A Topological Approach of Principal Component Analysis

Rafik Abdesselam

International Journal of Data Science and Analysis
2021; 7(2): 20-31
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20210702.11
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

A Topological Approach of Principal Component Analysis

Rafik Abdesselam

Department of Economics and Management, University Lumière of Lyon 2, Lyon, France

Email address:
rafik.abdesselam@univ-lyon2.fr

To cite this article:
Rafik Abdesselam. A Topological Approach of Principal Component Analysis. *International Journal of Data Science and Analysis*. Vol. 7, No. 2, 2021, pp. 20-31. doi: 10.11648/j.ijdsa.20210702.11

Received: January 10, 2021; **Accepted**: January 29, 2021; **Published**: April 20, 2021

Abstract: Large datasets are increasingly widespread in many disciplines. The exponential growth of data requires the development of more data analysis methods in order to process information more efficiently. In order to better visualize the data, many methods such as Principal Component Analysis (PCA) and MultiDimensional Scaling (MDS) allow to extract a low-dimensional structure from high-dimensional data set. The proposed approach, called 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. It is a topological method of data analysis that consists of comparing and classifying proximity measures from among some of the most widely used proximity measures for continuous data. Proximity measures play an important role in many areas of data analysis, the results strongly depend on the proximity measure chosen. So, among the many existing measures, which one is most useful? Are they all equivalent? How to identify the one that is most appropriate to analyze the correlation structure of a set of quantitative variables. TPCA proposes an appropriate adjacency matrix associated to an unknown proximity measure according to the data under consideration, then analyzes and visualizes, with graphic representations, the relationship structure of the variables relating to, the well known PCA problem. Its uses the concept of neighborhood graphs and compares a set of proximity measures for continuous data which can be more-or-less equivalent a topological equivalence criterion between two proximity measures is defined and statistically tested according to the topological correlation between the variables considered. An example on real data illustrates the proposed approach.

Keywords: Proximity Measure, Neighborhood Graph, Adjacency Matrix, Topological Equivalence, Correlation Matrix, MDS Graphical Representation

1. Introduction

Choosing a proximity measure from among the many available measures greatly influences the results of any data analysis method, moreover, these measures are more-or-less equivalent according to the concept of the neighborhood graph structure used.

A topological equivalence criterion is defined between proximity measures from the topological structure induced by each measure.

Large datasets are increasingly common and are often difficult to interpret. Principal component analysis (PCA) [16, 10, 5, 18] is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It is an exploratory tool for continuous data.

PCA is an adaptive technique for continuous data, variants of this technique have been developed and tailored to various different data types and structures.

In order to suitably interpret the large datasets, methods are needed are required to drastically reduce their dimensionality in an interpretable way. Many techniques have been developed for this purpose, but PCA is one of the oldest and most widely used. 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 so–called 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 the context of artificial intelligence, we often compare

des situations représented by a set of objects, for this, we must choose and specify the proximity measure between objects. The study context, the data type and other factors can help us to choose the proximity measure that might be suitable. However, the number of possible measures can still quite large.

Moreover, these measures which are still possible, are they all equivalent? Is there a measure more specific or more suited than another for the study considered? In information retrieval, the choose of proximity measure is an essential issue on which the results depend.

The present study proposes a new framework for comparing proximity measures in order to identify those that are similar, thus, we will no longer need to try all measures.

These comparisons are clarified by a proximity measure which evaluates the similarity or dissimilarity between two objects within a set. The proximity measure have mathematical properties and well specific axioms.

The best measure is selected according to the correlation structure of the set of quantitative variables to synthesize, the aim is to establish a topological PCA. The results of TPCA are different according to the selected proximity measure.

Several authors have studied the topological equivalence of proximity measures, in a general framework [4, 17, 13, 24], in the context of the discriminant analysis [3] and the correspondence analysis [2, 1], but none in the context of PCA. So, in this paper, we show how to built the appropriate adjacency matrix, induced by an unknown proximity measure but which takes in to account the correlation structure of the variables that we want to describe topologically.

In this article, 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 8 in Appendix summarizes some classic proximity measures used for continuous data [23], we give on R^n the definition of 15 of them.

We assume that we have at our disposal $\{x^k; k=1, \ldots, p\}$ 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

The notion of topological equivalence between two proximity measures is based on the concept of the neighborhood graph. Two measures are said topologically equivalent if their graphs induced on the set of objects remain identical. Measuring the similarity between two proximity measures amounts to measuring the similarity of their neighborhood graphs.

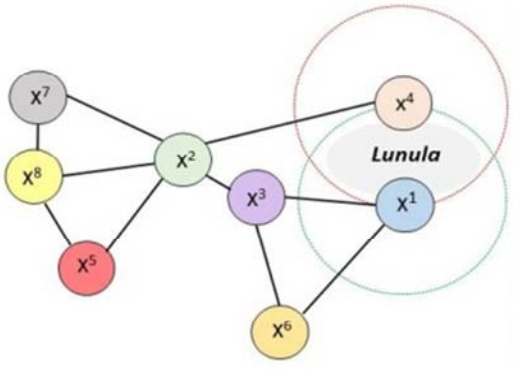
Consider a set $E=\{x^1, x^2, \ldots, x^k, \ldots, x^p\}$ of p objects in R^n, associated with the p quantitative variables.

Given a proximity measure noted u, we can define a neighborhood binary relationship on $E \times E$ noted Vu. Thus, we can build a neighborhood graph on a set of objects-variables, where the vertices are the variables and the edges are defined from the property of the neighborhood relationship. It is a binary symmetric matrix.

Many graph definitions are possible to build this binary matrix. 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].

So, given a proximity measure u, we can associated the adjacency matrix Vu of order p, where all pairs (x^k, x^l) of neighboring variables satisfy the following RNG expression:

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



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

Figure 1. *RNG example with eight variables - Aadjacency matrix.*

This means that if two variables x^k and xl 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 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_{ui} and V_{uj} are different or not. The degree of topological equivalence between two proximity measures is measured by the following definition of concordance. The topological equivalence between two adjacency matrices satisfy the following expression:

$$S(Vu_i, Vu_j) = \frac{1}{p^2}\sum_{k=1}^{p}\sum_{l=1}^{p}\delta_{kl}(x^k, x^l)$$

$$with\ \delta_{kl}(x^k, x^l) = \begin{cases} 1 & if\ Vu_i(x^k, x^l) = Vu_j(x^k, x^l) \\ 0 & otherwise \end{cases}$$

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 independence position.

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 correlation between the two variables.

We construct the adjacency matrix denoted by V_{u*}, 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. The adjacency matrix V_{u*} associated to reference measure u_* satisfy the following expression:

$$V_{u*}(x^k, x^l) = \begin{cases} 1 & if\ p - value = P\left[\ |T_{n-2}| > t - value\right] \leq \alpha \\ & \forall k = 1, p, \forall l = 1, p \\ 0 & otherwise \end{cases}$$

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: $t=\sqrt{(n - 2)\ (1 - r^2)}$ 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*}, is associated with an unknown proximity measure denoted u_* and called a reference measure. Thus, with this reference proximity measure we can establish (V_{ui}, V_{u*}), the topological equivalence between the two proximity measures u_i and u_*, by measuring the percentage of similarity between the adjacency matrix V_{ui} and the reference adjacency matrix V_{u*}.

In order to graphically describe the similarities between proximity measures, we can for example apply the notion of them scope [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: $[D]_{ij}=D\ (V_{ui}, V_{uj})=1 - S\ (V_{ui}, V_{uj})$ to partition them into homogeneous groups and to view their similarities.

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_*, we project the latter as a supplementary element into the two data analysis methods, positioned by the dissimilarity vector with 15 components $[D]_{*I}=1 - S\ (V_{u*}, V_{ui})$.

2.2. Statistical Comparisons Between 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 Demsar [7] and Schneider & Borlund [19, 20] 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_{ui} and V_{uj} 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 p (p + 1)/2 values, the p diagonal values and the 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 p (p + 1)/2.

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

2.3. Graphical Representations - Variables & 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 × 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 carry out the classical MDS [5], namely factorial analysis on similarity V_{u*} or dissimilarity $D_{u*}=U - V_{u*}$ table, where $U=1_p\,{}^t1_p$ is the p×p matrix of 1s and 1_p denotes the p indicator vector of 1s.

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

The TPCA can be performed from any adjacency matrix V_{ui} 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	N	Mean	Std. Dev.	Coef. Var. (%)	Min	Max
Debt	28	68.04	36.5	53.70	8.7	176.1
Deficit	28	-0.26	1.7	640.07	-3.1	3.5
Revenues	28	42.58	6.7	15.63	26.0	53.8
Expenditures	28	42.85	6.8	15.85	26.3	56.5

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], aggregation based on the criterion of the loss of minimal inertia, was used in order to characterize classes of proximity measure relative to their similarities. The reference measure u_* 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.

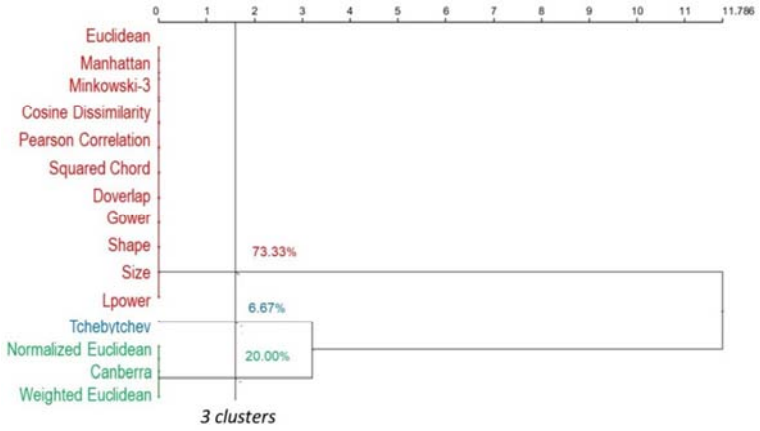


Figure 2. *Hierarchical tree of the proximity measures.*

Table 2. *Clusters composition - Assignment of the reference measure.*

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

Table 3. *Similarities and Fisher's Exact Test.*

u_i	u_j	S (u_i, u_j)	p-value
Cluster 1	Cluster 1	1.000	0.0083**
Cluster 1	Cluster 2	0.750	0.1833
Cluster 1	Cluster 3	0.750	0.1833
Cluster 2	Cluster 2	1.000	0.0083**
Cluster 2	Cluster 3	0.500	1.0000
Cluster 3	Cluster 3	1.000	0.0083**
u_*	Cluster 1	0.750	0.1833
u_*	Cluster 2	0.625	0.5000
u_*	Cluster 3	0.875	0.0333*

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

Cluster 2	**Cluster 1**	**Euclidean**	**Reference**	**Cluster 1**	**Euclidean**
Tchebychev	$V_{u1}=0$	$V_{u1}=1$	Measure	$V_{u1}=0$	$V_{u1}=1$
$V_{u2}=0$	2	1	$V_{u*}=0$	3	1
$V_{u2}=1$	1	6	$V_{u*}=0$	0	6
S (V_{u2}, V_{u1})=75%; p-value=0.183			S (V_{u*}, V_{u1})=75%; p-value=0.183		

Cluster 3	**Cluster 2**	**Tchebytchev**	**Reference**	**Cluster 2**	**Tchebytchev**
Canberra	$V_{u1}=0$	$V_{u1}=1$	Measure	$V_{u1}=0$	$V_{u1}=1$
$V_{u2}=0$	1	2	$V_{u*}=0$	2	2
$V_{u2}=1$	2	5	$V_{u*}=1$	1	5
S (V_{u3}, V_{u2})=50%; p-value=1.000			S (V_{u*}, V_{u2})=62.50%; p-value=0.500		

Cluster 1	**Cluster 3**	**Canberra**	**Reference**	**Cluster 3**	**Canberra**
Euclidean	$V_{u1}=0$	$V_{u1}=1$	Measure	$V_{u1}=0$	$V_{u1}=1$
$V_{u2}=0$	2	1	$V_{u*}=0$	3	1
$V_{u2}=1$	1	6	$V_{u*}=1$	0	6
S (V_{u1}, V_{u3})=75%; p-value=0.183			S (V_{u*}, V_{u3})=87.50%; p-value=0.033*		

Significance level α; **$\alpha \leq 1\%$; *$\alpha \in [1\%; 5\%]$

Moreover, in view of the results in Table 2, the reference measure u_* 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.

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 all the C^2_{15}=105 pairs of proximity measures formed with the 15 measures considered and the 15 pairs formed with the unknown reference measure u_*. The values below the diagonal correspond to the similarities S (V_{ui}, V_{uj}) 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_{ui}, V_{uj})=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 cluster of the partition presented in Table 2. The Table 4 illustrates the contingency tables 2 × 2 between the measures of each cluster: Euclidean, Tchebytchev, Canberra and reference measure u_* for the calculation of Fisher's exact test.

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

Table 5. *Pearson correlation matrix (p-value).*

Variable	Debt	Deficit	Revenues	Expenditures
Debt	1.000			
Deficit	-0.340 (0.08)	1.000		
Revenues	0.307 (0.11)	0.039 (0.843)	1.000	
Expenditures	0.385 (0.04*)	-0.209 (0.255)	0.969 (0.001**)	1.000

Significance level α; **α ≤ 1%; *α ∈ [1%; 5%]

The adjacency matrix Vu* associated to the adapted proximity measure u* to the considered data, is build from the correlations matrix Table 5. Figure 5 shows on the main first TPCA plane, the topological correlation between the Government finance variables.

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

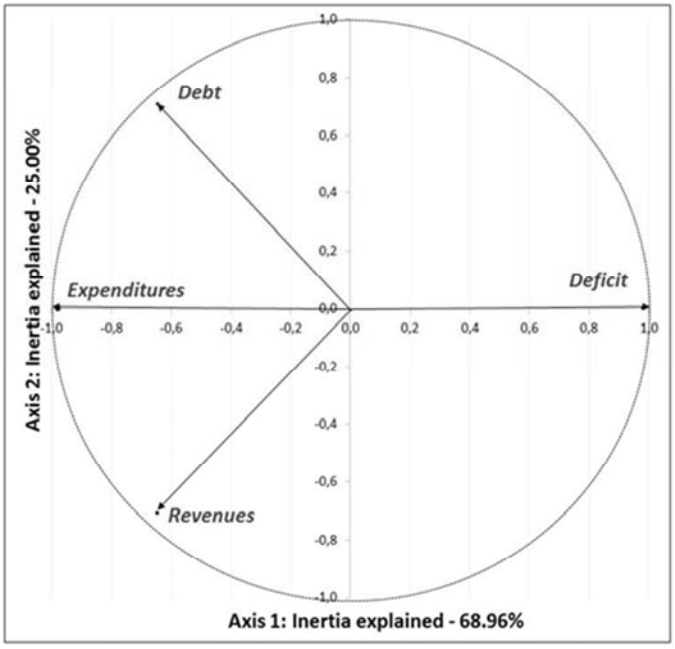


Figure 3. *TPCA - Adjacency matrix – The public finance variables on the first principal plane.*

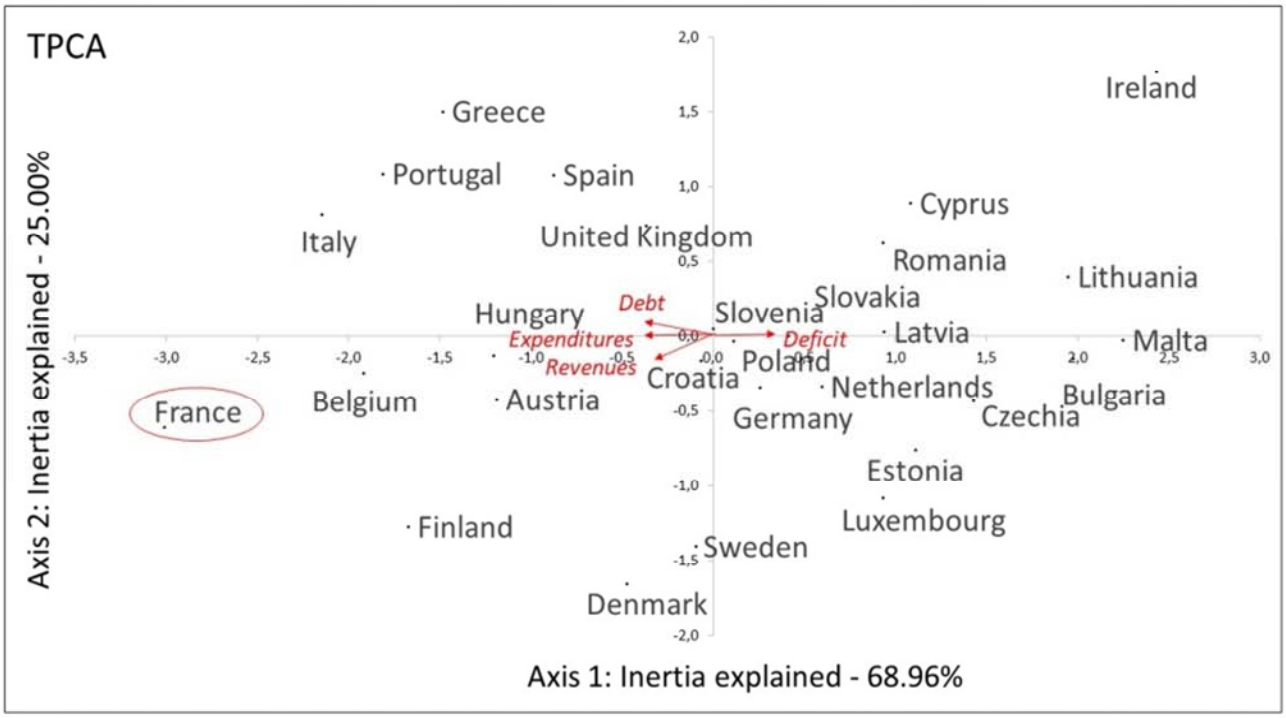


Figure 4. *TPCA - The EU-28 countries on the first principal plane.*

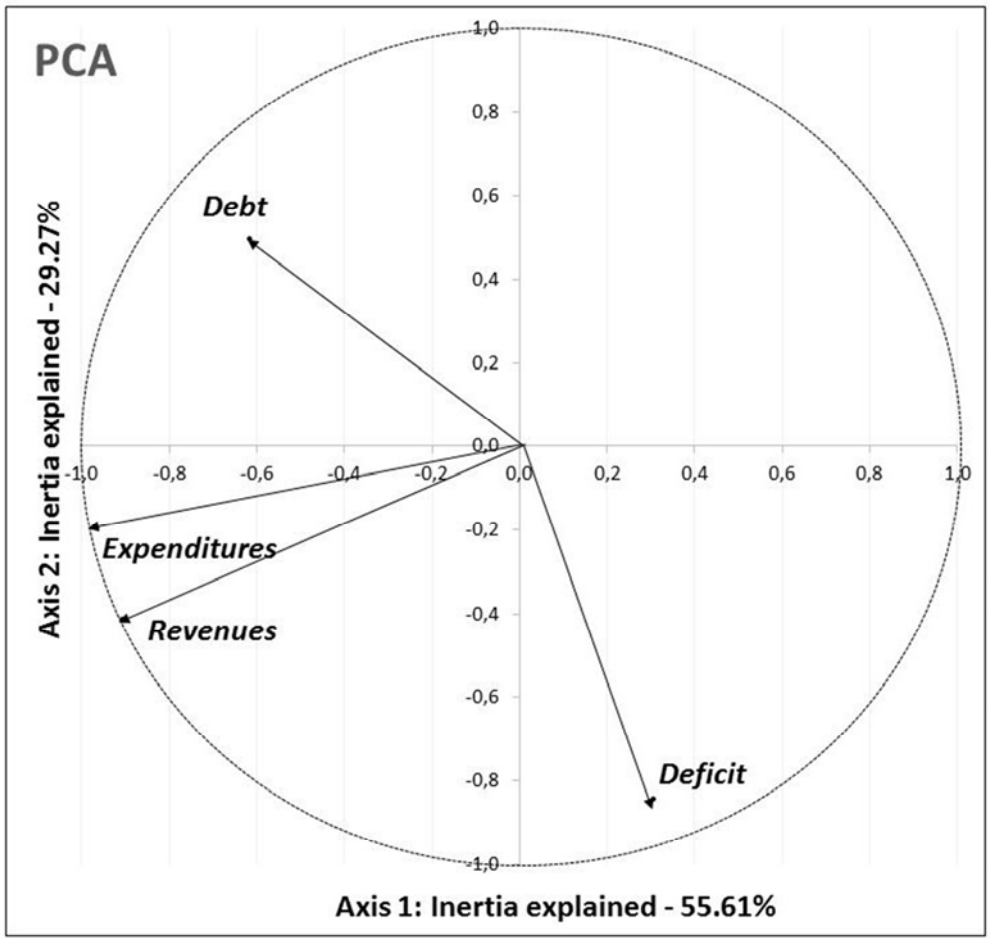


Figure 5. *PCA – The public finance variables on the first principal plane.*

The adjacency matrix V_{u*} associated to the adapted proximity measure u_* to the considered data, is build from the correlation matrix Table 5. Figure 5 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 - 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. Effectively, 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.

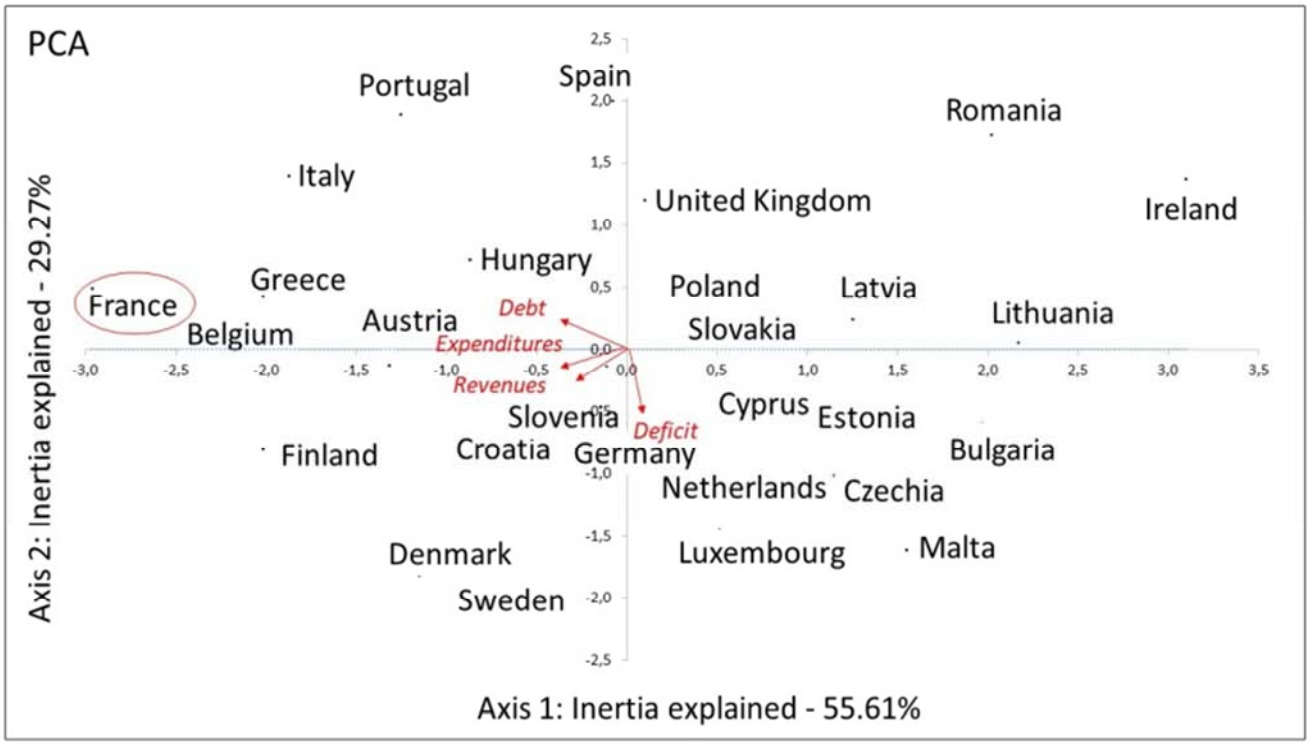


Figure 6. *PCA - The EU-28 countries on the first principal plane.*

Table 6. *TPCA and PCA eigenvalues and correlations Variables & Factors.*

TPCA	Eigenvalue	Proportion	Cumulative	Correlation	Factors	
	2.758	68.96%	68.96%	Variables	F1	F2
	1.000	25.00%	93.96%	Debt	0.645	0.707
	0.242	6.04%	100.00%	Deficit	0.982	0.000
	0.000	0.00%	100.00%	Revenues	0.645	-0.707
	4	100.00%	100.00%	Expenditures	0.982	0.000

PCA	Eigenvalue	Proportion	Cumulative	Correlation	Factors	
	2.224	55.61%	55.61%	Variables	F1	F2
	1.171	29.27%	84.88%	Debt	-0.615	0.497
	0.605	15.12%	100.00%	Deficit	0.307	-0.845
	0.000	0.00%	100.00%	Revenues	-0.907	-0.414
	4	100.00%	100.00%	Expenditures	-0.964	-0.196

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.

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, it's characterized by high Debts, high Expenditures and high Revenues, but the Deficit does not characterize the first factorial axis of the PCA.

We can represent the topological analysis of each of the 15 proximity measures considered, for example see the Euclidean TPCA in Figure 7. One can moreover give Figure 8, the graphical representation associated with a perfect no correlation between variables, from the identity adjacency matrix.

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

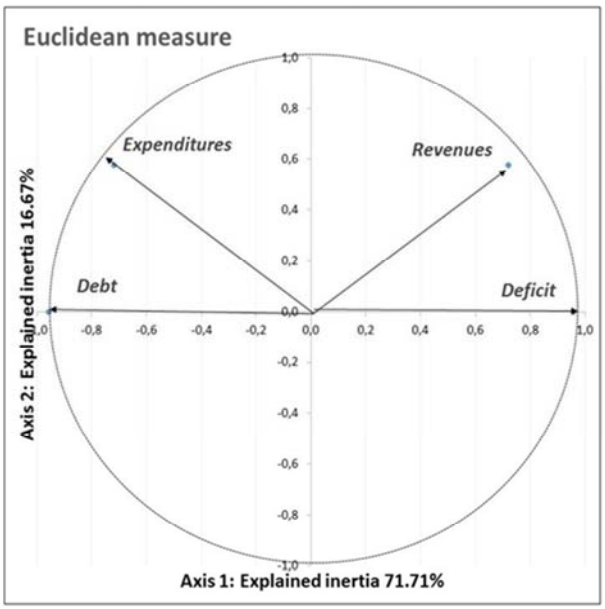


Figure 7. *Euclidean TPCA - The public finance variables on the first principal plane.*

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

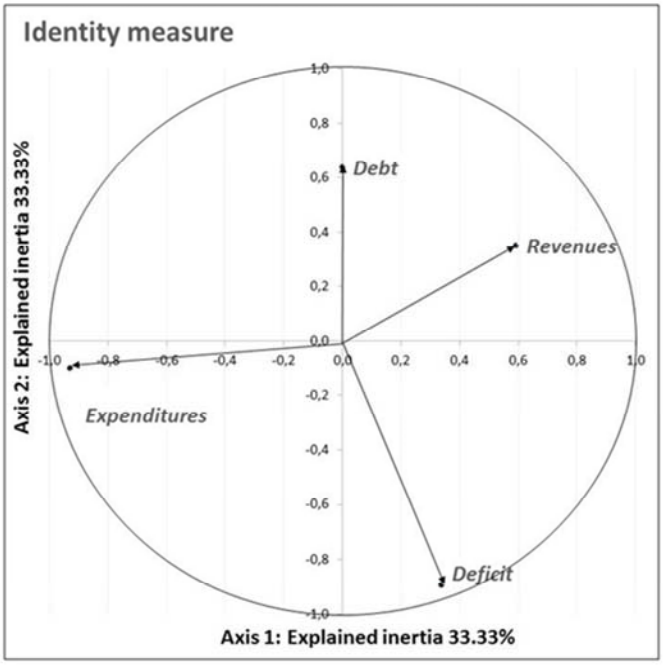


Figure 8. *Identity TPCA - The public finance variables on the first principal plane.*

4. Conclusion

This research work proposes a new approach that allows to synthesize and describe the correlation structure of 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 and information redundancy in a data set, it enriches the conventional quantitative data analysis methods. Future work involves extending this topological approach in three directions, to synthesize the relations existing between a set of a mixture of qualitative and quantitative variables, between two sets of continuous variables in the context of canonical analysis and also between several multidimensional data tables in the context of evolutionary data analysis.

Appendix

Table 7. *Similarities & Fisher's Exact Test p-values.*

	Euclidean	Manhattan	Minkowski-3	Dissimilarity	Correlation	Squared Chord	Doverlap	Gower
u* measure	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033
Euclidean	1	0.008	0.008	0.008	0.008	0.008	0.008	0.008
Manhattan	1	1	0.008	0.008	0.008	0.008	0.008	0.008
Minkowski-3	1	1	1	0.008	0.008	0.008	0.008	0.008
Dissimilarity	1	1	1	1	0.008	0.008	0.008	0.008
Correlation	1	1	1	1	1	0.008	0.008	0.008
Squared Chord	1	1	1	1	1	1	0.008	0.008
Doverlap	1	1	1	1	1	1	1	0.008
Gower	1	1	1	1	1	1	1	1
Shape	1	1	1	1	1	1	1	1
Size	1	1	1	1	1	1	1	1
Lpower	1	1	1	1	1	1	1	1
Tchebytchev	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
N. Euclidean	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
Canberra	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
W. Euclidean	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
u* measure	0.875	0.875	0.875	0.625	0.875	0.875	0.875	0.875

Table 7. *Continued.*

	Shape	Size	Lpower	Tchebytchev	N. Euclidean	Canberra	W. Euclidean
u* measure	0.033	0.033	0.033	0.5	0.033	0.033	0.033
Euclidean	0.008	0.008	0.008	0.183	0.183	0.183	0.183
Manhattan	0.008	0.008	0.008	0.183	0.183	0.183	0.183
Minkowski-3	0.008	0.008	0.008	0.183	0.183	0.183	0.183
Dissimilarity	0.008	0.008	0.008	0.183	0.183	0.183	0.183
Correlation	0.008	0.008	0.008	0.183	0.183	0.183	0.183
Squared Chord	0.008	0.008	0.008	0.183	0.183	0.183	0.183
Doverlap	0.008	0.008	0.008	0.183	0.183	0.183	0.183
Gower	0.008	0.008	0.008	0.183	0.183	0.183	0.183
Shape	1	0.008	0.008	0.183	0.183	0.183	0.183
Size	1	1	0.008	0.183	0.183	0.183	0.183
Lpower	1	1	1	0.183	0.183	0.183	0.183
Tchebytchev	0.75	0.75	0.75	1	1	1	1
N. Euclidean	0.75	0.75	0.75	0.5	1	0.008	0.008
Canberra	0.75	0.75	0.75	0.5	1	1	0.008
W. Euclidean	0.75	0.75	0.75	0.5	1	1	1
u* measure	0.875	0.875	0.875	0.875	0.875	0.875	0.875

Similarity: S (Tchebytchev; Euclidean)=75%.
Fisher's Exact Test: p-value (Euclidean; Tchebytchev)=0.183 > α=5%: not significant.

Table 8. *Some proximity measures for continuous data.*

Measure	Formula: Distance - Dissimilarity
Euclidean	$u_{Euclidean}(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$
Manhattan	$u_{Manhattan}(x,y) = \sum_{j=1}^{p}\lvert x_j - y_j\rvert$
Minkowski-ν	$u_{Minkowski}(x,y) = \left(\sum_{j=1}^{p}\lvert x_j - y_j\rvert^{\nu}\right)^{1/\nu}$
Cosine Dissimilarity	$u_{Cosine}(x,y) = 1 - \frac{\sum_{j=1}^{p} x_j\, y_j}{\sqrt{\sum_{j=1}^{p} {x_j}^2}\sqrt{\sum_{j=1}^{p} {y_j}^2}}$
Pearson Correlation	
Squared Chord	$u_{Chord}(x,y) = \sum_{j=1}^{p}(\sqrt{x_j} - \sqrt{y_j})^2$
Doverlap measure	$u_{Doverlap}(x,y) = \max\left(\sum_{j=1}^{p} x_j, \sum_{j=1}^{p} y_j\right) - \sum_{j=1}^{p}\min(x_j, y_j$
Gower	$u_{Gower}(x,y) = \frac{1}{p}\sum_{j=1}^{p}\lvert x_j - y_j\rvert$
Shape Distance	$u_{\text{Shape}}(x,y) = \sqrt{\sum_{j=1}^{p}\left[(x_j - \bar{x}_j) - (y_j - \bar{y}_j)\right]^2}$
Size Distance	$u_{Size}(x,y) = \left\lvert\sum_{j=1}^{p}(x_j - y_j)\right\rvert$
Lpower	$u_{Lpower}(x,y) = \sum_{j=1}^{p}\lvert x_j - y_j\rvert^{\nu}$
Tchebytchev	$u_{Tchebychev}(x,y) = max_{1\le j\le p}\lvert x_j - y_j\rvert$
Normalized Euclidean	$u_{\text{NEuclidean}}(x,y) = \sqrt{\sum_{j=1}^{p}\frac{1}{\sigma^2}\left[(x_j - \bar{x}_j) - (y_j - \bar{y}_j)\right]^2}$
Canberra	$u_{Canberra}(x,y) = \sum_{j=1}^{p}\frac{\lvert x_j - y_j\rvert}{\lvert x_j\rvert + \lvert y_j\rvert}$
Weighted Euclidean	$u_{\text{WEuclidean}}(x,y) = \sqrt{\sum_{j=1}^{p}\alpha_j(x_j - y_j)^2}$

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, x_j the mean, σ_j the Standard deviation, $\alpha_j=1\backslash\sigma_j^2$ and $\nu > 0$.

References

[1] R. Abdesselam, "A Topological Multiple Correspondence Analysis." Journal of Mathematics and Statistical Science, Science Signpost Publishing Inc., USA, Vol. 5, Issue 8, pp. 175-192, 2019.

[2] R. Abdesselam, "Selection of proximity measures for a Topological Correspondence Analysis." In a Book Series, 5th Stochastic Modeling Techniques and Data Analysis, International Conference, Chania, Greece, pp. 11-24, 2018.

[3] R. Abdesselam, "A Topological Discriminant Analysis." In book Chapter, Vol. 3, Data Analysis and Applications 2: Utilization of Results in Europe and 0ther Topics, ISTE Science Publishing, Wiley, pp. 167-178, 2018.

[4] V. Batagelj and M. Bren, "Comparing resemblance measures." In Journal of classification, 12, pp. 73-90, 1995.

[5] F. Cailliez and J-P Pagès "Introduction à l'Analyse des données.", S. M. A. S. H., Paris, 1976.

[6] J. Cohen, "A coefficient of agreement for nominal scales." Educational and Psychological Measurement, Vol. 20, pp. 27-46, 1960.

[7] J. Demsar, "Statistical comparisons of classifiers over multiple data sets." The journal of Machine Learning Research, Vol. 7, pp. 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, pp. 1-15, 2018.

[9] R-A. Fisher, "The Interpretation of chi2 from Contingency Tables, and the Calculation of P." Journal of the Royal Statistical Society, Published by Wiley, 85, 1, pp. 87-94, 1922.

[10] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components". Journal of Educational Psychology, Vol. 24, 6, pp. 417-441, 1933.

[11] J. H. Kim and S. Lee, "Tail bound for the minimal spanning tree of a complete graph." In Statistics & Probability Letters, Vol. 4, 64, pp. 425-430, 2003.

[12] L. Lebart, "Stratégies du traitement des données d'enquêtes." La Revue de MODULAD, 3, pp. 21-29, 1989.

[13] J. Lesot, M. Rifqi and H. Benhadda, "Similarity measures for binary and numerical data: a survey." In IJKESDP, Vol. 1, 1, pp. 63-84, 2009.

[14] N. Mantel, “A technique of disease clustering and a generalized regression approach.” In Cancer Research, Vol. 27, pp. 209-220, 1967.

[15] J-C. Park, H. Shin, and B-K. Choi, “Elliptic Gabriel graph for finding neighbors in a point set and its application to normal vector estimation.” In Computer-Aided Design Elsevier, Vol. 38, 6, pp. 619-626, 2006.

[16] K. Pearson, “On lines and Planes of Closest Fit to Systems of Points in Space.” In Philosophical Magazine, vol. 2, 11, pp. 559-572, 1901.

[17] J. Rifqi, M., Detyniecki, M. and Bouchon-Meunier, B. “Discrimination power of measures of resemblance.” IFSA'03 Citeseer, 2003.

[18] G. Saporta, “Probabilités, analyse des données et Statistique.” Editions TECHNIP, 2011.

[19] J-W. Schneider and P. Borlund, “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, Vol. 58, 11, pp. 1586-1595, 2007.

[20] J-W. Schneider and P. Borlund, “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, Vol. 11, 58, pp. 1596-1609, 2007.

[21] G-T. Toussaint, “The relative neighbourhood graph of a finite planar set.” In Pattern recognition, Vol. 12, 4, pp. 261-268, 1980.

[22] J-R. Ward, “Hierarchical grouping to optimize an objective function.” In Journal of the American statistical association JSTOR, Vol. 58, 301, pp. 236-244, 1963.

[23] M-J. Warrens, “Bounds of resemblance measures for binary (presence/absence) variables.” In Journal of Classification, Springer, Vol. 25, 2, pp. 195-208, 2008.

[24] D. Zighed, R. Abdesselam and A. Hadgu, “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, pp. 379-391, 2012.