doi.org/10.26398/IJAS.0029-017

MULTIPLE FACTOR ANALYSIS OF DISTRIBUTIONAL DATA

Rosanna Verde[1], Antonio Irpino

Department of Mathematics and Physics, University of Campania "Luigi Vanvitelli", Caserta, Italy

Abstract. *In the framework of Symbolic Data Analysis (SDA), distributional variables are a particular case of multi-valued variables: each unit is represented by a set of distributions (e.g., histograms, density functions, or quantile functions), one for each variable. Factor analysis methods are primary exploratory tools for dimension reduction and visualization. In the present work, we use a Multiple Factor Analysis (MFA) approach for the analysis of data described by distributional variables. Each distributional variable induces a set of new numeric variables related to the quantiles of each distribution. The set of quantile variables related to a distributional one is treated as a block in the MFA approach. Thus, a MFA is performed on juxtaposed tables of quantile variables. We show that the criterion decomposed in the analysis is an approximation of the variability based on a suitable metric between distributions: the squared L_2 Wasserstein distance. Applications on simulated and real distributional data corroborate the method. The interpretation of the results on the factorial planes is performed by new interpretative tools that are related to several characteristics of the distributions (location, scale, and shape).*

Keywords: *Multiple factor analysis, Distributional data, Wasserstein distance.*

1. INTRODUCTION

In the framework of Symbolic Data Analysis (SDA), *distributional* (or distribution-valued) data (DD) are data described by distributions (such as histograms, density, or quantile functions) for a numeric variable. According to Bock and Diday (2000), such data are realizations of a numeric *modal* symbolic variable, also named as a *distributional* variable. In general, distributional data can be expressed by a parametric or a non-parametric density function estimated on a set of observed values. Dimension reduction techniques have been extended to the analysis of multivalued variables to visualize the proximity between the individuals and the correlations between the variables onto lower dimensional spaces. Factorial methods,

[1] Corresponding author: Rosanna Verde, email: rosanna.verde@unicampania.it

such as Principal Component Analysis (PCA), are popular techniques for a dimensional reduction of a set of p numeric variables observed on n individuals. The aim is to extract a set of new *orthogonal* factors that explain the variance-covariance structure through few linear combinations of the original variables.

PCA, similarly to all the other factorial techniques, reduces the redundant information presented by the data such that the more the variables are correlated, the higher the dimensionality reduction is.

When data are distributions, the extraction of new factorial axes should take into account the characteristics of distributions and the variability among distributions. Although the meanings of orthogonality, variance, covariance, and correlation are consolidated for classical numeric variables, this is not the case for distributional ones.

In the framework of SDA, different PCA methods have been proposed for interval-valued data (see Lauro et al. (2008b) and Lauro et al. (2008a) for an extensive review), but only a few proposals exist for distributional ones.

Some proposals have been designed for histogram-valued data (Cazes, 2002; Le-Rademacher, 2008; Makosso-Kallyth and Diday, 2012; Nagabhushan and Kumar, 2007; Rodriguez et al., 2000), and they differ regarding what variability criterion is decomposed in the analysis.

Rodriguez et al. (2000) proposed a way to extend the PCA for interval data (Cazes et al., 1997) to include histogram data by considering intervals of relative frequencies. In this approach, it is supposed that the histograms share a common partition of the support (i.e., the same set of bins) and that analysis is conducted only on a transformation of the frequencies of the bins of histogram data. The decomposed covariance structure of the data takes into account only the covariance of the centers of the intervals of frequencies while the information related to the numerical support of the histograms is ignored. In a second contribution (Cazes, 2002), a variance-covariance matrix of a set of multivariate distributions is decomposed under the hypothesis of conditional independence. In this case, the conditional independence assumption leads a consideration of only the covariances between the means, and the variability related to the different sizes and shapes of the distributions is lost. Another proposal (Nagabhushan and Kumar, 2007) using a PCA for histogram-valued data considers only the empirical frequencies observed for each bin of the observed histograms, losing, like in Rodriguez et al. (2000), the information related to the support. Ichino (2008, 2011) proposed a PCA of *quantile representations* of symbolic data (they are particular transformations of the observed distributional data). However, the author did not define the geometric properties of the decomposed covariances

explicitly, and he did not give a suitable interpretation for the explained variability on the factorial sub-spaces. Le-Rademacher (2008) proposed an extension of the interval PCA for histogram-valued data (considered as weighted intervals). In this case, the eigenvalues of the PCA decompose an inertia measure corresponding to the sum of the variances of histogram-valued variables, as presented in Billard and Diday (2006).

Finally, an extension of the interval PCA to include histogram variables was proposed by Makosso-Kallyth and Diday (2012); here, a PCA is performed on the means of the histograms (similar to a centers PCA). Then, using the Tchebicheff inequality, the histograms are transformed into intervals and projected on the space spanned by the principal components. Unfortunately, the principal components are related only to the covariances of the means of the histograms. More recently, Wang et al. (2014) proposed an adaptation of the previous methods for the PCA of normal distribution-valued data.

All the above-mentioned methods do not explicitly require the definition of a measure of covariance between distributional variables in advance. Some basic statistics for histogram variables were presented by Bock and Diday (2000) and developed by Billard and Diday (2006). Recently, Verde and Irpino (2008) proposed new variance, covariance and correlation measures for distributional variables based on the *l*2 Wasserstein distance (Rüschendorf, 2011) between distributions.

Both approaches show that the variability of a distributional variable can be decomposed in several components: in the approach of Billard and Diday (2006), the data variability is expressed in the part related to the location and in the part related to the scale; althogh in the approach of Verde and Irpino (2008), the shape of the observed distributions is also taken into consideration. In particular, the results of Verde and Irpino (2008), consistent with the statistical modeling of the quantile functions proposed by Gilchrist (2000), show that the analysis conducted on empirical quantile functions (the inverse of the cumulative distribution functions) has two main interpretative advantages. First, working directly on the quantile functions associated with the empirical distributions, it is not necessary to consider a parametric hypothesis for the distributions. Second, it is possible to interpret the contribution to the results related to the variability of the location, scale, and shape of the distributions separately. More recently, Verde et al. (2016) proposed a PCA method for a single distribution variable using an approximation of the Wasserstein distance between distributions. The idea is to represent the distributional variables through a set of quantile variables and then to apply the PCA to matrix of quantiles.

The data are not standardized, so the information about the several characteristics, i.e., location, size, and shape of the distributions, is retrieved in the determination of the new factorial axes. The results of the analysis provide an interesting interpretation of the axes according to the different moments of the distributions.

The present paper aims to use Multiple Factor Analysis (MFA) to analyze data that are described by a set of distributional variables. MFA has been introduced by the works of Escofier and Pagès (1983, 1988, 1990) (recent overviews are available in Abdi et al. (2013) and Pagès (2014)). MFA is an extension of PCA on sets of variables (namely, blocks of variables), and it is one of the multi-tables techniques (e.g., STATIS, Multiblock Correspondence Analysis, SUM-PCA). MFA is conducted in two steps: first, it runs a PCA of each block of variables, and then, it normalizes each block by the respective first singular value, so that the first principal components have the same length. Second, it performs a common representation of the data sets, which is called a compromise or consensus representation. This compromise is obtained from a (non-normalized) PCA of a table obtained from the concatenation of the normalized blocks of variables.

Here, we propose to apply an MFA to a transformation of distributional variables in sets of quantile variables. Each set of quantile variables is related to a distributional variable and it is a block of variables in the MFA. The peculiarity of this approach comes in the decomposition of an approximation of the total variability of the distributional data according to the squared ℓ_2 Wasserstein distance (see Verde et al. (2016)). In this way, we preserve the coherence between the criterion optimized in the MFA and the distributional data space defined by the *1*2 Wasserstein metric. Finally, the proposal uses visualization and graphical interpretative tools to analyze the relationships between distributions according to their own characteristics, i.e., location, scale, and shape, on factorial planes.

The remainder of the paper is structured as follows: Section 2 introduces the data and the Wasserstein metric between distributions. Section 3 presents the extension of the MFA on quantile variables to analyze relationships between the distributional variables observed on the same set of individuals. Section 4 shows a new tool for the visualization of the distributional variables on the factorial planes. Sections 5 and 6 show the results of the applications on simulated and real data, respectively. Section 7 concludes the paper.

2. DISTRIBUTIONAL VARIABLES AND THE WASSERSTEIN DISTANCE

Let E be a set of n individuals described by a distributional variable Y, i.e., a modal-valued variable with a numerical domain $\mathscr{S} = [min(Y), Max(Y)] \subset \Re$. We denote with y_i ($i = 1, ..., n$), the realization of the variable Y for the $i-th$ individual (Verde and Irpino, 2008), expressed by a (empirical or estimated) probability density function $f_i(y)$. We denote by $F_i(y)$ a cumulative density function (cdf) and by $F_i^{-1}(t)$ (for $t \in [0;1]$) the corresponding quantile function (*qf*, i.e., the inverse of the cdf). According to Gilchrist (2000), several advantages can arise by working with *qf*s rather than with the distribution functions: all the *qf*s have a finite domain in $[0;1]$, the sum of quantile functions returns a *qf*, the product of a *qf* by a positive scalar returns a *qf*, and under certain conditions, it is possible to define the product between two *qf*s.

Several proposals have been formulated in the framework of SDA to define univariate (mean, variance, and standard deviation) and bivariate (covariance and correlation) statistics for histogram variables (Billard and Diday, 2006; Bock and Diday, 2000). Recently, Verde and Irpino (2008) introduced new measures based on the Wasserstein distance, which is a suitable metric to compare distributions. An overview of the family of Wasserstein metrics is presented by Rüschendorf (2011); Villani (2003).

According to Rüschendorf (2011), the ℓ_p Wasserstein distance between two (univariate) distribution functions can be expressed as follows:

$$d_{W_p}(y_i, y_{i'}) = \left[\int_0^1 \left| F_i^{-1}(t) - F_{i'}^{-1}(t) \right|^p dt \right]^{\frac{1}{p}} \tag{1}$$

where, $p \geq 1$, F_i and $F_{i'}$ are cumulative distribution functions (*cdf*s) associated with the y_i and $y_{i'}$ histograms, and F_i^{-1} and $F_{i'}^{-1}$ are the corresponding quantile functions (*qf*s). The ℓ_2 squared Wasserstein distance, also known as Mallow distance (Rüschendorf, 2011), between the *qsf* associated with two histograms is as follows:

$$d_{W_2}^2(y_i, y_{i'}) = \int_0^1 \left[F_i^{-1}(t) - F_{i'}^{-1}(t) \right]^2 dt. \tag{2}$$

The ℓ_2 Wasserstein metric can be considered a natural extension of the Euclidean metric between quantile functions.

Thus, the Wasserstein distance can be suitably computed for equi-depth histograms with a fixed number of bins equal to s. Given the histogram description

y_i, this can be partitioned into $s \geq 1$ bins as follows:

$$y_i = \{(I_{1i}, \pi_{1i}), ..., (I_{hi}, \pi_{hi}), ..., (I_{si}, \pi_{si})\},$$

where $I_{hi} = [\underline{y}_{hi}; \overline{y}_{hi}]$ is an interval of $\Re$, and $\pi_{hi} \geq 0$ such that $\sum_{h=1}^{s} \pi_{hi} = 1$. If y_i is an *equi-depth histogram*, then $\pi_{hi} = \frac{1}{s}$. The following quantities w_{li} represent the cumulative weights associated with the elementary intervals of $y(i)$:

$$w_{li} = \begin{cases} 0 & l = 0 \\ \sum\limits_{h=1,...,l} \pi_{hi} & l = 1, \ldots, s \end{cases} . \tag{3}$$

For simplicity, we consider two equi-depth histogram descriptions $y(i)$ and $y(i')$ that have the same number of bins equal to s implying that the weights $\pi_{li} = \pi_{li'} = w_{li} - w_{l-1i} = \frac{1}{s}$. In this case, being $w_{li} = w_{li'}$, we omit the second index. The squared Wasserstein distance between two equi-depth histogram descriptions is computed as follows:

$$d_{W_2}^2(y_i, y_{i'}) := \sum_{l=1}^{s} \int_{w_{l-1}}^{w_l} \left(F_i^{-1}(t) - F_{i'}^{-1}(t)\right)^2 dt. \tag{4}$$

Each couple (w_{l-1}, w_l) allows us to identify two uniformly dense intervals, one for i and one for i', having, respectively, the following bounds:

$$I_{li} = [F_i^{-1}(w_{l-1}); F_i^{-1}(w_l)] = [\underline{y}_{li}; \overline{y}_{li}] \text{ and}$$
$$I_{li'} = [F_{i'}^{-1}(w_{l-1}); F_{i'}^{-1}(w_l)] = [\underline{y}_{li'}; \overline{y}_{li'}].$$

The center and the radius of each interval are computed as follows:

$$c_{li} = (\underline{y}_{lu} + \overline{y}_{lu})/2 \quad r_{lu} = (\underline{y}_{lu} + \overline{y}_{lu})/2 \quad for\ u = i, i'.$$

Because the histograms are equi-depth, all the π_l are equal to $1/s$. The intervals that are uniformly distributed can be expressed as functions of their centers and radii. Hence, the equation (4) can be rewritten as follows:

$$d_{W_2}^2(y_i, y_{i'}) = \frac{1}{s} \sum_{l=1}^{s} \left[(c_{li} - c_{li'})^2 + \frac{1}{3} (r_{li} - r_{li'})^2 \right]. \tag{5}$$

According to the Wasserstein metric, the *mean histogram* y_b is defined as a Fréchet mean by solving the following minimization problem:

$$f(y_b) = \arg\min_{y} \sum_{i=1}^{n} d^2_{W_2}(y_i, y), \tag{6}$$

which is expressed as follows:

$$y_b = \left[\left((c_{1b} - r_{1b}; c_{1b} + r_{1b}), \frac{1}{s} \right); \ldots; \left((c_{sb} - r_{sb}; c_{sb} + r_{sb}), \frac{1}{s} \right) \right]; \tag{7}$$

where:

$$c_{lb} = n^{-1} \sum_{i=1}^{n} c_{li} \quad ; \quad r_{lb} = n^{-1} \sum_{i=1}^{n} r_{li}.$$

Therefore, the variance of the distributional variable Y, according to (6) and (7), is defined as follows:

$$Var(Y) = \frac{1}{n} \sum_{i=1}^{n} d^2_{W_2}(y_i, y_b) = \frac{1}{ns} \sum_{i=1}^{n} \sum_{l=1}^{s} \left[(c_{li} - c_{lb})^2 + \frac{1}{3} (r_{li} - r_{lb})^2 \right]. \tag{8}$$

In the next section, we will show that $Var(Y)$ can be approximated by the sum of the variances of the quantile variables of the distributional variable Y.

If histograms are not equi-depth, the computations are done in accordance with Irpino and Verde (2014). In such a case, a homogenization step is requested for comparing histograms through the Wasserstein distance. The support of the distributional data is shared according to a set of quantile values corresponding to the same set of density levels p_i for all the distributions.

2.1. OTHER PROPERTIES OF THE ℓ_2 WASSERSTEIN DISTANCE

The advantage of using the ℓ_2 Wasserstein distance for comparing distributional data is related to the property that it satisfies: the squared distance between two distributions can be decomposed according to the following three components (as shown by Irpino and Romano (2007)):

$$\begin{aligned} d^2_W(y_i, y_{i'}) &= \int_0^1 \left[F_i^{-1}(t) - F_{i'}^{-1}(t) \right]^2 dt = \\ &= \underbrace{(\mu_i - \mu_{i'})^2}_{Location} + \underbrace{\underbrace{(\sigma_i - \sigma_{i'})^2}_{Scale} + \underbrace{2\sigma_i \sigma_{i'}(1 - \rho_{i,i'})}_{Shape}}_{Variability}. \end{aligned} \tag{9}$$

where μ_u and σ_u (with $u = i, i'$) are, respectively, the means and the standard deviations of the distributions y_i and $y_{i'}$, while $\rho_{i,i'}$, is the Pearson correlation coefficient between two quantile functions $F_i^{-1}(t)$ and $F_{i'}^{-1}(t)$

Therefore, $\rho_{i,i'}$ can be considered a measure of shape similarity of two distribution functions. In fact, $\rho_{i,i'} = 1$ only if the two distributions have the same standardized quantiles (by the respective mean and standard deviation), which occurs when the two distributions have the same shape.

The decomposition of the squared ℓ_2 Wasserstein distance between two distribution functions allows for the evaluation of their deviation in terms of *location*, *scale*, and *shape*. The difference in *location* and *scale* is, respectively, expressed by the squared Euclidean distances between the means and between the standard deviations of the two distributions; while the difference in *shape* is related to the value of $\rho_{i,i'}$. The *scale* and *shape* components express together the difference of the *variability* structure between two distributions.

3. MULTIPLE FACTOR ANALYSIS ON THE QUANTILES OF DISTRIBUTIONAL VARIABLES

In this section, we present an MFA on a set of data tables containing the quantile representation of several distributional variables observed on the same individuals. MFA extnends PCA providing a set of common factors for projecting data described by blocks of variables onto a compromise subspace (Escofier and Pagès, 1983). The main idea is to extend the PCA methods for distributional data (Verde et al., 2016) to the case of multi-tables analysis.

According to the PCA strategy on a distributional variable Y, distributions are replaced by a set of predefined quantiles that are assumed to be values of the variables of the analysis.

Let E be the set of p histograms y_{ij} (for $j = 1, \ldots, p$) related to the description of the $i-th$ individual w.r.t. the $Y_1, \ldots, Y_j \ldots, Y_p$ variables. Each y_{ij} is the histogram of values that the $i-th$ individual assumes for the variable Y_j. We consider that all the histograms are equi-depth, so the bounds of the intervals $I_{li} = [\underline{y}_{li}; \bar{y}_{li}]$ (for $l = 1, \ldots, K_j - 1$), and they correspond to the K_j-quantiles, i.e., the values that divide the distribution in K_j equal parts). We have denoted by K_j the number of quantiles for each variable Y_j that can be also chosen as different for each of them ($K_1 \ldots K_j \ldots K_p$).

For the generic variable Y_j, we denote it with the following:

$$\begin{aligned} q_{i0,j} &= \underline{y}_{1i,j} = min(y_{ij}), \\ q_{il,j} &= \overline{y}_{li,j} \ (for \ l = 1,\ldots,K_j - 1) \text{ and} \\ q_{is,j} &= \overline{y}_{K_j i,j} = Max(y_{i,j}) \ \forall i = 1,\ldots,n \end{aligned}$$

To perform a PCA on quantiles, we consider the input a concatenation of classic $n \times (K_j + 1)$ data tables (with $j = 1,\ldots,p$), denoted with $\mathbf{Q}_j$ as follows:

$$\mathbf{Q} = [\mathbf{Q}_1|...|\mathbf{Q}_j|...|\mathbf{Q}_p] \tag{10}$$

Each row of the $j-th$ table $\mathbf{Q}_j$ is an individual representation expressed by the following order statistics: the minimum value (or *zero* quantile) q_{i0}; the l-quantiles q_{il}; and the maximum value or $K_j - th$ quantile, q_{iK_j}.

The generic $i-th$ individual (row) observed for the (single) distributional variable Y_j is described by a set of (K_j+1) quantiles (columns): $Q_{0j},\ldots,Q_{lj},\ldots,Q_{K_j j}$ with $1/K_j$ probability, or the relative frequency of the observed values between two consecutive quantiles.

$$\mathbf{Q}_j = \begin{bmatrix} q_{10,j} & q_{11,j} & \ldots q_{1l,j} & \ldots & q_{1K_j,j} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ q_{i0,j} & q_{i1,j} & \ldots q_{il,j} & \ldots & q_{iK_j,j} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ q_{n0,j} & q_{n1,j} & \ldots q_{nl,j} & \ldots & q_{nK_j,j} \end{bmatrix}$$

We assume that the elements of the matrix $\mathbf{Q}_j$ are centered by subtracting the means of the respective quantile variables Q_{lj} (for $l = 1,\ldots,K_j$).

For simplicity, we refer to the columns of the matrix $\mathbf{Q}_j$ as centered *quantile-variables*. The choice not to standardize the quantile variables preserves the approximation of the variance of a distributional variable based on the Wasserstein metric by the sum of the variances of the quantile variables (as shown hereafter). Particular care should be taken regarding to the lower and higher quantile variables. Indeed, the empirical evidence (see applications on simulated and real data) reveals that those quantile variables may have a higher variability with respect to the other ones. This can be checked before the analysis is performed. A practical solution is to consider the extreme quantile variables as supplementary in the analysis or to give lower weights with respect to the other ones.

We denote by $\mathbf{W}$ the matrix of the individual weights; assuming that all of them have the same weight, it is a diagonal matrix of elements $\frac{1}{n}$.

Moreover, we define the cross-product of quantiles matrix $\mathbf{Q}_j$, weighted by $\mathbf{W}$, as follows:

$$\mathbf{S_j} = \mathbf{Q}_j^\top \mathbf{W} \mathbf{Q}_j \tag{11}$$

where $\mathbf{S}_j$ is the variance-covariance matrix of the quantile variables of Y_j. Then, the cross-product of the matrix $\mathbf{Q}$, weighted by $\mathbf{W}$, is as follows:

$$\mathbf{S} = \mathbf{Q}^\top \mathbf{Q} = [\mathbf{Q}_1|...|\mathbf{Q}_j|...|\mathbf{Q}_p]^\top \mathbf{W} [\mathbf{Q}_1|...|\mathbf{Q}_j|...|\mathbf{Q}_p] = \sum_j \mathbf{Q}_j^\top \mathbf{W} \mathbf{Q}_j = \sum_j \mathbf{S}_j. \tag{12}$$

The $\mathbf{S}$ is the block variance-covariance matrix of the quantile variables Q_{lj} (for $l = 0, \ldots, s$) of the $Y_1, \ldots, Y_j \ldots, Y_p$.

The trace of the matrix $\mathbf{S}$ (denoted by $Tr(\mathbf{S})$) is equal to the sum of the variances of the quantile variables Q_{lj} (denoted by $Var(Q_{lj})$) (for $l = 0, \ldots, s$ and $j = 1, \ldots, p$).

Now, we show the relationship between the usual ℓ_2 Wasserstein metric used in the analysis of distributional data and the criterion decomposed in the MFA.

Verde et al. (2016) showed that the trace of $\mathbf{S}_j$ (denoted by $Tr(\mathbf{S}_j)$) approximates the variances of the distributional variable Y_j (denoted by $Var(Y_j)$), according to the ℓ_2 Wasserstein distance.

Denoted by Δ, the deviation is as follows:

$$\Delta = Tr(\mathbf{S}_j) - Var(Y_j). \tag{13}$$

This depends on the number of quantiles and on the number K_j (with $K = \sum_{j=1}^{p} K_j$) of the intervals (bins) of the supports of the n histogram data, as follows:

$$\Delta = \frac{\sum_{i=1}^{n} \sum_{l=0}^{K_j} (q_{il,j}^c)^2 - \sum_{i=1}^{n} \sum_{l=1}^{K_j} \left[\left(c_{il,j}^c \right)^2 + \frac{\left(r_{il,j}^c \right)^2}{3} \right]}{n \cdot K}.$$

with $c_{il,j}^c = c_{il,j} - \bar{c}_{l,j}$, $r_{il,j}^c = r_{il,j} - \bar{r}_{l,j}$, the rescaled center and radius on the respective means.

3.1. THE TWO STEPS OF MFA

The MFA is performed in two steps.

The first step consists of a PCA on each data table $\mathbf{Q}_j$. The results are obtained by the SVD decomposition, as follows:

$$\mathbf{Q}_j = \mathbf{U}_j \Lambda_j \mathbf{V}_j^{\mathsf{T}}. \tag{14}$$

subject to the usual ortho-normality constraints as follows

$\mathbf{U}_j^T \mathbf{U}_j = \mathbf{V}_j^T \mathbf{V}_j = \mathbf{I}$.

The factor scores are computed as follows:

$$\Psi_j = \mathbf{U}_j \Lambda_j \tag{15}$$

where Λ_j is the diagonal matrix of the eigenvalues of the matrix $\mathbf{Q}_j$.

In the MFA, each table $\mathbf{Q}_j$ is normalized by the respective squared first eigenvalue λ_{1j}, corresponding to the highest value of Λ_j, i.e., the following:

$$a_j = \frac{1}{\sigma} \tag{16}$$

where $\sigma = \lambda_{1j}^2$.

The a_j for $j = 1, \ldots, p$ can represent a system of weights for each matrix that can be arranged in a diagonal matrix $\mathbf{A}$, as follows:

$$\mathbf{A} = diag\{[a_1 \mathbf{1}_{[K_1]}^{\mathsf{T}}, \ldots, a_j \mathbf{1}_{[K_j]}^{\mathsf{T}}, \ldots, a_p \mathbf{1}_{[K_p]}^{\mathsf{T}}]\} \tag{17}$$

where $\mathbf{1}_{[K_j]}$ is a vector of ones, and K_j is the number of quantile-vectors of each block matrix $\mathbf{Q}_j$.

The second step of the MFA consists of a global PCA of the matrices $\mathbf{Q}_j$ normalized by the a_j, which is done by considering the weights of the individuals that are assumed to be all equal to $\frac{1}{n}$. The matrix of the weights of the individual is denoted as $\mathbf{W}$.

In such a way, the MFA is equivalent to an analysis of the triplet $(\mathbf{Q}, \mathbf{W}, \mathbf{A})$, according to the classical definition of the French school (see Lebart et al. (2006)).

The eigensolutions can be obtained by a Generalised SVD of the matrix $\mathbf{Q}$, as follows:

$$\mathbf{Q} = \mathbf{U} \Lambda \mathbf{V}^{\mathsf{T}} \tag{18}$$

under the following constraints:

$$\mathbf{U}^T \mathbf{W} \mathbf{U} = \mathbf{V}^{\mathsf{T}} \mathbf{A} \mathbf{V} = \mathbf{I}. \tag{19}$$

Note that for simplicity, the eigenvectors' and eigenvalues' matrices are denoted with the same letters as in SVD.

The factor scores of the single quantile vectors of $\mathbf{Q}_j$ are computed as follows:

$$\Psi_{j,\alpha} = a_j \mathbf{Q}_j^{\mathsf{T}} \mathbf{v}_\alpha \tag{20}$$

where $\mathbf{v}_\alpha$ is the eigenvector associated with the $\alpha - th$ eigenvalue ($\alpha = 1, \ldots, L$ where L is the rank of $\mathbf{S}$).

The factor scores represent a sort of compromise for a common representation in a reduced subspace of the variability structure of the matrices $\mathbf{Q}_j$ $j = 1, ..., n$.

The compromise factor score Ψ_α is the barycenter of the partial factor scores obtained as the average of the p partial scores factors, as follows:

$$\Psi_\alpha = \frac{1}{p} \sum_j a_j \mathbf{Q}_j^T \mathbf{v}_\alpha. \tag{21}$$

The representation of the individuals (the rows of $\mathbf{Q_j}$) can be obtained according to the classical biplot on the reduced subspaces, as follows:

$$\Phi_{j,\alpha} = \frac{1}{\lambda_\alpha} a_j \mathbf{Q}_j \mathbf{u}_\alpha. \tag{22}$$

4. TOOLS FOR THE INTERPRETATION: THE SPANISH FAN PLOT

Starting from the results of the MFA, it is interesting the interpretation of the proximities between the distributions according to the characteristics, i.e., location, scale and shape, that have contributed more to the determination of the axes.

Indeed, in the determination of the factorial axes, the components related to location, scale, and shape, into which the variance (based on the ℓ_2 Wasserstein distance) of the distributional variable Y can be decomposed, play a different role. In fact, each factor axis is oriented toward the direction of the variability of the means (location parameters), of the standard deviations (size parameters), and of the skewness and kurtosis (shape parameters), respectively. Therefore, the advantage of the proposed approach is its ability to interpret the axes according to the different characteristics of the distribution-valued data. If an MFA is performed on sets of four variables (namely, one set for for each distributional variable) representing the first four moments of the distributions, the results are not so evident.

As in a classical PCA, the representation of the quantile-variables on the factorial planes (e.g., the first plane for α equal to 1 and 2) is given by a circle of correlation given by the quantile vectors. For improving the interpretation of the plots, we connect the consecutive quantiles according to their natural order on the

factorial plan. This results in a nice representation of the quantile-vectors that looks like a *Spanish fan*. We call this representation a *Spanish-fan plot*. Each *Spanish-fan* allows the analysis of the structure of global variability and the visualization of the characteristics (variability and shape) of the distributional variable. We observe that the quantile variables representation usually follows a kind of order (being, in general, two consecutive quantile-variables more correlated w.r.t. two non-consecutive ones). We can explain the pattern of the *fan* with respect to the correlation (i.e., the angle) between two consecutive quantile-variables q_l and q_{l+k} ($k = l+1,\ldots,s$). For example, it is interesting to observe that when the distributions are almost symmetric, the correlation between q_l and q_{l+k} decreases as k ($k = 1,2,\ldots,s-l+1$) increases. Thus, the shape of a *Spanish-fan plot* impacts the interpretation of the factorial plans. When the distributions are different according to their first four moments, we show that the first plane better explains the variability of the locations and scales of the distributions: the more open the *fan* is, the higher the variability of the distributions is; while the second factorial plane (third and fourth axis) usually explains the variability in skewness and kurtosis of the distributions.

Other typical measures, such as the relative contribution, denoted by *cr*, can help interpret the axes. Similar to the classical PCA, the relative contribution of the $i-th$ distribution to the determination of the $\alpha-th$ axis is a measure of how much the variance explained by the $\alpha-th$ axis is because of the $l-th$ quantile variable.

Further, the quality of the representation of the individuals (distributions) and of the quantile variables is measured by the absolute contributions, denoted by *ca*. Similar to the classic PCA, absolute contributions sum to one for each distribution (respectively, for each quantile-variable), and the higher the contribution is, the better the distribution is represented on the axis (or on the plane, if we consider the sum of *ca*'s of each axes of the plane).

5. AN APPLICATION OF THE MFA ON SIMULATED DATA

In this section, we present an application of the proposed MFA on simulated data. For simplicity, we consider only two distributional variables. The simplicity of the proposed application aims to highlight the power of the method, especially as a visualization tool.

Recalling that the proposed MFA method provides the latent structure of the quantiles for each variable according to the first four moments of the distributions, we consider two sets of histogram data observed for the same n individuals.

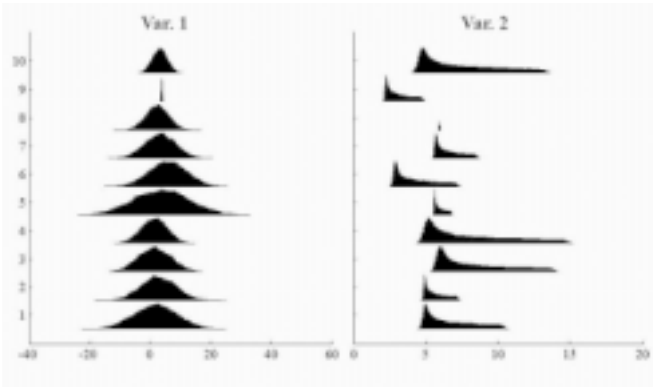

Figure 1: Representation of the two sets of distributional variables

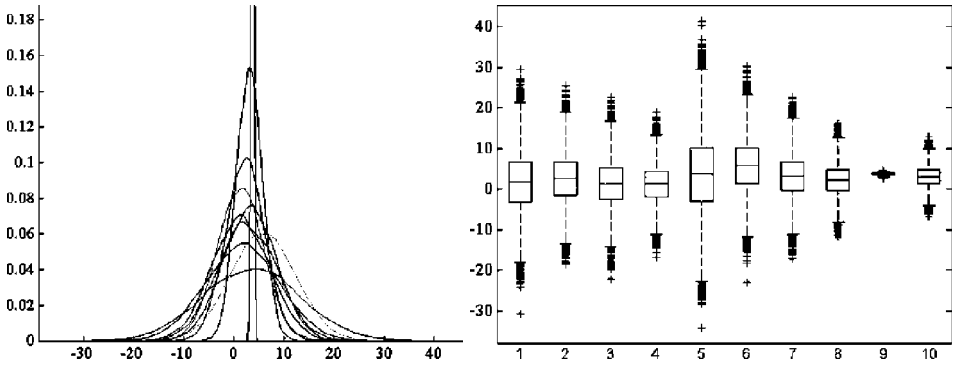

Figure 2: Characteristics of the point distributions of the first distributional variable Y_1

The data related to the first quantile variable are sampled from Gaussian distributions with the same mean but different standard deviations; the second ones are sampled from shifted and scaled Beta distributions. Ten histogram data for each variable were generated as follows: one thousand values were sampled for each distribution, and 19 quantiles are extracted, such that each bin, bounded by two consecutive quantiles, contained 5.55% of the sampled values. In this case, it is equivalent to set up an equi-depth histogram for each distribution that has 18 bins. Using smoothed representations, the two configurations are shown in Fig. 1. The box-plots of the sampled data for each distribution are represented in Fig. 2 and Fig. 3, respectively.

A partial PCA is performed on each block of quantile variables related to Y_1 and Y_2. $\mathbf{Q_1}$ and $\mathbf{Q_2}$ matrices comprise 19 quantile variables (including the min values), respectively. The quantile variables are centered w.r.t. the corresponding

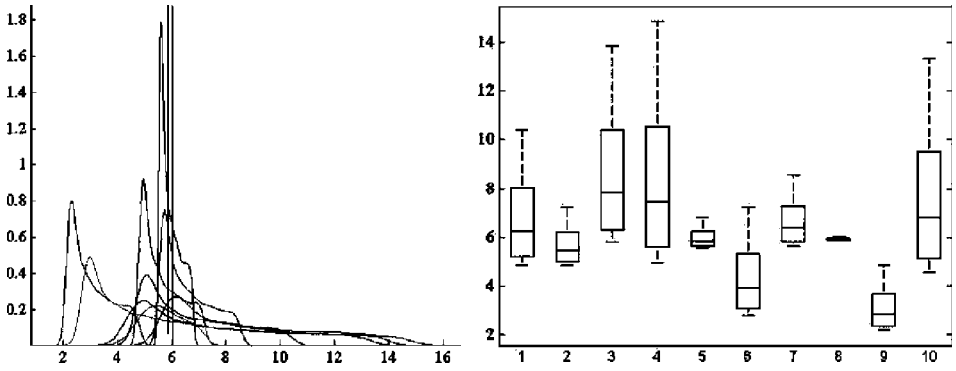

Figure 3: Characteristics of the point distributions of the second distributional variable Y_2

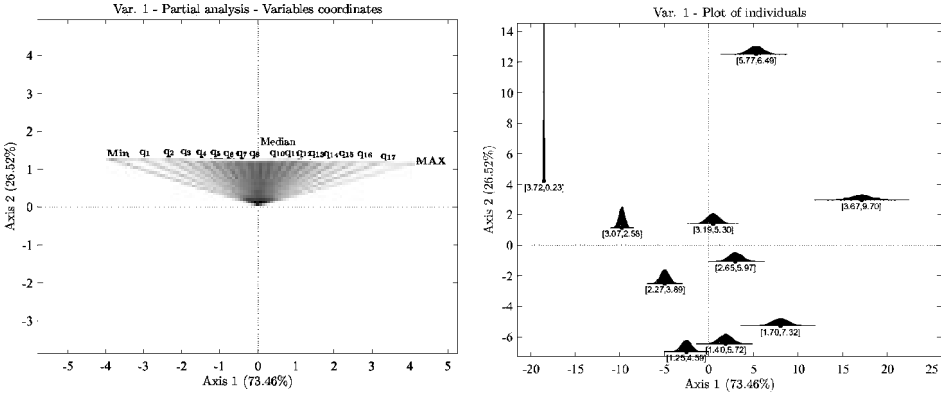

Figure 4: Representation of the quantile variables of Y_1 on the first factorial plan (Explained inertia 99.98%)

mean values, but are not scaled.

In this first step, MFA decomposes the covariance matrices $\mathbf{S_1}$ and $\mathbf{S_2}$, respectively. Consistent with the characteristics of the first distributional variable Y_1, the first latent factor is related to the variability of the standard deviations, as observed in the two plots in Fig. 4. Indeed, although the correlation is not strong with the central quantile-variables because all the Gaussians have the same mean (and median), we note that the first axis is strongly correlated with the extreme quantile-variables. Each distribution, suitably scaled horizontally and vertically, is placed at the point related to the individual, such that the mean corresponds to the abscissa of the point (right panel of Fig. 4). Following the first axis direction, it is worth noting that the distributions are ordered from lower to higher values of the standard deviations.

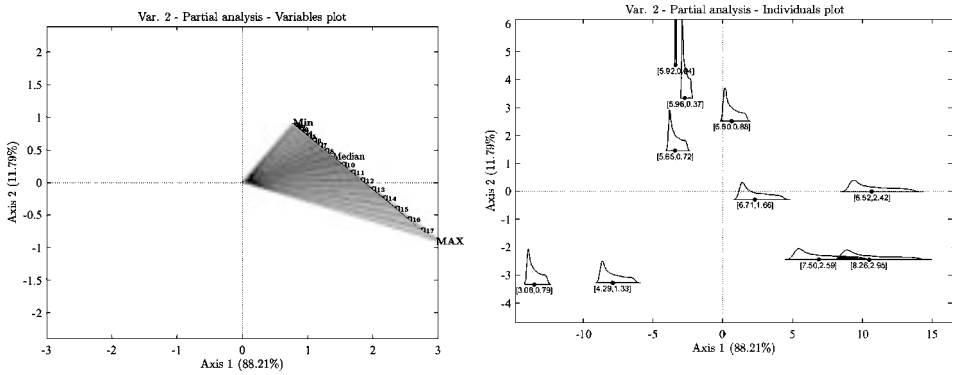

Figure 5: Representation of the quantile variables of Y_2 on the first factorial plan (Explained inertia=100%)

The partial PCA on the second set of quantile-variables associated with Y_2 is based on the decomposition of the covariance matrix denoted by $\mathbf{S_2}$. Figure 5 shows the representation of the variables by the *Spanish-fan* plot (on the left) and of the individuals by overlapping the distributions (on the right) on the first factorial plane.

Because the distributions are all skewed, the representation of the quantile-variables on the first plane (which explains 100% of the total inertia, with the first axis being 88.21%) is very different from the representation of the set of quantile variables associated with Y_1. Further, the shape of the *Spanish-fan* (the left panel of Fig. 5), a scalene triangle, is related to the fact that all the distributions are right skewed. Observing the representation of the individuals by their projected distributions (on the right side of Fig. 5), it is worth noting that along the first axis, the distributions are placed from the lower to the higher mean values (the first value between brackets at the bottom of each distribution) while the second factorial axis is the opposite(the second value between brackets).

The second step of the MFA is performed on the global matrix $\mathbf{Q}$. Figure 6 shows a simultaneous representation on the first factorial plane of the *Spanish-fans* of the two sets of quantile-variables (on the right side). The explained inertia of the first two factorial axes is 90.2%. Because in the partial analysis the first *Spanish-fan* of the set of quantile-variables in $\mathbf{Q_1}$ was strongly related to the variability component of the distributions (std), while the second *Spanish-fan* of the set of quantile-variables in $\mathbf{Q_2}$ was characterized by the values of the means on the first axis and by the values of the std on the second axis, in the global analysis, the first *Spanish-fan* plot appears rotated along the second dimension, which is related to the variability of the distributions, whereas the first axis is influenced by the values

of the means. This is also explained by the graphical representation (on the right side of Fig. 6) of the correlations between the factorial axes of the partial analyses and the ones from the global analysis.

The representation of the individuals on the first factorial plane is displayed in Fig. 7. The points labeled by numbers are the projection of the individuals on the first plane obtained with the MFA global phase. For interpreting the position of the individuals on the factorial plane, the respective distributions of the variables Y_1 and Y_2 are projected in supplementary.

In Fig. 8, a different representation of the individual on the first factorial plane is proposed. It is obtained by placing the distributions of each individual for the two variables in the same location points but one on the top and one at the bottom. This was possible because we had just two distributions for each individual.

Figure 9 shows the vectors corresponding to the two distributional variables. The correlation between the synthesis of the two sets of quantile-variables is expressed by the cosine of the angle on the factorial plane, according to the classical measure RV proposed by Escofier and Pagès (1988), which in this case is $RV = 0.2257$.

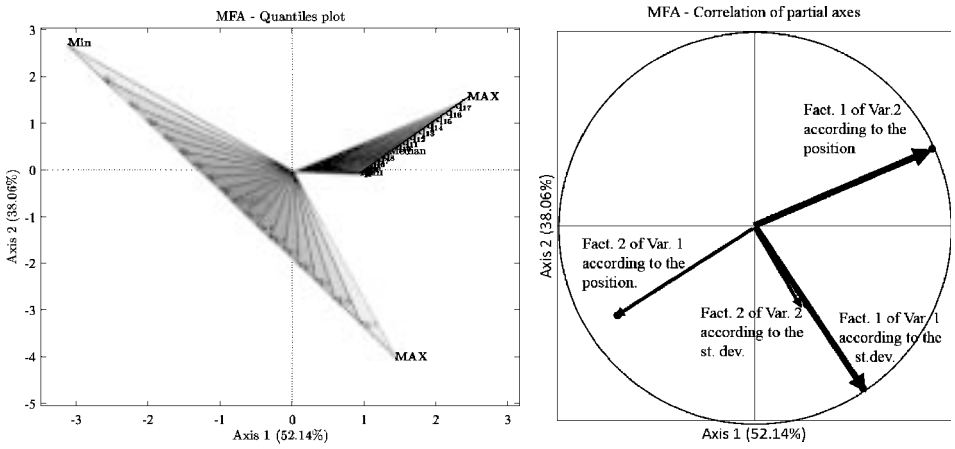

Figure 6: On the left side: Representation of the quantile variables of Y_1 and Y_2 on the first factorial plan (Explained inertia=90.2%).
On the right side: The circle of correlations between the factorial axes of the partial analyses and the factor axes of the global analysis.

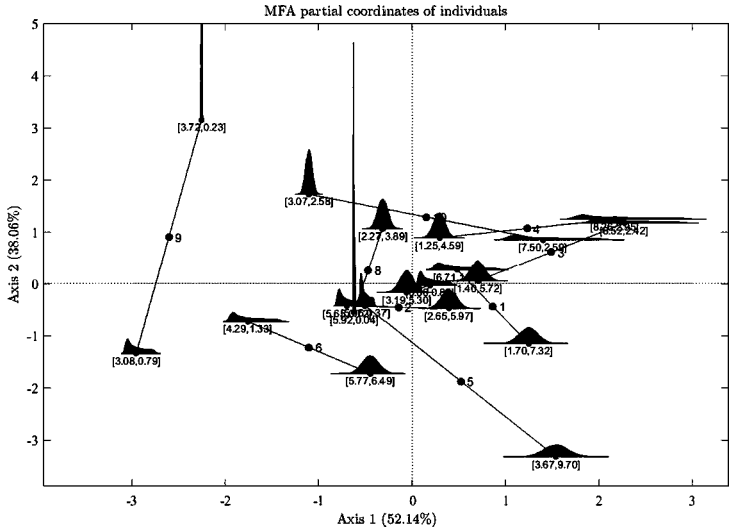

Figure 7: Representation of the individual distributions with respect to Y_1 and Y_2 on the first factorial plane (Explained inertia=90.2%). The distributions are drawn on the partial coordinates of individuals while the global coordinates of the individuals are labeled by the integers.

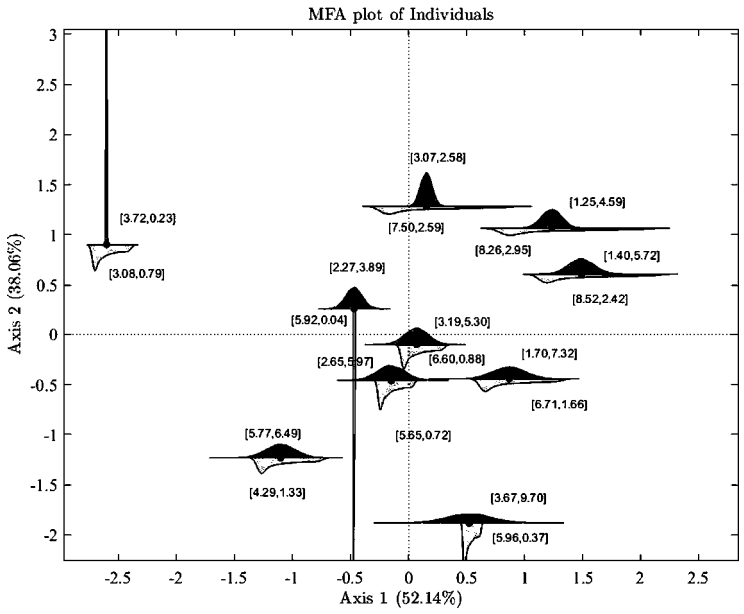

Figure 8: Representation of the individual distributions with respect to Y_1 and Y_2 on the first factorial plane (Explained inertia=90.2%). The distributions are drawn on the top and bottom of the global coordinates of the individuals.

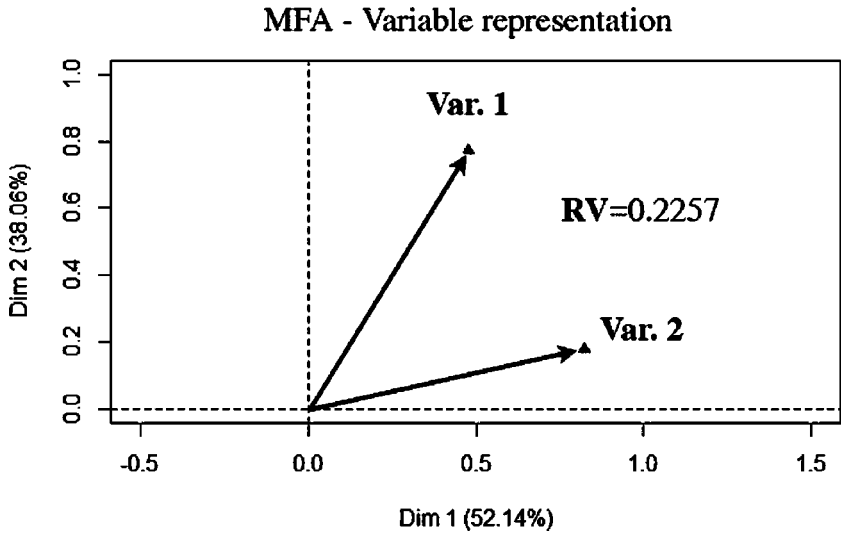

Figure 9: Correlation between the synthesis of the two sets of quantile-variables as expressed by the two vectors Var1 and Var2. RV=0.2257.

6. AN APPLICATION OF THE MFA ON REAL DATA

In this section, we present an application of the proposed MFA on a dataset described in (Billard and Diday (2006)). The dataset contains *Cholesterol*, *Hemoglobin* and *Hematocrit* levels observed for 14 groups of patients (each group is identified by a sex-age typology) using histograms of values. The size and the raw data of each group are not available, thus a classical PCA is not possible. The dataset is also available in the HistDAWass package[2] developed in R. The data table is shown in Fig. 10. The analysis is performed using 20 quantiles for each histogram. The MFA returns the components with their associated eigenvalues, as presented in Tab. 1. We note that the first two components synthesize 96.16% of the total variance, so we represent the main results using only the first factorial plane.

The variables are represented by *Spanish-fan* plots. Because each set of quantiles defines a block of variables in the MFA, we show the correlation plot of the *Spanish-fans* on the first factorial plane, which explains 92.54% of the total inertia. In Fig. 11, we show the *Spanish-fan* plots, while in Fig. 12, we observe the correlation between the axes of each partial PCA and each variable.

[2] `https://cran.r-project.org/web/packages/HistDAWass/index.html`

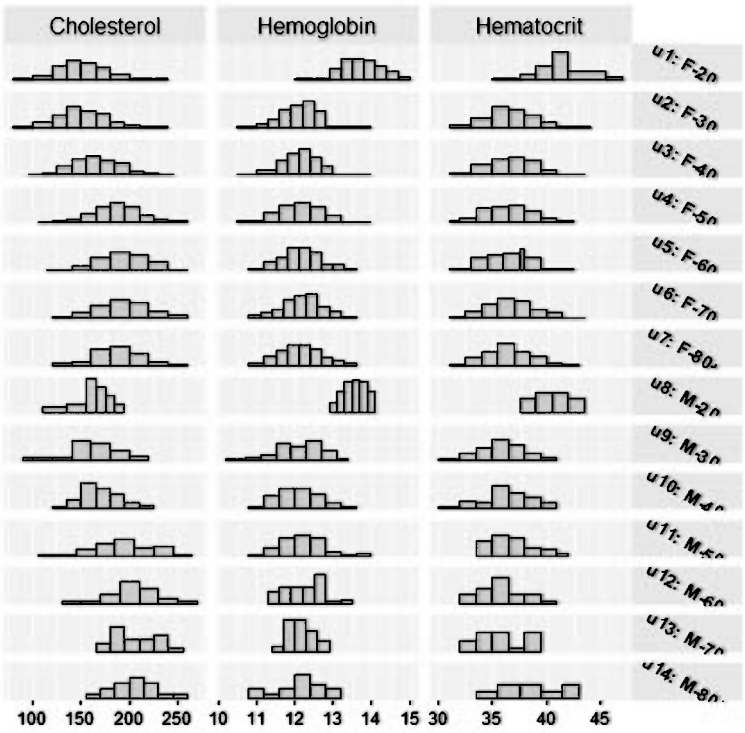

Figure 10: The BLOOD dataset

Table 1: BLOOD dataset: eigenvalues of each component

Components	Eigenvalue	% of variance	cum. % of variance
comp 1	2.28	71.55	71.55
comp 2	0.67	20.98	92.53
comp 3	0.12	3.63	96.16
comp 4	0.07	2.12	98.27
comp 5	0.02	0.58	98.85
comp 6	0.02	0.48	99.33
comp 7	0.01	0.25	99.59
comp 8	0.01	0.18	99.76
comp 9	0.00	0.13	99.89
comp 10	0.00	0.06	99.95
comp 11	0.00	0.03	99.98
comp 12	0.00	0.01	99.99
comp 13	0.00	0.01	100.00

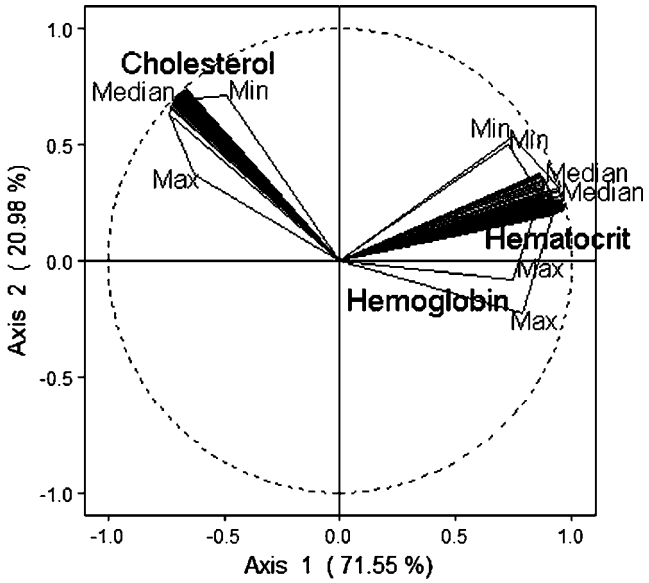

Figure 11: MFA first factorial plane: Spanish fan plots

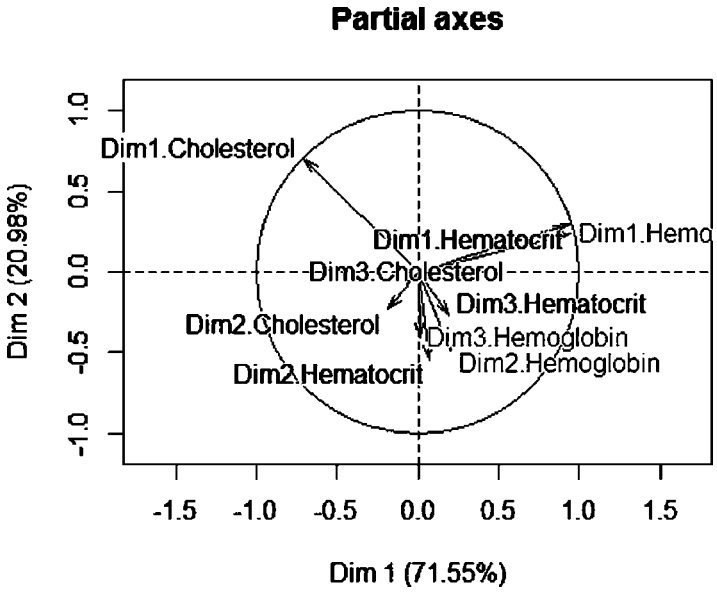

Figure 12: MFA first factorial plane: Correlation between the first three axes (dimensions DIM1, DIM2, and DIM3) of the partial PCA (for each distributional variable)

In Fig. 11, we note that the *Hemoglobin* and *Hematocrit Spanish-fans* almost overlap while the *Spanish-fan* of the *Cholesterol* variable is rather orthogonal to the others. Further, observing the spanning of the fans, it is worth noting that the distributions for the *Cholesterol* levels are less variable in scale (otherwise, the span should be more open) than those related for the *Hemoglobin* and *Hematocrit* levels. In Fig. 12, we observe that the MFA's first axis is positively correlated with the first axes of the partial analyses on the *Hemoglobin* and *Hematocrit* variables, while *Cholesterol* presents a higher correlation to the second MFA axis.

Further the second axes of the partial analyses present a low correlation on the MFA's first plane. Looking at both figures, we note that the first axes of the three partial analyses for each distributional variable, are oriented toward the direction of the central quantiles; thus, they are mainly related to the variability of positions. The other axes that are related to the variability of scales and shapes are associated with very small eigenvalues; thus, they poorly explain the variability of the distributional data.

The representation of individuals on the first factorial plane is performed by projecting the original distributions as supplementary variables on the factorial plane. The distributions are centred on the coordinates of each individual. To show the main characteristics of the individuals according their distribution for each variable, Figs. 13, 14, and 15 show the distributions for the *Cholesterol*, *Hemoglobin* and *Hematocrit* variables. Each plot is organized such that on the left, individuals are labeled according to their name, while on the right, individuals are labeled according to their mean value, and the darker the distributions are, the higher their mean is. In this way, Fig. 12, for each distributional variable, shows that the first axis opposes distributions with lower and higher mean values. Whereas the second axis opposes distributions according to lower and higher values of the scale and shape parameters.

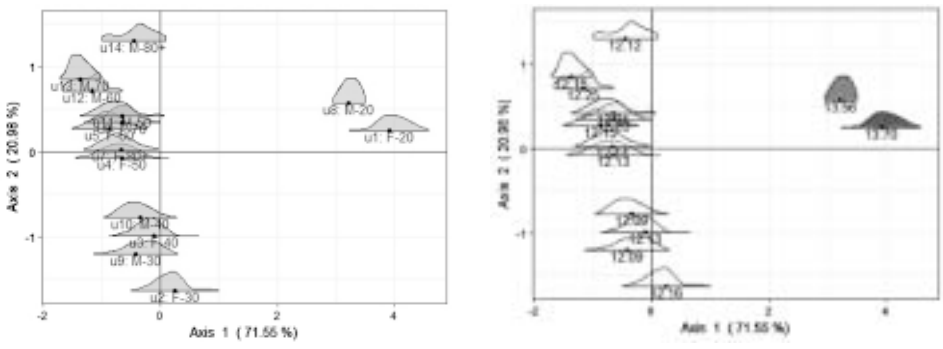

Figure 13: MFA first factorial plane: Plots of individuals for the *Cholesterol* variable, data are labeled with the object name on the left, and with the mean value on the right.

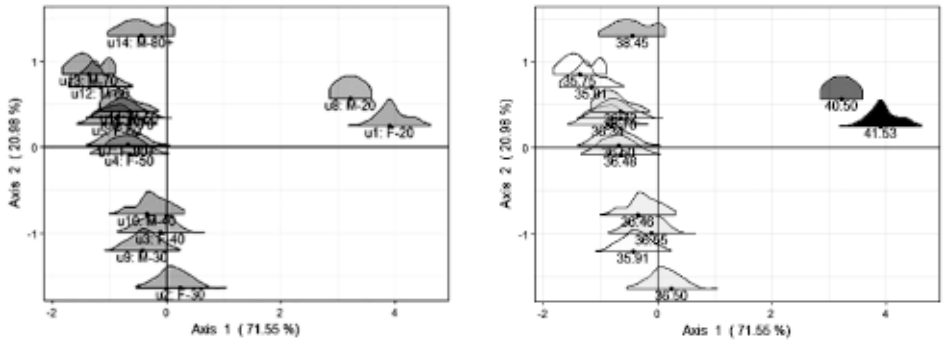

Figure 14: MFA first factorial plane: Plots of individuals for the *Hemoglobin* variable, data are labeled with the object name on the left, and with the mean value on the right.

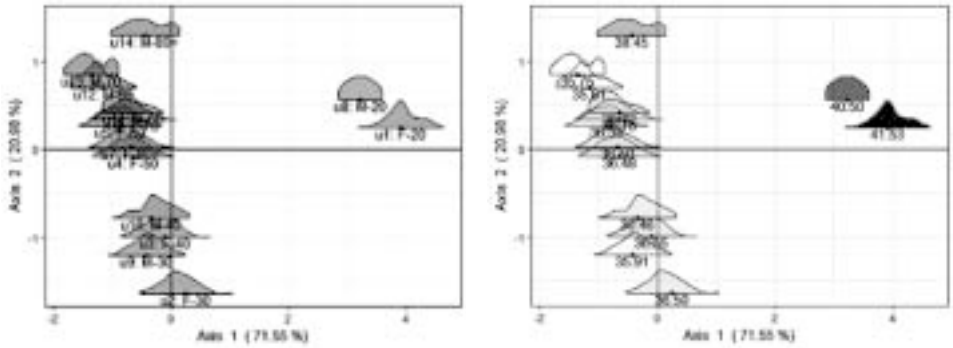

Figure 15: MFA first factorial plane: Plots of individuals for the *Hematocrit* variable, data are labeled with the object name on the left, and with the mean value on the right.

6.1 COMMENTS

As expected, being related to the quantity of iron in the blood cells, the *Hemoglobin* and *Hematocrit* are positively correlated to each other, and their values are higher in younger people. *Cholesterol* has a low correlation with the other two variables, and its mean value tends to increase from younger to older people. For the other scale and shape comparisons, there are very slight differences between the distributions. Thus the analysis is only slightly influenced by those aspects. It is possible to compare the scale and the shape of *M-20* and *F-20* on the factorial planes for the *Hemoglobin* and *Hematocrit* variables, observing that on the top of the

factorial plane, there are the distributions with low values for the standard deviation and kurtosis (Tab. 2).

Table 2: BLOOD dataset: kurtosis of each distribution computed as the fourth standardized moment.

Objects	Cholesterol	Hemoglobin	Hematocrit
u1: F-20	3.37	3.08	3.00
u2: F-30	3.23	4.10	2.72
u3: F-40	2.95	3.60	2.50
u4: F-50	3.18	2.81	2.51
u5: F-60	2.65	2.77	2.58
u6: F-70	2.65	2.91	2.61
u7: F-80+	3.04	2.72	2.73
u8: M-20	2.85	2.14	1.96
u9: M-30	3.16	2.83	2.73
u10: M-40	2.83	2.61	2.66
u11: M-50	2.74	3.44	2.49
u12: M-60	3.37	2.34	2.63
u13: M-70	1.92	2.37	1.86
u14: M-80+	2.56	2.42	1.94

7. CONLUSIONS

This paper represents an extension of the MFA for a PCA method for distributional data based on the ℓ_2 Wasserstein distance between distributions. We showed that the trace of the covariance matrix of the quantile-variables approximates the variance of a distributional variable computed with the Wasserstein metric. Previous approaches were not related to a particular metric between distributions; thus, a comparison could not be performed. Using quantile-variables, we observed that the proposed PCA enables us to identify the differences in the structure of the several sets of variables in the analysis according to the main characteristics of the distributional variables: position, scale, and shape. The classical MFA on standard data was enriched by the nature of the analyzed data. The characteristics of the observed distributions are emphasized by the peculiar tools for the interpretation. Further, a novel *Spanish-fan* plot was introduced to describe the relations among the quantile-variables projected on the factorial planes. We showed how to interpret the shape of a fan with respect to the characteristics of the distributions. Therefore, the similarity between the distributions (individuals in the analysis) is well interpreted according to the similarity between their parameters on each axis. The proposed applications on simulated and real data have shown how each axis is strongly related

to the variability of the parameters of position, scale, and shape. Aiming to show the advantages of the method and giving more readable factorial planes, only a few distributional variables were considered in the applications.

REFERENCES

Abdi, H., Williams, L. and Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. In *Wires Computational Statistics*, 5 (2): 97–179.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: conceptual statistics and data mining*. Wiley series in computational statistics. Wiley, Chichester, Hoboken (N.J.).

Bock, H.H. and Diday, E. (2000). *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Springer-Verlag Inc, Berlin; New York.

Cazes, P. (2002). Analyse factorielle d'un tableau de lois de probabilité. In *Revue de Statistique Appliquée*, L (3): 5–24.

Cazes, P., Chouakria, A., Diday, E. and Schektman, Y. (1997). Extension de l'analyse en composantes principales à des données de type intervalle. In *Revue de Statistique Appliquée*, XVI (3): 5–24.

Escofier, B. and Pagès, J. (1983). Methode pour l"analyse de plusieurs groupes de variables: application a la caràcterisation des vins rouges du val de loire. In *Revue de Statistique Appliquée*, 31: 43–59.

Escofier, B. and Pagès, J. (1988). *Analyses Factorielles Simples et Multiples: Objectifs, Methodes, Interprètation*. Dunod, Paris.

Escofier, B. and Pagès, J. (1990). Multiple factor analysis. In *Computational Statistics and Data Analysis*, 18: 121–140.

Gilchrist, W. (2000). *Statistical modelling with quantile functions*. Chapman & Hall/CRC, Boca Raton, Florida.

Ichino, M. (2008). Symbolic PCA for histogram-valued data. In *Proceedings IASC International Association of Statistical Computing*. Japanese Society of Computational Statistics, Yokohama, Japan.

Ichino, M. (2011). The quantile method for symbolic principal component analysis. In *Statistical Analysis and Data Mining*, 4 (2): 184–198.

Irpino, A. and Romano, E. (2007). Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation. In M. Noirhomme-Fraiture and G. Venturini, eds., *EGC*, vol. RNTI-E-9 of *Revue des Nouvelles Technologies de l'Information*, 99–110. Cépaduès-Éditions.

Irpino, A. and Verde, R. (2014). Basic statistics for distributional symbolic variables: a new metric-based approach. In *Advances in Data Analysis and Classification*, May 2014: 1–33. doi:10.1007/s11634-014-0176-4.

Lauro, N.C., Verde, R. and Irpino, A. (2008a). Generalized canonical analysis on symbolic objects. In M. Noirhomme-Fraiture and E. Diday, eds., *Symbolic data analysis and the SODAS software*, 313–330. John Wiley & Sons, Ltd.

Lauro, N.C., Verde, R. and Irpino, A. (2008b). Principal component analysis of symbolic data described by intervals. In M. Noirhomme-Fraiture and E. Diday, eds., *Symbolic Data Analysis and the SODAS Software*, 279–311. John Wiley & Sons, Ltd.

Le-Rademacher, J. (2008). *Principal Component Analysis for Interval-Valued and Histogram-Valued Data and Likelihood Functions and Some Maximum Likelihood Estimators for Symbolic Data*. Ph.D. thesis, Graduate School of The University of Georgia, Athens, Georgia, USA.

Lebart, L., Piron, M. and Morineau, A. (2006). *Statistique Exploratoire Multidimensionnelle*. Dunod, Paris.

Makosso-Kallyth, S. and Diday, E. (2012). Adaptation of interval PCA to symbolic histogram variables. In *Advances in Data Analysis and Classification*, 6 (2): 147–159.

Nagabhushan, P. and Kumar, R.P. (2007). Histogram PCA. In D. Liu, S. Fei, Z. Hou, H. Zhang, and C. Sun, eds., *Advances in Neural Networks ISNN 2007*, vol. 4492 of *Lecture Notes in Computer Science*, 1012–1021. Springer.

Pagès, J. (2014). *Multiple Factor Analysis by Example Using R*. Chapman & Hall/CRC, Boca Raton, Florida.

Rodriguez, O., Diday, E., and Winsberg, S. (2000). Generalization of the principal components analysis to histogram data. In *PKDD2000*. Lyon, France.

Rüschendorf, L. (2011). Wasserstein metric. In M. Hazewinkel, ed., *Encyclopedia of Mathematics*. Springer. URL http://www.encyclopediaofmath.org/index.php?title=Wasserstein_metric&oldid=32292.

Verde, R. and Irpino, A. (2008). Comparing histogram data using a Mahalanobis- Wasserstein distance. In P. Brito, ed., *COMPSTAT 2008*, 77–89. Physica-Verlag HD.

Verde, R., Irpino, A. and Balzanella, A. (2016). Dimension reduction techniques for distributional symbolic data. In *IEEE Transactions on Cybernetics*, 46 (2): 344–355. doi:10.1109/TCYB.2015.2389653.

Villani, C. (2003). *Topics in Optimal Transportation (Graduate Studies in Mathematics, Vol. 58)*. American Mathematical Society, Providence, Rhode Island.

Wang, H., Chen, M., Shi, X. and Li, N. (2014). Principal component analysis for normal-distribution-valued symbolic data. In *Cybernetics, IEEE Transactions on*, in press. doi:10.1109/TCYB.2014.2338079.