

**Fuzzy clustering of distribution-valued data using an adaptive $L_2$ Wasserstein distance**

Francisco de A.T. de Carvalho  
Centro de Informatica - CIn/UFPE  
Av. Jornalista Anibal Fernandes, s/n - Cidade Universitaria  
50.740-560, Recife-PE, Brazil  
Email:fatc@cin.ufpe.br

Antonio Irpino  
and Rosanna Verde  
Dept. of Political Sciences “J. Monnet  
Second University of Naples  
Caserta, CE, 81100, Italy

**Abstract**—In this paper, a fuzzy c-means algorithm based on an adaptive $L_2$-Wasserstein distance for histogram-valued data is proposed. The adaptive distance induces a set of weights associated with the components of histogram-valued data and thus of the variables. The minimization of the criterion in the fuzzy c-means algorithm is performed according three steps such that the representation, the allocation and the weights associated to the components of the variables are alternately computed until a the convergence of the solution to a local optimum. The effectiveness of the proposed algorithm is demonstrated through experiments with synthetic and real-world datasets.

I. INTRODUCTION

One of the current big-data age requirement is the possibility of represent groups of observed data by using summary structures that allows the minimum lose of information as possible. Empirical density estimations, like histograms, are useful tools for this aims. Examples of histogram representations can be find in the framework of image analysis, when the characteristics of images are described as distributions (or bar diagrams). Histograms can be also useful for privacy preserving matters, for representing the cash flows of a bank account, as well as for the dissemination of official statistics, usually expressed in aggregated form. A Histogram data formalization were introduced in the context of Symbolic Data Analysis (SDA) [1] as a particular set-valued description. A histogram variable is a variable whose observations are expressed by histograms.

Since the original definition, several data analysis techniques have been proposed for the treatment of these entities. More recently a new research field has emerged about distributional data, where the histogram data are analysed by using suitable metrics to compare distributions, as the family of Wasserstein distances [2]. The methods arose in such context present as peculiarity to work on the cumulative distribution functions and on the quantile functions associated to the histogram representations. Among them, a generalization of the Dynamic Clustering (DC) [3] has been proposed to partition a set of histogram data in homogeneous groups according to the characteristics of the distributions. The Dynamic Clustering, of which k-means is a particular case, is a two steps algorithm: it alternates a representation and an allocation step, such that a within homogeneity criterion is minimized. As usual in DC, the choice of a suitable distance plays a central role in the allocation and in the representation phases. Thus, consistently with the nature of the data, a Wasserstein distance (in norm $L_2$) has been firstly proposed by [4]–[7]. These authors posed also the problems of the representation of the clusters by prototypes as means of the distributions in the clusters, according to the Wasserstein metrics. More recently, [8] proposed a k-means clustering method using empirical joint distributions. [9] introduced a Dynamic Clustering algorithm based on the copula analysis aiming to take into account the relationship between the histogram variables.

A main issues in clustering analysis is to consider the different contribution of the several variables in the clustering process according to their characteristics that, in the case of distributions, are related to the location, variability and shapes of the observed data. Generally, clustering methods do not take into account the different relevance of the variables in the analysis, i.e., the algorithms consider all variables equally important in the clustering process. However, in most applications, some variables may be more discriminant of the clusters than others; in some other applications each cluster may have a different set of more relevant variables to group together the cluster data. In order to tackle this issue, in the clustering of classical data, [10] proposed to integrate adaptive distances. The use of adaptive distances in the clustering algorithm consists in introducing a weighting step in the optimization process. In this step a set of weights are obtained minimizing the total sum of squares criterion. Such weights are associated with each variable (for all the clusters or for each cluster) and represents a measure of the importance of a variable in the clustering process.

In the framework of SDA, [11]–[13] proposed several adaptive distances, based on Hausdorff, City-Block and Euclidean distances in dynamic clustering algorithm of set-valued data. A more recent work [14] extends these approaches to histogram data clustering, using an adaptive distance based on the $L_2$ Wasserstein metric. The authors propose two novel adaptive distances based on clustering schemes able to compute automatically the relevance of each histogram variable during the partitioning of the data set. Then, using a decomposition of the $L_2$ Wasserstein distance [15] and considering the variability measure introduced in [7], the distance between two histograms can be shared in two components: one related to the variability of averages of the histograms and the
second related to the shapes of the histograms. In such a way, the $L_2$ Wasserstein distance measures the diversity of two distributions according to two additive components of variability. Thus, adaptive distances take into account the two components of the variability of a set of histograms and the algorithm estimates two sets of weights for each variable and each component. In a local approach the authors consider also a different set of weights for each cluster. In [13] it is also presented a clustering algorithm with automatic weighting of the histogram variables. However, this approach is based on an Euclidean distance between two sets of weights related to a particular pre-processing of the set-valued data.

The most clustering algorithms proposed for histogram data (adaptive distance based or not) are hard clustering methods, that means they partition a set of observed data in a predefined $K$ number of no-overlapped clusters. Therefore, each histogram belongs to only one cluster according to the characteristics of their distributions (i.e.: mean value, variability and shape). Using the $L_2$ Wasserstein distance all the characteristics of pair of distributions to be compared are taken into consideration. However, particular structure of the observed histogram data could give clusters not well separated and with a high internal variability due to the presence of some data that are forced to belong to only one cluster. In presence of this kind of problem, a more appropriate algorithm is the fuzzy clustering. According to that, an observation can be assigned to more than one cluster with a membership degree that expresses the similarity of this element to the representative element (prototype) of each cluster. Usually the membership degrees of an observation to the several clusters are valued in $[0,1]$ and, on all clusters, sum to 1 [16].

The present paper introduces a fuzzy c-means (FCM), in a more general scheme of Dynamical Clustering algorithm, for histogram data, based on the $L_2$ Wasserstein distance, denoted as: Fuzzy c-means with non adaptive $L_2$ Wasserstein distance (FCM-D). A second algorithm is also proposed as the FCM-D extension to the adaptive $L_2$ Wasserstein distance [14], denoted as: Fuzzy c-means with adaptive $L_2$ Wasserstein distance (AFCM-D).

Taking into consideration the $L_2$ Wasserstein distance decomposition in two additive components [7] we propose adaptive distances that take into account the two components of the variability of a set of histograms. We propose to associate two sets of weights to each variable and to each component. The two sets are locally estimated for each cluster of the partition.

The proposed fuzzy clustering algorithm, based on adaptive distances, is an alternating three steps procedure that estimates, step by step, the membership values of the observed distributions to the clusters, the weights for each variable and each component, as well as the cluster prototypes.

This paper is organized as follows: in Section II, we introduce the definitions of histogram data and the Wasserstein distance between histograms. Sections III and IV present, respectively, the Fuzzy c-means with non adaptive $L_2$ Wasserstein distance (FCM-D) and the Fuzzy c-means with adaptive $L_2$ Wasserstein distance (FCM-D) Section V shows the results achieved by the proposed algorithms on synthetic and real data sets. A comparison is also proposed with the hard Fuzzy c-means with non adaptive and adaptive distances. Section VI concludes the paper with some final considerations and open perspectives.

II. HISTOGRAM DATA AND WASSERSTEIN DISTANCE

Histogram allows a suitable summary of huge data or confidential data that cannot be available as individual values. SDA formalized histogram data as realizations of a histogram variable (a special case of modal numerical multivalued variable). In this case, a continuous variable $Y$ is a histogram-valued variable if each observation $i$ is expressed by a frequency distribution in form of histogram [1]. Formally, let $y_i$ a realization of $Y$ such that $S(i) = [\min(y); \max(y)] \subset \mathbb{R}$ is the support, that is partitioned into a set of contiguous and no-overlapped intervals (bins) $\{I_{1i}, \ldots, I_{hi}, \ldots, I_{Ni}\}$ (where $I_{hi} = [a_{hi}; b_{hi}]$ with $\min(y) = a_{1i}$ and $\max(y) = b_{Ni}$) and each $I_{hi}$ is associated with a (non negative) weight $\pi_{hi}$ that represents an empirical (or theoretical) relative frequency.

Hereafter, we denote with $f_i(y)$ the empirical density function associated with the description $y_i$, with $F_i(y)$ the cumulative distribution function and with $Q_{it}(t)$ the quantile function. It is possible to define the $i-th$ histogram, realization of the variable $Y$, as:

$$y_i = [(I_{1i}, \pi_{1i}), \ldots, (I_{ui}, \pi_{ui}), \ldots, (I_{Hi}, \pi_{Hi})]$$

such that

$$\forall I_{ui} \in S(i), \pi_{ui} = \int_{I_{ui}} f_i(y)dy \geq 0 \text{ and } \int_{S(i)} f_i(y)dy = 1.$$ 

In the following, we use $y_i$ to denote the histogram associated with the $i-th$ unit when a single histogram variable is observed. If we observe $p$ variables, we denote with $y_{ij}$ (for $i = 1, \ldots, n$ and $j = 1, \ldots, p$) the histogram of the values of the $j-th$ variable taken by the $i-th$ unit.

Thus, considering the classic data analysis approach, the individuals $\times$ variables input data table contains in each cell a histogram.

The comparison of histogram data is a particular case of comparison of distribution functions. Several distances between distribution functions [17] can be used for comparing the density or frequency distributions of two random variable. Among them the family of distances based on Wasserstein metric [2] permits to obtain interesting interpretable results about the characteristics of the distributions (see [7] for details).

According to [2], the $L_2$ squared Wasserstein distance between the qsf associated with the two (univariate) histograms is:

$$d_2^{2}(y_i, y_{i'}) = \int_0^1 [Q_i(t) - Q_{i'}(t)]^2 dt. \quad (1)$$

If the quantile functions $Q_i(t)$ and $Q_{i'}(t)$ are centered about their respective means, i.e., $Q_i'(t) = Q_i(t) - \bar{y}_i$ and $Q_{i'}'(t) = Q_{i'}(t) - \bar{y}_{i'}$. 

For all the other cases [18], in general it is not possible to obtain analytically a Wasserstein distance between two multivariate distributions. In this paper, we assume that no knowledge exists about the joint histograms for each individual, but it is known only the simple distributions expressed by the histograms for each variable. So, for the Euclidean distance, we consider the multivariate squared $L_2$ Wasserstein distance as follows:

$$d^2_W (y_i, y_i') = \sum_{j=1}^{p} d^2_W (y_{ij}, y_{i'j}).$$

(3)

In order to consider the variables with a different weight, we generalize the concept of adaptive distances [10] to the $L_2$ Wasserstein distance. Let us consider a vector of weights $\lambda = [\lambda_1, \ldots, \lambda_p]$ such that $\lambda_j > 0$. According to [11] and [10], a general formulation for an Adaptive Single Variable (squared) Wasserstein distance is:

$$d^2_W (y_i, y_i' | \lambda) = \sum_{j=1}^{p} \lambda_j d^2_W (y_{ij}, y_{i'j}).$$

(4)

In this formulation, the weights induce a linear transformation of the original space. Several approaches have been proposed (see for examples [13], [12]), where the weights are associated to global partition of the set of data or they are local weights on each cluster in which it is partitioned the set of data. Because we have shown that the $L_2$ Wasserstein distance can be decomposed in two components, a proposal by [14] is to introduce a suitable system of weights on such of the components: the first approach consists to computed the weights for the variables, defining the adequacy criterion on a Globally Component-wise Adaptive Wasserstein Distance (GC-AWD); in the second, the weights are computed for each component of the Wasserstein Distance, defining the adequacy criterion on a Cluster Dependent Component-wise Adaptive Wasserstein Distance (CDC-AWD).

### III. Fuzzy c-means with non-adaptive $L_2$ Wasserstein Distance

This section introduces the fuzzy c-means clustering algorithms with non-adaptive $L_2$ Wasserstein distances (hereafter denoted FCM-D).

#### A. Objective function of FCM-D

The set $E$ of $n$ individuals, described by $y_i = (y_{i1}, \ldots, y_{ip})$ is clustered according to a FCM method.

Each fuzzy cluster $C_k (k = 1, \ldots, K)$ has a representative or prototype $g_k = (g_{k1}, \ldots, g_{kp})$, where each $g_{kj}$ is associated with an (empirical) histogram $f_k(y_j)$, a distribution function $F_k(y_j)$ and a quantile function $Q_k(y_j)$.

This FCM algorithm aims to provide a fuzzy partition $P = (C_1, \ldots, C_K)$ of $E$ into $K$ fuzzy clusters, represented by a matrix of membership degrees $U = (u_{11}, \ldots, u_{nk})$ with $u_{ik} = (u_{i1}, \ldots, u_{in})$; and a matrix of prototypes $G = (g_{11}, \ldots, g_{Kp})$ representing the fuzzy clusters in the fuzzy partition $P$, by (locally) minimizing a suitable objective function. This function measures the adequacy between clusters and their representatives and it is defined as:

$$J(G, U) = \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m d_W (y_i, g_k)$$

(5)

$$= \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m \sum_{j=1}^{p} (y_{ij} - \bar{y}_{g_kj})^2$$

$$+ \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m \sum_{j=1}^{p} d^2_W (y_{ij}, g_{kj})$$

(6)

in which

$$d(y_i, g_k) = \sum_{j=1}^{p} (\bar{y}_{ij} - \bar{y}_{g_kj})^2 + \sum_{j=1}^{p} d^2_W (y_{ij}, g_{kj})$$

is the non-adaptive $L_2$ Wasserstein distances computed between the object $e_i$ and the prototype $g_k$ of the fuzzy cluster $C_k$. $u_{ik}$ is the membership degree of the object $e_i$ on the fuzzy cluster $C_k$ and $m \in ]1, +\infty[ $ is a parameter that controls the fuzziness of membership for each object $e_i$.

#### B. The FCM-D algorithm

From an initial solution, the algorithm iterates two steps until the convergence to a local minimum. These steps are the following:

1) Computation of the prototypes: The first step consists in the computation of the matrix of prototypes $G$. With $U$ fixed, by taking the derivative of $J$ with respect to the prototypes, the components of the matrix of prototypes $G$ are computed from the following optimization problem:

$$\sum_{i=1}^{n} (u_{ik})^m (\bar{y}_{ij} - \bar{y}_{g_kj})^2 + \sum_{i=1}^{n} (u_{ik})^m d^2_W (y_{ij}, (g_{kj})^c) \rightarrow \text{Min}$$
Setting this derivative to zero, they are calculated by means of the quantile functions associated with each histogram $g_k$:

$$Q_{g_k}^c (t_j) + \frac{n}{\sum_{i=1}^{n} u_{ik}^m} F_{i}^{-1} (t_j) = 0$$

(7)

2) Computation of the membership degrees: With $G$ fixed, the second step computes the matrix of membership degrees $U$ assuming the following constraints:

$$\sum_{k=1}^{K} u_{ik} = 1, \ for \ k = 0, 1$$

Let $A = \{ k \in \{ 1, \ldots, K \} : d (y_i, g_k) = 0 \}$.

- If $A = \emptyset$ (i.e., no object coincides with any of the representatives), the minimization of the criterion $J$ is obtained from the method of Lagrange Multipliers:

$$\mathcal{L} = \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m d (y_i, g_k) - \sum_{i=1}^{n} \theta_i \left( \sum_{k=1}^{K} u_{ik} - 1 \right)$$

By setting the derivatives of $\mathcal{L}$ with respect to $u_{ik}$ and $\theta_i$ to zero, we obtain the components of the matrix $U$ of membership degrees:

$$u_{ik} = \left[ \sum_{k=1}^{K} \left( \frac{d (y_i, g_k)}{d (y_i, g_k)} \right)^{1/m} \right]^{-1}$$

(8)

where

$$d (y_i, g_k) = \sum_{j=1}^{p} (\bar{y}_{ij} - \bar{y}_{g_k})^2 + \sum_{j=1}^{p} d_W^2 (y_{ij}', (g_k)')$$

- If $A \neq \emptyset$ then

$$u_{ik} = \frac{1}{|A|}, \ \forall k \in A
\quad u_{is} = 0, \ \forall s \notin A$$

(9)

These two steps are repeated until the convergence is obtained, i.e., the $J$ value change is small. The Algorithm 1 summarizes these steps.

**Algorithm 1** The FCM-D algorithm

**Input:** $D$, $K$, $T$, $0 < \epsilon < < 1$ and $m$

**Initialization:** $t \leftarrow 1$; random initialization of $G^{(t)}$, $U^{(t)}$

**Repeat**

1. $t \leftarrow t + 1$; Compute $G^{(t)}$ using equation (7)
2. Compute $U^{(t)}$ using equation (8) (or equation (9)); Compute $J (G^{(t)}, U^{(t)})$

**Until** $|J (G^{(t)}, U^{(t)}) - J (G^{(t-1)}, U^{(t-1)})| < \epsilon$ or $t > T$

**Output:** $G$, $U$

IV. Fuzzy C-means with Adaptive L2 Wasserstein Distance

This section introduces the fuzzy c-means clustering algorithms with adaptive L2 Wasserstein distances (hereafter denoted AFCM-D).

A. Objective function of AFCM-D

This clustering algorithm aims to provide a fuzzy partition $P = \{ C_1, \ldots, C_K \}$ of $E$ into $K$ fuzzy clusters, represented by a matrix of membership degrees $U = \{ u_{11}, \ldots, u_{nk} \}$ with $u_i = \{ u_{i1}, \ldots, u_{ik} \} (i = 1, \ldots, n)$, a matrix of prototypes $G = \{ g_1, \ldots, g_K \}$ representing the fuzzy clusters in the fuzzy partition $P$, and a matrix of relevance weights $\Lambda$, by (locally) minimizing a suitable objective function. This function measures also the adequacy between the clusters and their respective representatives and it is now defined as:

$$J_A (G, \Lambda, U) = \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m d (\lambda_{k, g_k}, \lambda_{k, C_k} (y_i, g_k)^{c})$$

(10)

$$= \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m \sum_{j=1}^{p} \lambda_{k, g_k} (y_{ij} - \bar{y}_{g_k})^2$$

$$+ \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m \sum_{j=1}^{p} \lambda_{k, C_k} d_W^2 (y_{ij}', (g_k)')$$

in which

$$d (\lambda_{k, g_k}, \lambda_{k, C_k} (y_i, g_k)) = \sum_{j=1}^{p} \lambda_{k, g_k} (y_{ij} - \bar{y}_{g_k})^2$$

(11)

is the adaptive L2 Wasserstein distances computed between the object $e_i$ and the prototype $g_k$ of fuzzy cluster $C_k$, parameterized by the relevance weight vectors $\lambda_{k, g_k} = (\lambda_{k1, g_k}, \ldots, \lambda_{kp, g_k})$ and $\lambda_{k, C_k} = (\lambda_{k1, C_k}, \ldots, \lambda_{kp, C_k})$ that assign weights to the components of the distribution-valued data and thus to the variable in each cluster. Furthermore, the matrix of relevance weights is

$$\Lambda = \begin{bmatrix} \Lambda_{1, g_k} & \Lambda_{1, C_k} \\ \vdots & \vdots \\ \Lambda_{K, g_k} & \Lambda_{K, C_k} \end{bmatrix}$$

B. The AFCM-D algorithm

From an initial solution, the algorithm iterates three steps until the convergence to a local minimum. These steps are the following:

1) Computation of the prototypes: The first step consists in the computation of the matrix of prototypes $G$. With $\Lambda$, $U$ fixed, by taking the derivative of $J_A$ with respect to the prototypes, the components of the matrix of prototypes $G$ are computed again from the following optimization problem:

$$\sum_{i=1}^{n} (u_{ik})^m (\bar{y}_{ij} - \bar{y}_{g_k})^2 + \sum_{i=1}^{n} (u_{ik})^m d_W^2 (y_{ij}', (g_k)') \rightarrow \text{Min}$$

(12)
Setting this derivative to zero, they are calculated by means of the quantile functions associated with each histogram \( g_{kj} \) according to equation (7).

2) Computation of the relevance weights: With \( G \) and \( U \) fixed, this step aims the computation of the matrix of relevance weights \( \Lambda \) assuming the following constraints:

\[
\prod_{j=1}^{p} \lambda_{kj,\bar{y}} = 1, \quad \lambda_{kj,\bar{y}} > 0
\]

\[
\prod_{j=1}^{p} \lambda_{kj,Disp} = 1, \quad \lambda_{kj,Disp} > 0.
\]

The choice of these constraints is done accordingly to [10], where the product is used in order to consider the relevance of each variable with respect to the volume of the ellipses containing the data, thus the product of the magnitude of their axes. The minimization of the criterion \( J_A \) is obtained from the method of Lagrange Multipliers:

\[
L_{1A} = \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m d(y_i, g_k)
\]

\[
- \sum_{k=1}^{K} \theta_{1i} \left( \prod_{j=1}^{p} \lambda_{kj,\bar{y}} - 1 \right)
\]

\[
- \sum_{k=1}^{K} \theta_{2i} \left( \prod_{j=1}^{p} \lambda_{kj,Disp} - 1 \right)
\]

By setting the derivatives of \( L_{1A} \) with respect to \( \lambda_{kj,\bar{y}}, \theta_{1i}, \lambda_{kj,Disp} \) and \( \theta_{2i} \) to zero, we obtain the components of the matrix \( U \):

\[
\lambda_{kj,\bar{y}} = \left[ \prod_{h=1}^{p} \sum_{i=1}^{n} (u_{ih})^m (\bar{y}_{ih} - \bar{y}_{gh})^2 \right]^{\frac{1}{m}}
\]

(12)

and

\[
\lambda_{kj,Disp} = \left[ \prod_{h=1}^{p} \sum_{i=1}^{n} (u_{ih})^m d_w^2(y_{ih}^c, y_{gh}^c) \right]^{\frac{1}{m}}
\]

(13)

3) Computation of the membership degrees: With \( G \) and \( \Lambda \) fixed, the last step computes the matrix of membership degrees \( U \) assuming again the following constraints:

\[
\sum_{k=1}^{K} u_{ik} = 1, \quad u_{ik} \in [0, 1]
\]

Let \( A = \{k \in \{1, \ldots, K\} : d(\lambda_{kj,\bar{y}},\lambda_{kj,Disp}) = 0\} \).

- if \( A \neq \emptyset \) (i.e., no object coincides with any of the representatives), the minimization of the criterion \( J_A \) is obtained from the method of Lagrange Multipliers:

\[
L_{2A} = \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m d(\lambda_{kj,\bar{y}},\lambda_{kj,Disp})(y_i, g_k)
\]

\[
- \sum_{i=1}^{n} \theta_i \left( \sum_{k=1}^{K} u_{ik} - 1 \right)
\]

By setting the derivatives of \( L_{2A} \) with respect to \( u_{ik} \) and \( \theta_i \) to zero, we obtain the components of the matrix \( U \) of membership degrees:

\[
u_{ik} = \left[ \sum_{h=1}^{K} \frac{d(\lambda_{kh,\bar{y}},\lambda_{kh,Disp})(y_i, g_h)}{d(\lambda_{h,\bar{y}},\lambda_{h,Disp})(y_i, g_h)} \right]^{-1}
\]

(14)

where

\[
d(\lambda_{kh,\bar{y}},\lambda_{kh,Disp})(y_i, g_k) = \sum_{j=1}^{p} \lambda_{kj,\bar{y}} (\bar{y}_{ij} - \bar{y}_{gk})^2
\]

\[
+ \sum_{j=1}^{p} \lambda_{kj,Disp} d_w^2(y_{ij}^c, y_{gk}^c)
\]

- if \( A \neq \emptyset \)

\[
\left\{ \begin{array}{l}
  u_{ik} = 1/|A|, \forall k \in A \\
  u_{is} = 0, \forall s \notin A
\end{array} \right.
\]

(15)

These three steps are repeated until the convergence is obtained, i.e., the \( J_A \) value change is small. The Algorithm 2 summarizes these steps.

Algorithm 2 The AFCM-D algorithm

\textbf{input:} \( D, K, T, 0 < \varepsilon << 1 \) and \( m \)

\textbf{initialization:} \( t \leftarrow 1; \) random initialization of \( G^{(t)}, \Lambda^{(t)}, U^{(t)} \)

\textbf{repeat}

(1) \( t \leftarrow t + 1; \) Compute \( G^{(t)} \) using equation (7)

(2) Compute \( \Lambda^{(t)} \) using equations (12) and (13)

(3) Compute \( U^{(t)} \) using equation (14) (or equation (15)); Compute \( J(G^{(t)}, \Lambda^{(t)}, U^{(t)}) \)

\textbf{until} \( |J(G^{(t)}, \Lambda^{(t)}, U^{(t)}) - J(G^{(t-1)}, \Lambda^{(t-1)}, U^{(t-1)})| < \varepsilon \) or \( t > T \)

\textbf{output:} \( G, \Lambda, U \)

V. EXPERIMENTAL RESULTS

To show the usefulness of the proposed fuzzy clustering methods we considered different configurations of sets of histogram-valued data according to a different number of clusters having different position, scale and shape described by two histogram variables. Applications are considered also for real data. Our aim is to compare the hard and the fuzzy clustering algorithm with the adaptive distance-based ones. For synthetic data sets, external validity indices are used to compare the partitions returned by the algorithms. The \textit{a priori} classes are defined according to the model generating the data. For comparing the agreement between the obtained fuzzy partitions and the \textit{a priori} classes we use the modified version of the Rand, Jaccard, Folkes-Mallows and Hubert indices for fuzzy clustering algorithms proposed in [20]. For real valued data, we analyze a histogram-valued dataset where data are histograms: population pyramids by sex for 228 countries in the year 2014. For choosing an optimal number of cluster, considering the main results of [21], a survey paper on validity indices for fuzzy c-means algorithms, and we used the \( V_{MPC} \)
A way for modelling the quantile function of a position parameter, $\lambda$, skewness parameter taking value in [0,1], was introduced by Gilchrist [25].

The generation of the quantile functions was done by using a model of quantile functions such that each couple of quantile functions belongs to two different distributional variables. It generalizes the ratio of the between sum of squared distances and the total sum of squared distances, thus, it is a compactness index of a fuzzy partition.

### A. Synthetic data

Wasserstein distance, even it was used for density functions, can be see as a norm between two quantile functions. With this in mind, we have generated several datasets of couples of quantile functions such that each couple of quantile functions belongs to two different distributional variables. The generation of the quantile functions was done by using a model of quantile function proposed by Gilchrist [25].

Denoting with $Q(p)$ the quantile observed for a level of $p \in [0,1]$, Ref. [25] introduced a way for modelling the quantile function of a skew logistic distribution, depending on three parameters: $\lambda$ a position parameter, $\eta > 0$ a scale parameter, and $\delta$ a skewness parameter taking value in $[-1;1]$ (negative, resp. positive, values are associated to left-skewed, resp. right-skewed distributions). The explicit formula of such quantile function is:

$$Q(p) = \lambda + \eta \left[1 - \delta \left(\ln(p) - \ln(1 - p)\right)\frac{1 + \delta}{2}\right].$$

In Fig. 1 and Fig. 2 are shown the quantile functions and the associated density functions of three examples of skew logistic distributions. For the sake of brevity, we considered a number of clusters $k = 4$ only, each one having 50 objects described by two quantile functions. For avoiding numerical problems related to the extreme values we considered 51 equally spaced quantiles from $p = 0.01$ to $p = 0.99$ that well approximate the theoretic quantile functions. This approximation is required for computing the results into a reasonable time without a significantly loss of precision. The settings of the experiment is shown in Tab. I, where we reported the intervals in which the parameters have been sampled. For each dataset we repeated 10 times the standard dynamic clustering method (DC-D) for histogram data, the adaptive-distance based dynamic clustering (ADC-D), the fuzzy c-means (FCM-D) and the adaptive-distance based fuzzy c-means (AFCM-D).

Considering the different dispersion structure of the parameters generating the quantiles (see the length of the intervals for the parameters in Tab. I), we observe that in general algorithms based on adaptive distance have better performance with respect to those based on the standard Wasserstein distance. Even if the DC-D and the FCM-D have similar performances in terms of agreement, the difference is greater for the adaptive distance based algorithms, where the AFCM-D seems the best.

### B. Real data

For testing the proposed algorithm on a real-world dataset, we considered population age-sex pyramids data collected by...
the Census Bureau of USA in 2014. A population pyramid is a common way to represent jointly the distribution of sex and age of people living in a given administrative unit (city, region or country, for instance). In this dataset, each country is represented by two histograms describing the age distribution for the male and the female population. Both distributions are represented by vertically juxtaposing, and the representation is similar to a pyramid. The shape of this pyramids vary according to the distribution of the age in the population that is considered as a consequence of the development of a country. In Fig. 3 is shown the age pyramid of the World. For discovering a number of clusters for the fuzzy partitioning of the 228 countries, after fixing \( m = 2 \) as fuzzyness parameter, \( \varepsilon = 10^{-5} \), we used the \( V_{M_{PC}} \) for \( c = 2, \ldots, 8 \). In Table III, we observe that using the \( V_{M_{PC}} \) index a good number of cluster is \( c = 3 \) for both the FCM-D and the AFCM-D algorithm. We reported also the \( QPI \) indices giving a measure of the compactness of the clusters based on the variability of data. Observing \( V_{M_{PC}} \), AFCM-D returns partitions that are more compact compared with FCM-D, even if the QPI is slightly higher for FCM-D. However, a comparison between two \( QPI \)'s computed using two different distances (adaptive or not) can be questionable.

For the sake of brevity, we report only the results obtained for AFCM-D algorithm. In Fig. 4 are represented the \( c = 3 \) mean histograms for the observed variables.

In Tab. IV we reported the first ten countries (ordered by their membership value) to each cluster, while in Tab. V are shown the most ambiguous countries, i.e., those countries whose population structure by age is in transition from a cluster to another. In Tab. VI are reported the weights for each variable component and for each cluster. We note that the male average age has an higher importance in the distance computation compared to the female one for clusters 1 and 2, while the female dispersion of age is more relevant for cluster 1 and 3 compared to the male one. In Fig. 4 are shown the centers of clusters described by their distributions for the male and female age population density. It is easy to see that the discovered clusters correspond to well known population structures. Sex-age population densities that are very right-skewed (like Cluster 3) show populations with a high number of young dependants and a low life expectancy. Population that have fairly uniform densities (like Cluster 2) show populations with a falling birth rate and a rising life expectancy, but also with an high rate of juvenile emigration (like in Poland and Romania). Over time, as a country develops, the shape changes from triangular to barrel-like where the intermediary situation are shaped like Cluster 1. In this cluster is easy to find countries into a developing phase like Brazil and Turkey.

VI. Conclusion

Two new algorithms for fuzzy clustering a set of distribution-valued data have been presented. The first one, FCM-D, is an extension of the fuzzy c-means algorithm to distribution-valued data, using the L2 Wasserstein distance between distributions. A second algorithm, AFCM-D, extends the classical fuzzy c-means to distribution-valued data using adaptive L2 Wasserstein distance. The adaptive distances induce a system of weights on the components of the distance for each distributional variable, showing the importance of
TABLE V
AFCM-D MOST AMBIGUOUS COUNTRIES: MEMBERSHIPS

| Countries            | Clust. 1 | Clust. 2 | Clust. 3 |
|----------------------|----------|----------|----------|
| Virgin Islands, Brit | 0.4895   | 0.4918   | 0.0186   |
| United Arab Emirates | 0.5381   | 0.0225   | 0.4394   |
| Armenia              | 0.4764   | 0.5067   | 0.0168   |
| Trinidad and Tobago  | 0.5318   | 0.4505   | 0.0177   |
| Singapore            | 0.4318   | 0.5512   | 0.0170   |
| Egypt                | 0.4722   | 0.0091   | 0.5187   |
| Bangladesh           | 0.4594   | 0.0089   | 0.5317   |
| Nicaragua            | 0.4241   | 0.0090   | 0.5669   |
| Saudi Arabia         | 0.3832   | 0.0086   | 0.6082   |
| Cabo Verde           | 0.3471   | 0.0086   | 0.6443   |

TABLE VI
AFCM-D WEIGHTS OF THE COMPONENTS OF THE VARIABLES.

| Male population | λ1, g | λ1, Disp | λ2, g | λ2, Disp |
|-----------------|-------|----------|-------|----------|
| Cluster 1       | 1.1757| 0.9117   | 0.8505| 1.0968   |
| Cluster 2       | 1.1366| 1.0565   | 0.8798| 0.9466   |
| Cluster 3       | 0.9934| 0.9517   | 1.0067| 1.0508   |

Pyramids of clusters centers

![Pyramids of clusters centers](image)

Fig. 4. Centers of clusters description

each component to the clustering criterion. Applications on simulated and real data confirm the usefulness of the proposed algorithms.

ACKNOWLEDGMENT

The first author would like to thank FACEPE (Research Agency from the State of Pernambuco, Brazil), CNPq (National Council for Scientific and Technological Development, Brazil) and Second University of Naples for their financial support.

REFERENCES

[1] H. H. Bock and E. Diday, Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data. Berlin: Springer, 2000.
[2] L. Rüschendorf, “Wasserstein metric,” in Encyclopedia of Mathematics. Springer, 2001.
[3] E. Diday and J. C. Simon, “Clustering analysis,” in Digital Pattern Classification, K. Fu, Ed. Berlin: Springer, 1976, pp. 47–94.
[4] A. Irpino and R. Verde, “A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data,” in Data Science and Classification, V. Batenjeli, H. Bock, A. Felfuga, and A. Ziberna, Eds. Berlin: Springer, 2006, pp. 185–192.
[5] R. Verde and A. Irpino, “Dynamic clustering of histograms using Wasserstein metric,” in Proceedings in Computational Statistics, COMPSTAT 2006, A. Rizzi and M. Vichi, Eds., Compstat 2006. Heidelberg: Physica Verlag, 2006, pp. 869–876.
[6] ——, “Comparing histogram data using a Mahalanobis-Wasserstein distance,” in Proceedings in Computational Statistics, COMPSTAT 2008, P. Brito, Ed., Compstat 2008. Heidelberg: Springer Verlag, 2008, pp. 77–89.
[7] ——, “Dynamic clustering of histogram data: using the right metric,” in Selected contributions in data analysis and classification, P. Brito, P. Bertrand, G. Cucumel, and F. De Carvalho, Eds. Berlin: Springer, 2008, pp. 123–134.
[8] Y. Terada and H. Yadohisa, “Non-hierarchical clustering for distribution-valued data,” in Proceedings of COMPSTAT 2010, Y. Lechevallier and G. Saporta, Eds. Berlin: Springer, 2010, pp. 1653–1660.
[9] M. Vrac, L. Billard, E. Diday, and A. Chédin, “Copula analysis of mixture models,” Computational Statistics, vol. 27, pp. 427–457, 2012.
[10] E. Diday and G. Govaert, “Classification automatique avec distances adaptatives,” R.A.I.R.O. Informatique Computer Science, vol. 11, no. 4, pp. 329–349, 1997.
[11] F. A. T. De Carvalho and Y. Lechevallier, “Partitioning clustering algorithms for symbolic interval data based on single adaptive distances,” Pattern Recognition, vol. 42, no. 7, pp. 1223–1236, 2009.
[12] ——, “Dynamic clustering of interval-valued data based on adaptive quadratic distances,” Trans. Sys. Man Cyber. Part A, vol. 39, no. 6, pp. 1295–1306, 2009.
[13] F. A. T. De Carvalho and R. M. C. R. De Souza, “Unsupervised pattern recognition models for mixed feature—symbolic data,” Pattern Recognition Letters, vol. 31, pp. 430–443, 2010.
[14] A. Irpino, R. Verde, and F. de A.T. De Carvalho, “Dynamic clustering of histogram data based on adaptive squared wasserstein distances,” Expert Systems with Applications, vol. 41, no. 7, pp. 3351 – 3366, 2014.
[15] A. Irpino and E. Romano. “Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation,” Revue des Nouvelles Technologies de l’Information, vol. RNTI-E, pp. 99–110, 2007.
[16] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Norwell, MA, USA: Kluwer Academic Publishers, 1981.
[17] A. L. Gibbs and F. E. Su, “On choosing and bounding probability metrics,” Int. Stat. Rev., vol. 7, no. 3, pp. 419–435, 2002.
[18] J. A. Cuesta-Albertos, C. Matrán, and A. Tuero-Díaz, “Optimal transportation plans and convergence in distribution,” Journal of Multiv. An., vol. 60, pp. 72–83, 1997.
[19] R. G. Clark and M. S. Rae, “A class of Wasserstein metrics for probability distributions,” Michigan Math. J., vol. 31 (2), pp. 231–240, 1984.
[20] H. Frigui, C. Hwang, and F. C.-H. Rhee, “Clustering and aggregation of relational data with applications to image database categorization,” Pattern Recognition, vol. 40, no. 11, pp. 3053 – 3068, 2007.
[21] W. Wang and Y. Zhang, “On fuzzy cluster validity indices,” Fuzzy Sets and Systems, vol. 158, no. 19, pp. 2095 – 2117, 2007.
[22] R. N. Dave, “Validating fuzzy partitions obtained through c-shells clustering,” Pattern Recognition Letters, vol. 17, no. 6, pp. 613 – 623, 1996.
[23] F. de A.T. de Carvalho, “Fuzzy c-means clustering methods for symbolic interval data,” Pattern Recognition Letters, vol. 28, no. 4, pp. 423 – 437, 2007.[Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865506002297
[24] G. Celeux, E. Diday, G. Govaert, Y. Lechevalier, and H. Ralaibondrainy, Classification Automatique des Données. Paris: Bordas, 1989.
[25] W. Gilchrist, Statistical Modelling with Quantile Functions. Abingdon: CRC Press, 2000.