Metodološki zvezki, Vol. 9, No. 2, 2012, 107-118

Mallows' L^2 Distance in Some Multivariate Methods and its Application to Histogram-Type Data

Katarina Košmelj[1] and Lynne Billard[2]

Abstract

Mallows' L^2 distance allows for decomposition of total inertia into within and between inertia according to Huygens theorem. It can be decomposed into three terms: the location term, the spread term and the shape term; a simple and straightforward proof of this theorem is presented. These characteristics are very helpful in the interpretation of the results for some distance-based methods, such as clustering by k-means and classical multidimensional scaling. For histogram-type data, Mallows' L^2 distance is preferable because its calculation is simple, even when the number and length of the histograms' subintervals differ. An illustration of its use on population pyramids for 14 East European countries in the period 1995-2015 is presented. The results provide an insight into the information that this distance can extract from a complex dataset.

1 Introduction

1.1 Previous work

An important class of metrics was invented several times in different fields, such as probability, statistics and computer science, and can be found under different names. According to Rüschendorf (2001), from the historic point of view, the name Gini–Dall'Aglio –Kantorovich–Vasershtein–Mallows metric would be correct for this class of metrics. In statistics, Mallows' distance between two distributions is defined as the minimum expected distance among all pairs of random variables having those two distributions as marginal distributions (Mallows, 1972; Zhou and Shi, 2011). In computer science, the term Earth Movers' Distance (EMD) is very often used. In Levina and Bickel (2002), it is shown that Mallows' distance is exactly the same as the EMD when it is used to measure the difference between two probability distributions.

In the literature, several distances for histogram-type data can be found. Irpino and Verde (2006) proposed a "new Wasserstein based" distance, which is identical to Mallows' L^2 distance. The paper presents several useful properties of this distance under the assumption that the distribution within each histogram subinterval is uniform. They

[1] Biotechnical Faculty, University of Ljubljana, Slovenia; katarina.kosmelj@bf.uni-lj.si
[2] University of Georgia, Athens, USA; lynne@stat.uga.edu

show, that in the histogram setting, this distance is easily calculable, even when the number and the length of the histograms' subintervals differ. Moreover, this distance allows for identification of the *barycentric (centroid)* histogram. A measure of total inertia can be defined and can be decomposed into the within and between inertia according to the Huygens theorem.

Verde and Irpino (2007) analyzed different metrics in the dynamic clustering of histogram data. Korenjak-Černe et al. (2008) used Euclidean distance to cluster population pyramids. In the symbolic data setting (Billard and Diday, 2006), several distances for histograms are presented by Kim and Billard (2011).

1.2 Motivation for the present work

Košmelj and Billard (2011) applied Mallows' L^2 distance in a clustering procedure on 14 East European described by the population pyramids. The data were for the years 1995, 2000, 2005, 2010, and the "predicted" data for 2015; the data were obtained from http://www.census.gov/ipc/www/idb/informationGateway.php. The objective was to partition these 14 countries into homogenous groups in each particular year, and to observe time-trends from the obtained results. As in Irpino and Verde (2006), Ward's agglomerative clustering method was used; the results for each year under study are presented.

The present paper expands the application of Mallows' L^2 distance to some other multivariate distance-based methods, using the same dataset. Given the Huygens theorem for the Mallows' L^2 distance, a natural choice for the clustering method is the k-means method which minimizes the within cluster inertia for a given number of clusters in a partition. A nice overview of the history of k-means in given by Bock (2007); this method is in some settings called dynamic clusters method or dynamic clustering (Irpino et al., 2006; Verde and Irpino, 2007). We want to implement clustering by k-means using Mallows' L^2 distance; for each year under study the best partition of countries according to this criterion will be sought. A comparison with previous results based on Ward's agglomerative clustering method is of interest as well.

A set of distance-based multivariate methods is multidimensional scaling, which is often used in data visualization. These methods were introduced in the field of psychometrics (see Shepard, 1962; Kruskal, 1964), a relevant book on this topic is Cox and Cox (2001). Gower (1966) described a method to obtain a Euclidean representation of a set of objects whose relationships are measured by any dissimilarity or distance chosen by the user. This algebraic method, known as *principal coordinate analysis*, *metric multidimensional scaling* (in contrast to the nonmetric method), or *classical multidimensional scaling*, allows us to position objects in a space of reduced dimensionality while preserving their distance relationships as well as possible. These methods are also often used in ecology. In Legendre and Legendre (1998) the computational procedure is presented.

An additional motivation for the use of classical multidimensional scaling (CMDS) is given in an interesting paper on distances between empirical distributions by Zhou and Shi (2011). These authors consider several dissimilarity measures including Mallows' L^2 distance. They present a nice overview of its important properties. In a simulation study, they apply CMDS to datasets drawn from a family of univariate two-component mixture Gaussian distributions that are described by its mixing proportion, mean and standard deviation. For Mallows' L^2 distance, they found that the first dimension of CMDS perfectly

corresponds to the mean parameter, the other two leading dimensions together represent the standard deviation and shape of the distributions in a nonlinear fashion. They summarize that Mallows' L^2 distance is preferred in applications where support dependent features of a distribution, such as mean and variance, are important, especially when the distributions have potentially different supports. It also has an advantage of a well defined metric which will be convenient when rigorous modeling is needed.

An other important feature for Mallows' L^2 distance is the decomposition theorem, first presented by Irpino and Romano (2007). In their paper, the proof of the decomposition theorem is presented; however, it is rather complex and long.

To sum up, the motivation for our work is as follows. First, we try to find a simpler alternative for the decomposition theorem proof. Second, other distance-based multivariate methods such as clustering by k-means and CMDS are applied using Mallows' L^2 distance to give a deeper insight into the population pyramids under study. We are interested in how the location, spread and shape of the population pyramids are reflected in k-means and CMDS results, i.e., what information can be extracted from the dataset.

This paper is organized as follows: in the methodology section, we present the definition for Mallows' L^q distance, and in particular Mallows' L^2 distance with its decomposition theorem. Section 3 gives k-means clustering and CMDS results and its analysis. In the last section, some conclusions are presented.

2 Methodology

Here we present the formula for Mallows' L^q distance, $q \in [1, \infty)$. Let us denote the univariate variable Y for the object u by Y_u, and for the object v by Y_v. Mallows' L^q distance for the variable Y between objects u and v is expressed in the form of the corresponding inverse distribution functions F_u^{-1} and F_u^{-1}:

$$d_M^q(u, v) = d_M^q(F_u, F_v) = \int_0^1 |F_u^{-1}(t) - F_v^{-1}(t)|^q dt. \tag{2.1}$$

The inverse distribution function is called the quantile function. Some of its characteristics are presented in the Appendix.

An interesting review of the Mallows' distance properies is given in Levina and Bickel (2002).

2.1 Mallows' L^2 distance and its decomposition

Mallows'distance for $q = 2$, i.e., Mallows' L^2 distance is commonly used:

$$d_M^2(u, v) = d_M^2(F_u, F_v) = \int_0^1 (F_u^{-1}(t) - F_v^{-1}(t))^2 dt. \tag{2.2}$$

It can be considered as a natural extension of the Euclidean distance from point data to distribution data. For this case, the decomposition theorem first presented in Irpino and Romano (2007), holds, as:

$$d_M^2(F_u, F_v) = (\mu_u - \mu_v)^2 + (\sigma_u - \sigma_v)^2 + 2\sigma_u\sigma_v - 2Cov_{QQ}(u, v), \tag{2.3}$$

or equivalently,

$$d_M^2(F_u, F_v) = (\mu_u - \mu_v)^2 + (\sigma_u - \sigma_v)^2 + 2\sigma_u\sigma_v(1 - Corr_{QQ}(u, v)), \tag{2.4}$$

where

$$Corr_{QQ}(u, v) = \frac{\int_0^1 (F_u^{-1}(t) - \mu_u)(F_v^{-1}(t) - \mu_v)dt}{\sigma_u\sigma_v} \tag{2.5}$$

is the correlation of the quantiles as represented in a classical QQ plot.

In (2.4), the first term is the squared difference of the mean values and describes the *location component*; the second term is the squared difference of the standard deviations and describes the *spread component*; the last term is the residual term and presents the *shape component*. It should be pointed out that the means are in one term of (2.4) only; however the standard deviations are in both the second and in the third term of (2.4). Hence, the means present an independent component of the distance; however the standard deviations are interrelated with the correlation.

Irpino and Romano (2007) presented the proof for the decomposition theorem in their Appendix, however it is rather complicated. Here we present a simple and straightforward alternative for the proof. Mallows' distance between two distributions is the minimum expected distance among all pairs of random variables having those two distributions as marginal distributions (Zhou and Shi, 2011:3-4):

$$d_M^2(u, v) = \int_0^1 (F_u^{-1}(t) - F_v^{-1}(t))^2 dt = \inf E[(Y_u - Y_v)^2]. \tag{2.6}$$

The infimum is over all joint distributions having F_u and F_v as marginals. Let us develop $E[(Y_u - Y_v)^2]$ further. Using a standard approach often used in analysis of variance, we can write:

$$\begin{aligned} E[(Y_u - Y_v)^2] &= E[((Y_u - \mu_u) - (Y_v - \mu_v) + (\mu_u - \mu_v))^2] \\ &= E[(Y_u - \mu_u)^2 + (Y_v - \mu_v)^2 + (\mu_u - \mu_v)^2 - \\ &\quad -2(Y_u - \mu_u)(Y_v - \mu_v) + 2(Y_u - \mu_u)(\mu_u - \mu_v) - 2(Y_v - \mu_v)(\mu_u - \mu_v)] \\ &= E[(Y_u - \mu_u)^2] + E[(Y_v - \mu_v)^2] + (\mu_u - \mu_v)^2 - \\ &\quad -2E[(Y_u - \mu_u)(Y_v - \mu_v)] + 2(\mu_u - \mu_v)E[(Y_u - \mu_u - Y_v + \mu_v)]. \end{aligned} \tag{2.7}$$

The last term is zero; therefore, upon taking the infimum over all joint distributions with marginals F_u and F_v, and considering that marginal quantites μ_u, μ_v, σ_u, σ_v are not affected by the infimum, the decomposition theorem (2.3) is obtained:

$$\begin{aligned} d_M^2(u,v) &= (\mu_u - \mu_v)^2 + \inf E[((Y_u - \mu_u) - (Y_v - \mu_v))^2] \\ &= (\mu_u - \mu_v)^2 + \sigma_u^2 + \sigma_v^2 - 2\sup E[(Y_u - \mu_u)(Y_v - \mu_v)] \\ &= (\mu_u - \mu_v)^2 + (\sigma_u - \sigma_v)^2 + 2\sigma_u\sigma_v - 2Cov_{QQ}(u,v). \end{aligned}$$

3 Results

In our dataset, there are 14 East European countries (Albania (AL), Bosnia and Herzegovina (BA), Bulgaria (BG), Czeck Republic (CZ), Croatia (HR), Hungary (HU), Kosovo (KO), Montenegro (ME), Macedonia (MK), Poland (PL), Romania (RO), Serbia (RS), Slovenia (SI), Slovakia (SK)). Each country is described by the population pyramids for the years 1995, 2000, 2005, 2010 and the "predicted data" for the year 2015. The population pyramid is a double histogram, two gender histograms are plotted horizontally back-to-back, on the left for males and on the right for females. On the *y*-axis are age groups: the subintervals are five-year age groups; on the *x*-axis are the corresponding proportions.

For each year under study, we calculated the distance matrix using Mallows' L^2 distance, the formula is given in Irpino and Verde (2006) and in (2.10) in Košmelj and Billard (2011); this distance matrix is the input for clustering by k-means and for CMDS.

3.1 Clustering by k-means

We adapted the `kmeans` procedure in R (R Development Core Team, 2012) for the histogram setting. The adaptation was made taking into account Mallows' L^2 distance for the histogram setting; for each cluster the barycentric histogram was calculated, as described Irpino and Verde (2006) and Košmelj and Billard (2011). This clustering approach is equivalent to dynamic clustering of histograms, as presented in Irpino et al.(2006) and Verde and Irpino (2007).

The clustering procedure was as follows: for each particular year under study, we started with the partition into $k = 2$ clusters. The initial partition was random, the clustering procedure was repeated hundred times. Then the procedure was repeated for $k = 3,\ 4,$ and 5 clusters. The best partition into k clusters was assessed in term of the Calinski-Harabanz pseudo F-statistics (Calinski and Harabanz, 1974):

$$CH(k) = \frac{BI(k)/(k-1)}{WI(k)/(n-k)}, \tag{3.1}$$

where $BI(k)$ and $WI(k)$ denote the between and within inertia for the number of clusters k, respectively, and n the number of objects. In Table 1, we present these values; higher values of $CH(k)$ are preferred. From Table 1, it is evident that in 1995, 2000 and 2005, the partition into three clusters turned out to be the best. In 2010 a major change occurred, the best partition has 5 clusters. In 2015, the best partition has 4 clusters.

Table 2 presents the best partitions. In 1995, 2000 and 2005 the best partitions were the same, specifically: Cluster 1: [AL, KO]; Cluster 2: [BA, ME, MK, PL, RO, SK];

Table 1: Calinski-Harabanz pseudo F-statistics for the partitions into 2, 3, 4, and 5 clusters, for each year under study; the preferred value is the maximal value (presented in bold).

	Number of clusters			
Year	2	3	4	5
1995	33.40	**63.62**	53.77	45.16
2000	33.51	**61.86**	53.41	44.71
2005	37.38	**51.18**	44.55	37.52
2010	36.38	39.09	47.48	**53.67**
2015	34.96	32.99	**50.92**	40.82

Table 2: Left: the obtained optimal partition for each year according to k-means clustering and Calinski-Harabanz criterion (see Table 1). Right: total inertia (TI), within (WI) and between inertia (BI) for the best partition in each year.

Year	Cluster 1	Cluster 1a	Cluster 2	Cluster 2a	Cluster 3	TI	WI	BI
1995	AL, KO		BA, ME, MK, PL, RO, SK		BG, CZ, HR, HU, RS, SI	448.8	35.7	413.1
2000	AL, KO		BA, ME, MK, PL, RO, SK		BG, CZ, HR, HU, RS, SI	422.3	34.5	387.8
2005	AL, KO		BA, ME, MK, PL, RO, SK		BG, CZ, HR, HU, RS, SI	394.6	38.3	356.3
2010	KO	AL	MK, SK	BA,ME, PL, RO	BG, CZ, HR, HU, RS, SI	362.2	14.6	347.6
2015	KO	AL, MK		ME, PL, RO, SK	BA, BG, CZ, HR, HU, RS, SI	342.2	21.0	321.2

Cluster 3: [BG, CZ, HR, HU, RS, SI]. In 2010, changes started: Cluster 1 and Cluster 2 split into two subclusters: [AL, KO] split into singletons [KO] and [AL], Cluster 2 split into subcluster [MK, SK] and subcluster [BA, ME, PL, RO]; however Cluster 3 did not change. In 2015, the changes continued: MK joined [AL], SK joined [ME, PL, RO], and BA joined Cluster 3.

Table 2 gives additional information on the total inertia (TI), and on the within (WI) and between inertia (BI) for the best partition in each year. An important feature of the obtained partitions is the decreasing value of TI in time: in 1995 its value is 448.8, in 2005 it is 394.6, and in 2015 it is 342.2. The entity TI could be regarded as a measure of total variability, therefore these results suggest that countries become more and more similar in time.

Analysis of the WI allows comparison with the Ward's results presented in Košmelj and Billard (2011). For the year 1995, k-means results present a slight improvement (Ward positions RO in Cluster 3, the corresponding WI equal to 37.6). The results obtained by Ward's method and k-means for 2000, 2005 and 2015 are the same. However, for 2010 the difference between k-means and Ward's results is substantial: Ward's partition has 4 clusters with WI equal to 24.6, while k-means best partition has 5 clusters with WI equal to 14.6.

3.2 Multidimensional scaling

Multidimensional scaling takes a set of distances and returns a set of points in a low-dimensional Euclidean space such that the distances between the points are approximately

equal to the original distances. For our purposes, we used the classical multidimensional scaling procedure `cmdscale` in R (R Development Core Team, 2012).

The CMDS results for each particular year under study are conceptually similar. From 14 eigenvalues, five or six are positive; their order of magnitude decreases from 10^5 to 10^2, 10^1, etc. What is the maximum dimension of the output space? The original data matrix has 14 rows (countries), and the number of variables equals two (age for females, age for males). With the Euclidean-type distance, when there are more objects than variables, the maximum dimension of the output space equals the number of variables (Legendre and Legendre, 1998:426). Therefore, for each year, the countries may be represented in a two-dimensional Euclidean space. The order of magnitude of the first two eigenvalues suggests that the first CMDS axis is by far the most important.

Graphical visualization of the CMDS results is given in Figure 1. It is known that the representation is determined up to location, rotation and reflection. For 1995 and 2000, the first CMDS axis, denoted by $MSD1$ (see Figure 1), was reversed to obtain comparable results in the observed time-period. The years 1995, 2005, 2010 and 2015 are chosen for the presentation; the year 2000 is omitted, because the results are very similar to the neighboring two years 1995 and 2005.

Figure 1 highlights the proximity among countries which are described by population pyramids, for each particular year. Each country is represented by a point, as in classical multidimensional scaling approach. Additionally, the k-means results (Table 2) are added onto the plots; clusters are schematically represented by rectangles. Hence, on Figure 1 we can see the position of the countries in two-dimensions obtained by classical multidimensional scaling and the k-means results, simultaneously. The series of panels reveals the changes in time.

Let us try to interpret the results along the classical multidimensional axes $MSD1$ and $MSD2$. For each year under study, Mallows' distance is calculated as the sum of Mallows' distance for males and for females. Therefore the decomposition theorem gives two location components, two spread components and two shape components. For each year under study, we calculated the correlation coefficient between $MDS1$ and the average age over gender. The results show nearly perfect correlation: absolute value of the correlation coefficient was over 0.99, for each year under study. Therefore, we can conclude that $MSD1$ *reflects the average age*. In Figure 1, the average age is decreasing with increasing value of $MSD1$; the average age for AL and for KO are the lowest, however they are increasing in time (e.g., the average age for males and females in AL in 1995 is 27.8 years and 28.8 years, respectively; in 2015 these values increased to 34.9 years and 36.6 years). However, the age-trend is not revealed in panels of Figure 1.

Given two output dimensions only, $MSD2$ *reflects the spread and shape components*, for males and females simultaneously; these components can not be disentangled. However, the countries with biggest differences in spread/shape can be identified. As an illustration, let us consider AL and SI in 2015; their difference in $MSD2$ is considerable (see Figure 1). In Figure 2, their population pyramids are presented. The different shapes of the pyramids are evident; however, the spread of the histograms is relatively similar (the age standard deviation for AL is 21.6 years for males and 21.7 years for females, and for SI the corresponding values are 21.6 years and 23.0 years.

Figure 1 shows as well the position of cluster-rectangles in the space of $MDS1$ and $MDS2$. Clusters are separated in the $MDS1$ dimension, which presents the average

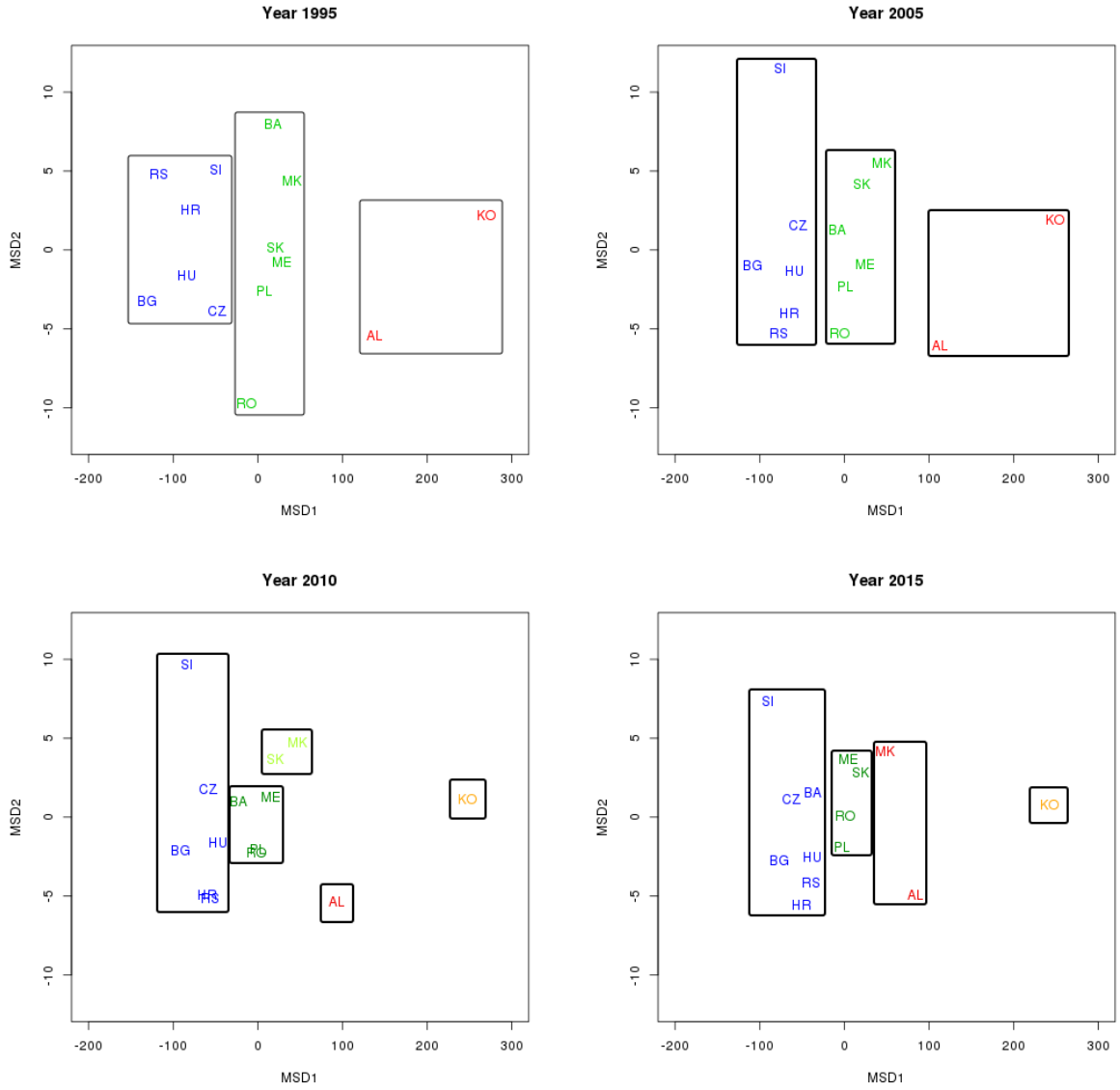

Figure 1: Results of classical multidimensional scaling in two-dimensional space of MDS1 and MDS2. The k-means results are added onto the plots (see Table 2); clusters are schematically represented by rectangles. Year 2000 is omitted.

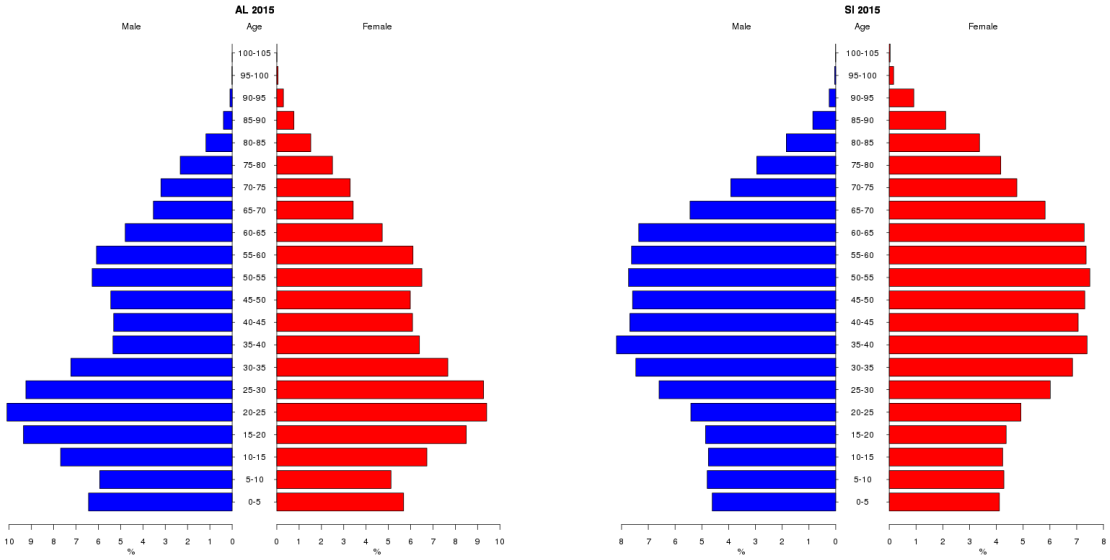

Figure 2: Population pyramid for Albania 2015 (left) and Slovenia 2015 (right). The scales on the horizontal axes are different.

age. Thus, we may conclude that clusters incorporate countries with similar average age. As already mentioned, the eigenvalues obtained in the CMDS analysis show that the information in $MDS1$ is by far the most important.

4 Conclusion

We used Mallows' L^2 distance with two distance-based multivariate methods, k-means and classical multidimensional scaling, in order to obtain a deeper insight into the 14 population pyramids in the time period 1995-2015. The main interest of our analysis was to find out what information this distance can extract from our dataset. The results obtained uncover a lot of information. This may be explained by the fact that Mallows' L^2 distance is a Euclidean-type of distance, fulfilling the Huygens theorem and the decomposition theorem.

Huygens theorem allows for the total inertia to be decomposed into the within and between cluster inertia. The k-means clustering method, which minimizes the within cluster inertia for a given number of clusters in a partition, is a natural choice for the clustering procedure when Mallows' L^2 distance is used. The k-means results in connection with Calinski-Harabanz pseudo F-statistics reveal the best partitions for each year under study and the corresponding time dynamics. For 1995, 2000 and 2005, the best partition has three clusters and is stationary. In 2010 changes started. Analysis of total inertia show that countries become more and more similar in time.

Multidimensional scaling is a data visualization technique.In our context, it is used as in classical data analysis. In general, it is not possible to find a simple interpretation of the corresponding axes, as for example in principal component analysis (PCA). We used CMDS with Mallows' L^2 distance: the results reveal the location component in the first

axis and the spread/shape component in the second axis. Given the fact that the spread of the observed population pyramids is relatively similar, the second axis extracts the shape components. Graphical presentation of CMDS results (Figure 1) incorporates also the clustering results; it is very powerful, time-dynamics can also be observed. Our results are consistent with the findings presented in Zhou and Shi (2011).

Acknowledgment

We thank Batagelj, V. for his advice concerning adapting `kmeans` procedure in R.

References

[1] Billard L. and Diday E. (2006): *Symbolic Data Analysis: Conceptual Statistics and Data Mining.* Wiley Series in Computational Statistics.

[2] Bock, H.H. (2007): A history of the k-means algorithm. In Brito. P. et al. (Ed.): *Selected Contributions in Data Analysis and Classification*, 161-172. Heidelberg: Springer Verlag.

[3] Cox, T.F. and Cox, M.A.A. (2001) *Multidimensional Scaling*. Second edition. Chapman and Hall.

[4] Calinski, T. and Harabasz, J. (1974): A dendrite method for cluster analysis. *Communications in Statistics*, **3**, 1-27.

[5] Gower, J.C. (1966): Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325-328.

[6] Irpino A. and Verde R. (2006): A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In Batagelj, V., Bock, H.H., Ferligoj, A., and Žiberna, A. (Eds): Data Science and Classification, 185-192. Berlin: Springer.

[7] Irpino, A., Verde, R., and Lechevallier, Y. (2006): Dynamic clustering of histograms using Wasserstein metric. In Rizzi, A. and Vichi, M. (Eds): *COMPSTAT 2006*, 869-876. Berlin: Physica-Verlag.

[8] Irpino, A. and Romano, E. (2007): Optimal histogram representation of large data sets: Fisher vs piecewise linear approximations. *Revue des nouvelles technologies de l'information*, **RNTI-E-9**, 99-110.

[9] Kim, J. and Billard, L. (2012): Dissimilarity measures for histogram-valued observations. *Communications in Statistics: Theory and Methods*, in press.

[10] Korenjak-Černe, S., Kejžar, N. and Batagelj, V. (2008): Clustering of population pyramids. *Informatica*, **32**, 157-167.

[11] Košmelj, K. and Billard, L. (2011): Clustering of population pyramids using Mallows' L^2 distance. *Metodološki zvezki*, **8**, 1-15.

[12] Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, **29**, 1-27.

[13] Legendre, P. and Legendre, L. (1998): *Numerical Ecology*. Amsterdam: Elsevier.

[14] Levina, E. and Bickel, P. (2002): The EarthMover's Distance is the Mallows Distance: Some Insights from Statistics (http://www.stat.lsa.umich.edu/ elevina/EMD.pdf).

[15] Mallows, C.L. (1972): A note on asymptotic joint normality. *Annals of Mathematical Statistics*, **43**, 508-515.

[16] R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org/.

[17] Rüschendorf, L. (2001): Wasserstein metric. In M. Hazewinkel (Ed): *Encyclopaedia of Mathematics*. Berlin: Springer.

[18] Shepard, R.N. (1962). The analysis of proximities: multidimensional scaling with unknown distance function. Psychometrika, **27**, 125-140.

[19] Verde, R. and Irpino, A. (2007): Dynamic clustering of histogram data: using the right metric. In Brito, P., Bertrand, P., Cucumel, G., and de Carvalho, F. (Eds.): *Selected Contributions in Data Analysis and Classification*, 123-134, Berlin: Springer.

[20] Williams, C.K.I. (2002): On a connection between Kernel PCA and Metric Multidimensional Scaling. *Machine Learning*, **46**, 11-19.

[21] Zhou, D. and Shi, T. (2011): Statistical Inference Based on Distances Between Empirical Distributions with Applications to AIRS Level-3 Data. *CIDU 2011*, 129-143 (http://www.stat.osu.edu/ taoshi/research/papers/2011_Zhou_Shi_CIDU.pdf).

A Appendix

Let Y be a random variable with finite moments of the first and the second order, and let $F(y)$ be its distribution function, $F(y) = P(Y \leq y)$. The inverse distribution function is called the quantile function. Let us present how the mean and the variance of Y are expressed in terms of the quantile function, given it exists.

The mean of the variable Y is given as follows:

$$\mu = E(Y) = \int_{-\infty}^{\infty} y dF(y) = \int_0^1 F^{-1}(t)dt. \tag{A.1}$$

Equation (A.1) can be proved using the substitution $t = F(y)$. Then, $F^{-1}(t) = F^{-1}(F(y)) = y$, and $dt = dF(y)$.

Similarly,

$$E(Y^2) = \int_{-\infty}^{\infty} y^2 dF(y) = \int_0^1 (F^{-1}(t))^2 dt. \tag{A.2}$$

Consequently, the variance of the variable Y is:

$$\sigma^2 = \int_0^1 (F^{-1}(t))^2 dt - (\int_0^1 F^{-1}(t)dt)^2. \tag{A.3}$$