's Kappa test [6] also but it is in general an asymptotic test. The Kendall or Spearman coefficient compares matched continuous data.

It makes it possible in this context to measure the agreement or the concordance of the binary values of two adjacency matrices associated with two proximity measures. The Fisher's exact test between two adjacency matrices evaluates the topological equivalence between their proximity measures.

Let V_{u_i} and V_{u_j} be adjacency matrices associated with two proximity measures u_i and u_j. To compare the degree of topological equivalence between these two measures, we propose to test if the associated adjacency matrices are statistically different or not, using a non-parametric test of paired data. These binary and symmetric matrices of order p, are unfolded in two vector-matched components, consisting of $\frac{p(p+1)}{2}$ values: the p diagonal values and the $\frac{p(p-1)}{2}$ values above or below the diagonal.

The degree of topological equivalence between two proximity measures is evaluated from the Fisher's exact test, computed on the 2×2 contingency table formed by the two binary vectors of order $\frac{p(p+1)}{2}$.

We also test the topological equivalence between each proximity measure $u_{i=1,15}$ and the reference measure $u_\star$ by comparing the adjacency matrices V_{u_i} and $V_{u_\star}$.

2.3 Graphical representations - Variables and Individuals

In order to represent graphically the possible topological links between the p quantitative variables, we use MultiDimensional Scaling (MDS) which makes it possible to find, for any distance matrix (similarity or dissimilarity) of size $p \times p$, a set of p points identified by their Euclidean coordinates whose distance matrix is equal to or very close to the given distance matrix. We propose to carry out the classical MDS [5], namely factorial analysis on similarity $V_{u_\star}$ or dissimilarity $D_{u_\star} = U - V_{u_\star}$ table, where $U = 1_p \, {}^t1_p$ is the $p \times p$ matrix of 1s and 1_p denotes the p indicator vector of 1s.

Definition 4:

TPCA consist to perform the standardized PCA of the triple $\{V_{u_\star} \; ; \; M \; ; \; D_p\}$, where, $V_{u_\star}$ is the adjacency matrix associated with the proximity measure $u_\star$, the most appropriate measure for the considered data, $M = I_p$ is the identity matrix of order p and $D_p = \frac{1}{p} I_p$ is the weighted diagonal matrix of variable weights.

The TPCA can be performed from any adjacency matrix V_{u_i} associated with each of the 15 proximity measures u_i considered.

Aid for the interpretation of TPCA results are those of PCA. Graphical representations on factorial plans allow to visualize and identify the topological structure of the variables. As in PCA, for representations of variables, we consider the most significant variables on the axes, that is the variables highly correlated with factors, having a strong contribution and a good quality of representation, measured by the square cosine of the angle between main axes and initial axes.

For representations of active individuals, these are projected as illustrative elements. The quality of representation of these individuals on the factorial axes is measured by their squared cosine.

3 Illustrative example and Empirical results

To illustrate the TPCA, we use Eurostat data [8] on government finance of the 28 European Union (EU) countries in 2017. We examine how key government finance statistics have developed in the EU-28. Specifically, it considers general government gross debt, deficit/surplus, total revenue and total expenditure. Simple statistics of the considered variables are displayed in Table 1.

Table 1. Summary statistics of public finances

Variable	Frequency	Mean	Standart Deviation (N)	Coefficient of variation (%)	Min	Max
Debt	28	68.043	36.539	53.70	8.70	176.10
Deficit	28	-0.264	1.692	640.07	-3.10	3.50
Revenues	28	42.579	6.654	15.63	26.00	53.80
Expenditures	28	42.850	6.793	15.85	26.30	56.50

In a metric and classical context, we simply have to apply a standardised PCA on the homogeneous set of the 4 characteristics of the government finance of the EU-28.

In a topological context, the main results of the proposed method are presented in the following tables and graphs, which allow us to visualize proximity measures close to each other and to select the one that best describes and synthesis, the government finance of the EU-28.

The objective here is to give a topological synthesis of the public finances of the EU countries in 2017.

An HAC algorithm based on the Ward criterion [22] was used in order to characterize classes of proximity measure relative to their similarities. The reference measure $u_\star$ is projected as a supplementary element. The dendrogram of Figure 2 represents the hierarchical tree of the 15 proximity measures considered. Table 2 describes the final composition of each class of proximity measures, the results of the chosen partition into three homogeneous classes, obtained from the cut of the hierarchical tree of Figure 2.

Aggregation based on the criterion of the loss of minimal inertia.

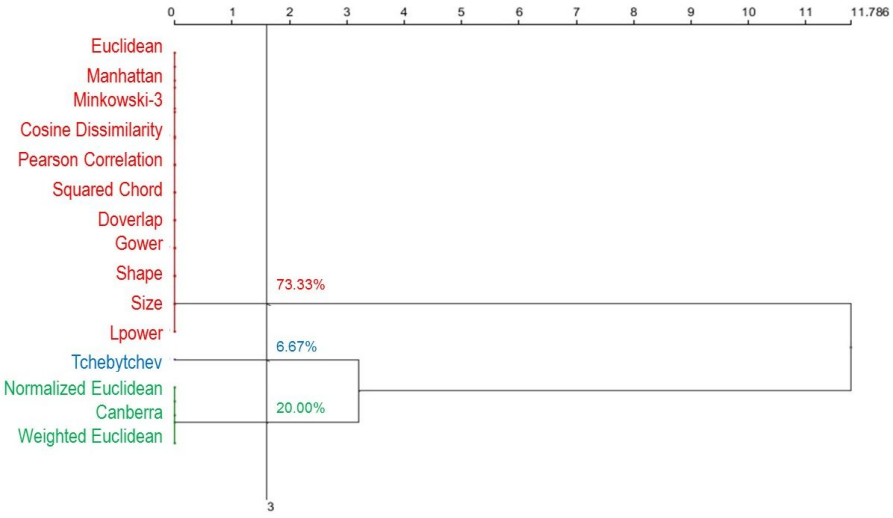


Fig. 2. Hierarchical tree of the proximity measures

Moreover, in view of the results in Table 2, the reference measure $u_\star$ is closer to the third class consisting of Normalized Euclidean, Canberra and Weighted Euclidean measures for which there is a strong topological association between the variables of government finance of EU-28 among the 15 proximity measures considered.

Table 2. Clusters composition - Assignment of the reference measure

Cluster number	Cluster 1	Cluster 2	Cluster 3
Frequency	11	1	3
Proximity measure	*Euclidean* *Manhattan* *Minkovski* $-$ 3 *Cosine Dissimilarity* *Pearson Correlation* *Squared Chord* *Doverlap, Gower* *Shape, Size, Lpower*	*Tchebytchev*	*Canberra* *Normalized Euclidean* *Weighted Euclidean*
Reference measure			$u_\star$

It was shown in [24], by means of a series of experiments, that the choice of proximity measure has an impact on the results of a supervised or unsupervised classification.

In a topological framework, Table 3 summarizes all the results of Table 8 given in the Appendix, the similarities and Fisher's Exact p-values between

Table 3. Similarities and Fisher's Exact Test

u_i	u_j	$S(u_i\ ,\ u_j)$	$p-value$
Cluster 1	Cluster 1	1.0000	0.0083**
Cluster 1	Cluster 2	0.7500	0.1833
Cluster 1	Cluster 3	0.7500	0.1833
Cluster 2	Cluster 2	1.0000	0.0083**
Cluster 2	Cluster 3	0.5000	1.0000
Cluster 3	Cluster 3	1.0000	0.0083**
$u_\star$	Cluster 1	0.7500	0.1833
$u_\star$	Cluster 2	0.6250	0.5000
$u_\star$	Cluster 3	0.8750	0.0333*

Significance level α ; $^{**}\alpha \leq 1\%$; $^{*}\alpha \in]1\%; 5\%]$

Table 4. 2×2 Contingency Table - Similarity - Fisher's Exact Test

Cluster 2 Cluster 1 : Euclidean

Tchebytchev	$V_{u_2}=0$	$V_{u_2}=1$
$V_{u_1}=0$	2	1
$V_{u_1}=1$	1	6

$S(V_{u_2}\ ,\ V_{u_1}) = 75\%$; $p\text{-}value = 0.1833$

Mesure Cluster 1 : Euclidean

Reference	$V_{u_2}=0$	$V_{u_2}=1$
$V_{u_\star}=0$	3	1
$V_{u_\star}=1$	0	6

$S(V_{u_\star}\ ,\ V_{u_1}) = 75\%$; $p\text{-}value = 0.183$

Cluster 3 Cluster 2 : Tchebytchev

Canberra	$V_{u_2}=0$	$V_{u_2}=1$
$V_{u_1}=0$	1	2
$V_{u_1}=1$	2	5

$S(V_{u_3}\ ,\ V_{u_2}) = 50\%$; $p\text{-}value = 1.000$

Mesure Cluster 2 : Tchebytchev

Reference	$V_{u_2}=0$	$V_{u_2}=1$
$V_{u_\star}=0$	2	2
$V_{u_\star}=1$	1	5

$S(V_{u_\star}\ ,\ V_{u_2}) = 62.50\%$; $p\text{-}value = 0.500$

Cluster 1 Cluster 3 : Canberra

Euclidean	$V_{u_2}=0$	$V_{u_2}=1$
$V_{u_1}=0$	2	1
$V_{u_1}=1$	1	6

$S(V_{u_1}\ ,\ V_{u_3}) = 75\%$; $p\text{-}value = 0.1833$

Mesure Cluster 3 : Canberra

Reference	$V_{u_2}=0$	$V_{u_2}=1$
$V_{u_\star}=0$	3	1
$V_{u_\star}=1$	0	6

$S(V_{u_\star}\ ,\ V_{u_3}) = 87.50\%$; $p\text{-}value = 0.0333^{*}$

Significance level α ; $^{**}\alpha \leq 1\%$; $^{*}\alpha \in]1\%; 5\%]$

all the $C_{15}^2 = 105$ pairs of proximity measures formed with the 15 measures considered and the 15 pairs formed with the unknown reference measure $u_\star$. The values below the diagonal correspond to the similarities $S(V_{u_i}, V_{u_j})$ and the values above the diagonal are the Fisher's Exact test p-values.

The similarities in pairs between the 15 proximity measures differ somewhat: some are closer than others. Some measures are in perfect topological equivalence $S(V_{u_i}, V_{u_j}) = 1$ with a significant Fisher's exact test p-value $< 5\%$; these are therefore identical for the data considered, as is the case with the measures in each class of the partition presented in Table 2.

The Table 4 illustrates the contingency tables 2×2 between the measures of each class: Euclidean, Tchebytchev, Canberra and reference measure $u_\star$ for the calculation of Fisher's exact test.

Only the topological equivalence between the reference measure and the Canberra measure is significant, p-value $= 0.0034 < \alpha = 5\%$, the null hypothesis H_0 of independence is rejected.

Table 5. Pearson correlation matrix (p-value)

Variables	Debt	Deficit	Revenues	Expenditures
Debt	1.000			
Deficit	-0.3403	1.000		
	(0.076)			
Revenues	0.3071	0.0393	1.000	
	(0.112)	(0.8428)		
Expenditures	0.3845	-0.2092	0.9689	1.000
	(0.0434*)	(0.2853)	(0.0001**)	

Significance level α ; $^{**}\alpha \leq 1\%$; $^{*}\alpha \in]1\%; 5\%]$

$$V_{u_*} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

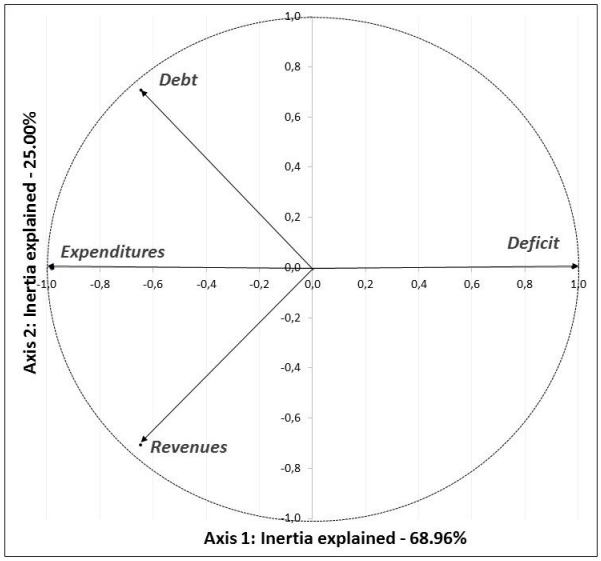


Fig. 3. TPCA - Adjacency matrix - The public finance variables on the first principal plane

The adjacency matrix $V_{u_\star}$ associated to the adapted proximity measure $u_\star$ to the considered data, is build from the correlation matrix Table 5 according

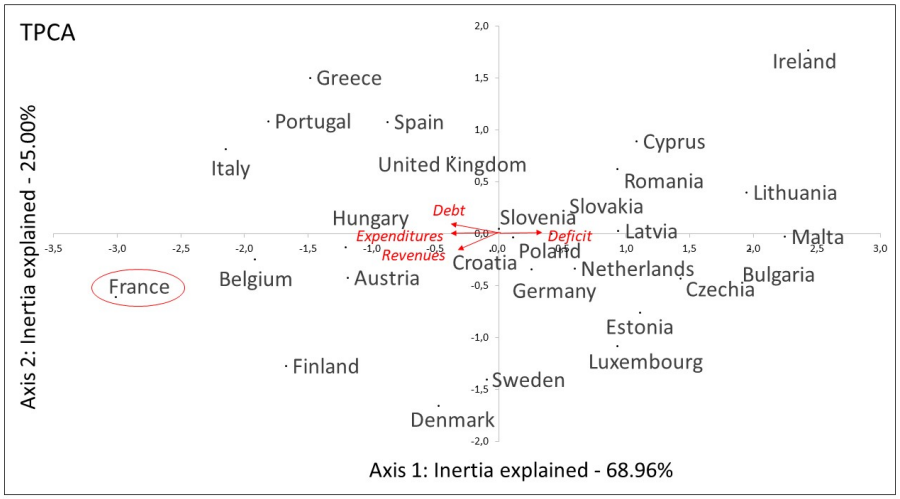


Fig. 4. TPCA - The EU-28 countries on the first principal plane

to definition 3. Figure 3 shows on the main first TPCA plane, the topological correlation between the Government finance variables.

The corresponding representation for individuals is given in Figure 4. It is thus possible to suggest which are the variables - government finance are responsible for the proximities between the individuals - the 28 EU countries.

The main numerical and graphical results of the proposed TPCA are given in the following Tables and Figures, and are compared to those of the classical PCA.

Figure 5 presents, for comparison on the first factorial plane, the correlations between principal components - Factors and the original variables. We can see that these graphical representations of the variables are slightly different. Effectivelly, the percentage of inertia explained on the first principal plane of the Topological PCA is greater than that of Classical PCA and the significant correlations variables-factors are also different.

Table 6 shows that the two first factors of TPCA explain 68.96% and 25.00%, respectively, they account for 93.96% of the total variation in the dataset, while the two first factors of classical PCA sum up that 84.88%. Thus, the first two factors provide an adequate summary of the data, i.e. of government finance of EU-28 countries, we restrict the comparison of the graphical representations to the first factorial plane.

The correlation tables show that the original variables are strongly correlated with the factors, those that contribute the most to the achievement of this principal component.

While the first PCA factor (55.61%) is strongly correlated with three of the original variables, expenditures, revenues and debt, the first TPCA factor (68.96%) opposes these three variables to the deficit. As for the second PCA (29.27%) and TPCA (25.00%) factors, they oppose the debt to revenues.

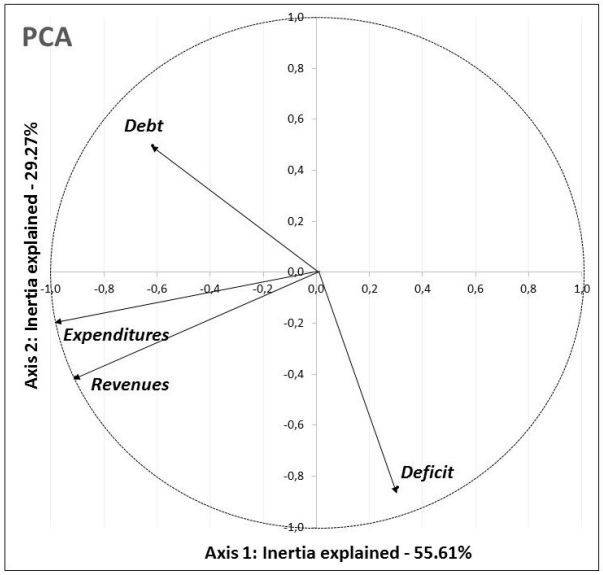


Fig. 5. PCA: The public finance variables on the first principal plane

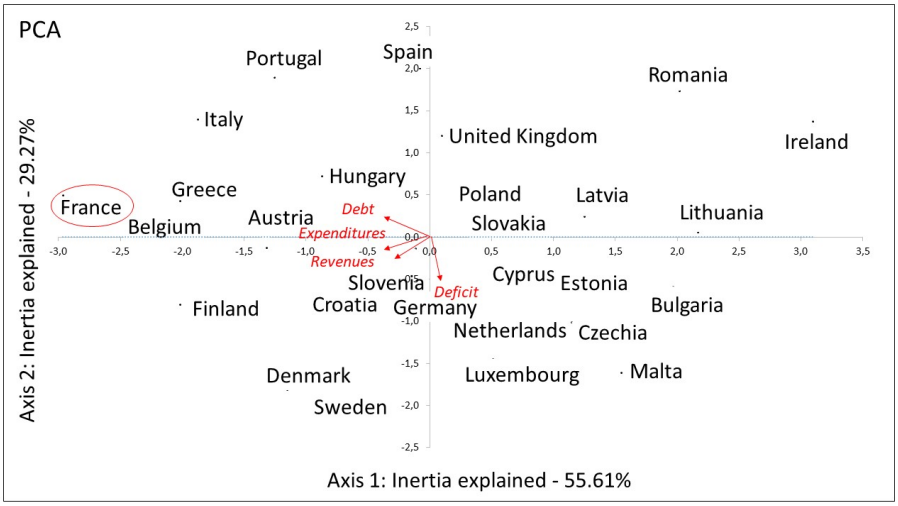


Fig. 6. PCA - The EU countries on the first principal plane

The representations of the countries presented in Figures 4 and 6 are of course slightly different, indeed, for example, for France which contributes to the realization of the first TPCA axis, it is characterized by high Debts, high Expenditures, high Revenues and a low Deficit. France also contributes on the first PCA axis, its characterized by high Debts, high Expenditures and high Revenues, but the Deficit does not characterize the first factorial axis of the PCA.

Table 6. TPCA and PCA - Eigenvalues and Correlations Variables & Factors

TPCA - Eigenvalue	Proportion	Cumulative
2.758	68.96%	68.96%
1.000	25.00%	93.96%
0.242	6.04%	100.00%
0.000	0.00%	100.00%
4	100.00%	100.00%

Correlations Variables	Factors F1	F2
Debt	0.645	0.707
Deficit	0.982	0.000
Revenues	0.645	-0.707
Expenditures	0.982	0.000

PCA - Eigenvalue	Proportion	Cumulative
2.224	55.61%	55.61%
1.171	29.27%	84.88%
0.605	15.12%	100.00%
0.000	0.00%	100.00%
4	100.00%	100.00%

Correlations Variables	Factors F1	F2
Debt	-0.615	0.497
Deficit	0.307	-0.845
Revenues	-0.907	-0.414
Expenditures	-0.964	-0.196

$$V_{u_{Euclidean}} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

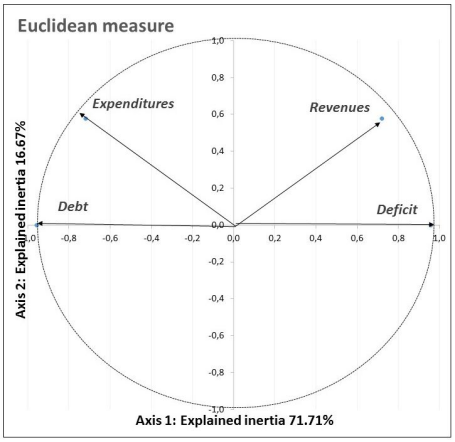


Fig. 7. Euclidean TPCA - Adjacency matrix - The public finance variables on the first principal plane

We can represent the topological analysis of each of the 15 proximity measures considered, for example see the Euclidean TPCA in Figure 3. One can moreover give Figure 8, the graphical representation associated with a perfect no correlation between variables, from the identity adjacency matrix.

4 Conclusion

This research work proposes a new approach that allows to synthesize and describe a set of quantitative variables in a topological context. Like PCA, the proposed TPCA is a multidimensional topological exploratory method that can be useful for dimension reduction, it enriches the conventional quantitative data analysis methods. Future work involves extending this topological approach to synthesize the relations existing between a set of mixed (quantitatives & qualitatives) variables - Topological Mixed Principal Component Analysis -

$$V_{u_\circ} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

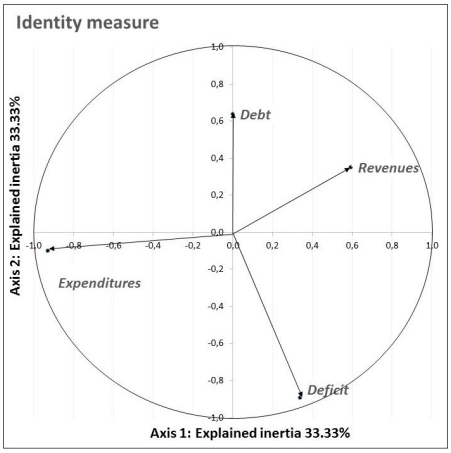


Fig. 8. Adjacency identity matrix - Total absence of correlation between variables

TMPCA, between two groups of continuous variables - Topological Canonical Analysis - TCA and between several multidimensional data tables - Topological Analysis of Evolutionary Data - TAED.

5 Appendix

Table 7. Some proximity measures for continuous data

Measures	Formula : Distance - Dissimilarity
Euclidean	$u_{Euc}(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$
Manhattan	$u_{Man}(x,y) = \sum_{j=1}^{p} \lvert x_j - y_j \rvert$
Minkowski	$u_{Min_\gamma}(x,y) = (\sum_{j=1}^{p} \lvert x_j - y_j \rvert^\gamma)^{\frac{1}{\gamma}}$
Tchebychev	$u_{Tch}(x,y) = \max_{1\leq j\leq p} \lvert x_j - y_j \rvert$
Normalized Euclidean	$u_{NE}(x,y) = \sqrt{\sum_{j=1}^{p} \frac{1}{\sigma_j^2}[(x_j - \overline{x}_j) - (y_j - \overline{y}_j)]^2}$
Cosine dissimilarity	$u_{Cos}(x,y) = 1 - \frac{\sum_{j=1}^{p} x_j y_j}{\sqrt{\sum_{j=1}^{p} x_j^2}\sqrt{\sum_{j=1}^{p} y_j^2}} = 1 - \frac{<x,y>}{\|x\|\|y\|}$
Canberra	$u_{Can}(x,y) = \sum_{j=1}^{p} \frac{\lvert x_j - y_j \rvert}{\lvert x_j \rvert + \lvert y_j \rvert}$
Pearson Correlation	$u_{Cor}(x,y) = 1 - \frac{(\sum_{j=1}^{p}(x_j - \overline{x})(y_j - \overline{y}))^2}{\sum_{j=1}^{p}(x_j - \overline{x})^2 \sum_{j=1}^{p}(y_j - \overline{y})^2} = 1 - \frac{(<x-\overline{x}, y-\overline{y}>)^2}{\|x-\overline{x}\|^2 \|y-\overline{y}\|^2}$
Squared Chord	$u_{Cho}(x,y) = \sum_{j=1}^{p}(\sqrt{x_j} - \sqrt{y_j})^2$
Doverlap measure	$u_{Dev}(x,y) = max(\sum_{j=1}^{p} x_j, \sum_{j=1}^{p} y_j) - \sum_{j=1}^{p} min(x_j, y_j)$
Weighted Euclidean	$u_{WEu}(x,y) = \sqrt{\sum_{j=1}^{p} \alpha_j (x_j - y_j)^2}$
Gower's Dissimilarity	$u_{Gow}(x,y) = \frac{1}{p} \sum_{j=1}^{p} \mid x_j - y_j \mid$
Shape Distance	$u_{Sha}(x,y) = \sqrt{\sum_{j=1}^{p}[(x_j - \overline{x}_j) - (y_j - \overline{y}_j)]^2}$
Size Distance	$u_{Siz}(x,y) = \mid \sum_{j=1}^{p}(x_j - y_j) \mid$
Lpower	$u_{Lpo_\gamma}(x,y) = \sum_{j=1}^{p} \lvert x_j - y_j \rvert^\gamma$

Where, p is the dimension of space, $x = (x_j)_{j=1,\ldots,p}$ and $y = (y_j)_{j=1,\ldots,p}$ two points in R^p, $\overline{x}_j$ the mean, σ_j the Standard deviation, $\alpha_j = \frac{1}{\sigma_j^2}$ and $\gamma > 0$.

Table 8. Similarities $S(V_{u_i}, V_{u_j})$ & Fisher's Exact Test p-values

Measure	Euclidean	Manhattan	Minkovski	Cosine dissimilarity	Pearson Correlation	Squared Chord	Doverlap measure	Gower's Dissimilarity	Shape Distance	Size Distance	Lpower	Tchebytchev	Normalized Euclidean	Canberra	Weighted Euclidean	Measure
$u_\star$ measure	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.500	0.033	0.033	0.033	Weighted Euclidean
$u_{Euclidean}$	1	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.183	0.183	0.183	0.183	Canberra
$u_{Manhattan}$	1	1	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.183	0.183	0.183	0.183	Normalized Euclidean
$u_{Minkovski}$	1	1	1	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.183	0.183	0.183	0.183	Tchebytchev
u_{Cosine}	1	1	1	1	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.183	0.183	0.183	0.183	Lpower
$u_{Pearson}$	1	1	1	1	1	0.008	0.008	0.008	0.008	0.008	0.008	0.183	0.183	0.183	0.183	Size Distance
u_{Chord}	1	1	1	1	1	1	0.008	0.008	0.008	0.008	0.008	0.183	0.183	0.183	0.183	Shape Distance
u_{Dover}	1	1	1	1	1	1	1	0.008	0.008	0.008	0.008	0.183	0.183	0.183	0.183	Gower's Dissimilarity
u_{Gower}	1	1	1	1	1	1	1	1	0.008	0.008	0.008	0.183	0.183	0.183	0.183	Doverlap measure
u_{Shape}	1	1	1	1	1	1	1	1	1	0.008	0.008	0.183	0.183	0.183	0.183	Squared Chord
u_{Size}	1	1	1	1	1	1	1	1	1	1	0.008	0.183	0.183	0.183	0.183	Pearson Correlation
u_{Lpowe}	1	1	1	1	1	1	1	1	1	1	1	0.183	0.183	0.183	0.183	Cosine dissimilarity
$u_{Tchebytch}$	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	1	1	1	1	Minkovski
$u_{Neuclidean}$	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.500	1	0.008	1	Manhattan
$u_{Canbera}$	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.500	1	1	0.008	Euclidean
$u_{Weuclidean}$	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.500	1	1	1	
$u_\star$ measure	0.875	0.875	0.875	0.625	0.875	0.875	0.875	0.875	0.875	0.875	0.875	0.875	0.875	0.875	0.875	
Measure	Euclidean	Manhattan	Minkovski	Cosine dissimilarity	Pearson Correlation	Squared Chord	Doverlap measure	Gower's Dissimilarity	Shape Distance	Size Distance	Lpower	Tchebytchev	Normalized Euclidean	Canberra	Weighted Euclidean	

Similarity: $S(u_{Tchebytchev}\ ,\ u_{Euclidean}) = 75\%$.
Fisher's Exact Test: $p - value_{(u_{Euclidean}\ ,\ u_{Tchebytchev})} = 0.1833 > \alpha = 5\%$: not significant.

References

1. Abdesselam, R.: A Topological Multiple Correspondence Analysis. *Journal of Mathematics and Statistical Science*, Science Signpost Publishing Inc., USA, Vol.5, Issue 8, 175–192, 2019.
2. Abdesselam, R.: Selection of proximity measures for a Topological Correspondence Analysis. *In a Book Series*, 5th Stochastic Modeling Techniques and Data Analysis, International Conference, Chania, Greece, C.H. Skiadas (Ed), 11–24, 2018.
3. Abdesselam, R.: A Topological Discriminant Analysis. *In book Chapter, Volume 3, Data Analysis and Applications 2: Utilization of Results in Europe and 0ther Topics*, J. Bozeman and C. Skiadas Editors, ISTE Science Publishing, Wiley, 167–178, 2018.
4. Batagelj, V., Bren, M.: Comparing resemblance measures. *In Journal of classification*, 12, 73–90, 1995.
5. Cailliez, F. and Pagès, J.P.: Introduction à l'Analyse des données", *S.M.A.S.H.*, Paris, 1976.
6. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol 20, 27–46, 1960.
7. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. *The journal of Machine Learning Research*, Vol. 7, 1–30, 2006.
8. Eurostat, Data source: Government finance statistics - Statistics explained, *https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Government finance statistics* 1–15, 2018.
9. Fisher, R-A.: The Interpretation of χ^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, Published by Wiley, 85, 1, 87–94, 1922.
10. Hotelling H.: Analysis of a Complex of Statistical Variables into Principal Components. *In Journal of Educational Psychology*, vol. 24, 417–441, 1933.
11. Kim, J.H. and Lee, S.: Tail bound for the minimal spanning tree of a complete graph. *In Statistics & Probability Letters*, 4, 64, 425–430, 2003.
12. Lebart, L.: Stratégies du traitement des données d'enquêtes. *La Revue de MODULAD*, 3, 21–29, 1989.
13. Lesot, M. J., Rifqi, M. and Benhadda,H.: Similarity measures for binary and numerical data: a survey. *In IJKESDP*, 1, 1, 63-84, 2009.
14. Mantel, N.: A technique of disease clustering and a generalized regression approach. *In Cancer Research*, 27, 209–220, 1967.
15. Park, J. C.,Shin, H. and Choi, B. K.: Elliptic Gabriel graph for finding neighbors in a point set and its application to normal vector estimation. *In Computer-Aided Design Elsevier*, 38, 6, 619–626, 2006.
16. Pearson K.: On lines and Planes of Closest Fit to Systems of Points in Space. *In Philosophical Magazine*, vol. 2, 11, 559–572, 1901.
17. Rifqi, M., Detyniecki, M. and Bouchon-Meunier, B.: Discrimination power of measures of resemblance. *IFSA'03 Citeseer*, 2003.
18. Saporta, G.: Probabilités, analyse des données et Statistique, *Editions TECHNIP*, 2011.
19. Schneider, J. W. and Borlund, P.: Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *In Journal of the American Society for Information Science and Technology*, 58, 11, 1586–1595, 2007.
20. Schneider, J. W. and Borlund, P.: Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by use of the Mantel and Procrustes statistics. *In Journal of the American Society for Information Science and Technology*, 11, 58, 1596–1609, 2007.

21. Toussaint, G. T.: The relative neighbourhood graph of a finite planar set. *In Pattern recognition*, 12, 4, 261–268, 1980.
22. Ward, J. R.: Hierarchical grouping to optimize an objective function. *In Journal of the American statistical association JSTOR*, 58, 301, 236–244, 1963.
23. Warrens, M. J.: Bounds of resemblance measures for binary (presence/absence) variables. In Journal of Classification, *Springer*, 25, 2, 195–208, 2008.
24. Zighed, D., Abdesselam, R., and Hadgu, A.: Topological comparisons of proximity measures. *In the 16th PAKDD 2012 Conference*. In P.-N. Tan et al., Eds. Part I, LNAI 7301, Springer-Verlag Berlin Heidelberg, 379–391, 2012.