Adaptive Dynamic Clustering Algorithm for Interval-valued Data based on Squared-Wasserstein Distance

Rong Guan*, Yves Lechevallier**
Huiwen Wang*

*School of Economics and Management, Beihang University, Beijing, 100191, China
rongguan77@gmail.com (R. Guan)
wanghw@vip.sina.com (H. Wang)
**INRIA-Institut National de Recherche en Informatique et en Automatique Domaine de Voluceau, Rocquencourt B.P.105, 78153 Le Chesnay Cedex, France
Yves.Lechevallier@inria.fr

Abstract. Wide applications of interval-valued data in various domains have triggered the call for more powerful analytical tools. In light of this, this paper has presented an adaptive dynamic clustering algorithm for interval-valued data, using squared-Wasserstein distance. Experiments on both synthetic data and real data have unveiled the merits of the proposed algorithm.

1 Introduction

The technique of clustering deals with finding a structure in a collection of objects, which groups objects of similar kind into respective categories. As a main task of explorative statistical analysis, clustering has been widely used in machine learning, pattern recognition, image analysis and other fields of data mining.

The development of computer science in recent decades has enabled us to record immense amount of data. Data sets with a large number of objects are commonly seen in clustering. In some cases, however, analysts may prefer to concentrate on higher level conceptual objects rather than massive and too-specific individual objects. For example, it makes more sense to perform clustering on consumer groups, say male and female, or the young and the old, in order to analyze their buying behaviors. Potential applications also exist in complex-structured database or privacy-preserved census data, where conceptual observations should be employed to prevent identification of specific individuals. Symbolic Data Analysis (Diday, 1989; Bock and Diday, 2000; Billard and Diday, 2003; Diday and Noirhomme-Fraiture, 2008) has directed an innovative way for solving this problem. The technique aims to generalize large-scale individuals to conceptual objects described by symbolic data, such as categorical multi-valued data, interval-valued data, modal data, etc., and to extend classical statistical methods or develop new approaches for multivariate analysis on symbolic data. As a main topic in symbolic data analysis, clustering methods on symbolic data, especially on interval-valued data, has aroused much attention in recent years (Diday and Brito, 1989; Bock, 2002; Chavent et al., 2006; De Carvalho, 2007; Costa et al., 2010).

One of the well-discussed clustering methods is dynamic clustering algorithm (DCA), firstly proposed by Diday and Simon (1976). Recent years have witnessed an increasing number of literatures on DCA for interval-valued data based on different dissimilarity measures. Chavent and Lechevallier (2002) have used Hausdorff distance in the algorithm. De Souza and De Carvalho (2004) have extended this algorithm to city-block distance. An optimality criterion based on squared Euclidean distance has been proposed by De Carvalho et al. (2006a). Irpino and Verde (2008) presented a Wasserstein-based distance for DCA on interval-valued data, and investigated its properties in the clustering algorithm.

In order to recognizes different shapes and sizes of clusters, Diday and Govaert (1977) have proposed an adaptive dynamical clustering algorithm (ADCA), which associates a distance to each cluster. De Souza and De Carvalho (2004) and De Carvalho et al. (2006b) respectively provided ADCA based on city-block distance and Hausdorff distance for interval-valued data. Both non-adaptive and adaptive algorithms based on Mahalanobis distance have been proposed by De Souza et al. (2004). De Carvalho and Lechevallier (2009b) presented a comparison between ADCA using city-block distance and Hausdorff distance. A novel ADCA using quadratic distances was proposed by De Carvalho and Lechevallier (2009a).

In this paper, we intend to present ADCA with squared-Wasserstain distance for clustering interval-valued data. This algorithm could be considered as an extension of DCA based on Wasserstain-based distances proposed by Irpino and Verde (2008). The remainder of this paper is structured as follows: Section 2 and Section 3 will respectively introduce preliminaries on ADCA and Wasserstain distance for interval-valued data; squared-Wasserstain-distance-based ADCA will be proposed in Section 4; to demonstrate the merits of the proposed algorithm, Section 5 will conduct an experiment with synthetic data sets; two cases of real life applications will be used to further unveil the usefulness of the proposed algorithm in Section 6; and we will conclude our work in Section 7.

2 Preliminaries

Suppose there exists a set of n objects $\Omega = \{1, 2, \cdots, n\}$, each of which is represented by a p-dimensional vector $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{ip})$. Given a predefined cluster number K, dynamic clustering algorithm (Diday and Simon, 1976), hereinafter referred to as DCA for short, aims to find out a partition $P = \{C_1, C_2, \cdots, C_K\}$ that classifies Ω into K clusters with a given criterion. If each cluster is represented by a prototype, denoted as $\mathbf{y}_h' = (y_{h1}, y_{h2}, \cdots, y_{hp})$, a good partition shall achieve that each observation in a cluster is similar to the prototype that represents the cluster, while dissimilar to the prototype of any other clusters. Consequently, the clustering criterion is to minimize the following function

$$\Delta(P, L) = \sum_{h=1}^{K} \sum_{i \in C_h} \delta(\mathbf{x}_i, \mathbf{y}_h), \tag{1}$$

where $L = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_K\}$ represents the prototype set and $\delta(\cdot, \cdot)$ is a dissimilarity measure.

As proposed by Diday and Simon (1976), DCA implements by iteratively performing a two-stage algorithm as follows:

1. *Representative stage*
   Given P fixed, L that minimizes $\Delta(P, L)$ can be obtained by finding for $h = 1, 2, \cdots, K$, the prototype $\mathbf{y}_h$ that minimizes the criterion $\sum_{i \in C_h} \delta(\mathbf{x}_i, \mathbf{y}_h)$.
2. *Allocation stage*
   Given L fixed, P that minimizes $\Delta(P, L)$ can be obtained by finding for $h = 1, 2, \cdots, K$, the cluster $C_h = \{i \in \Omega | \delta(\mathbf{x}_i, \mathbf{y}_h) \leq \delta(\mathbf{x}_i, \mathbf{y}_m), \forall m = 1, 2, \cdots, K\}$.

The dissimilarity measure $\delta(\cdot, \cdot)$ in Equation (1) is usually expressed as

$$\delta(\mathbf{x}_i, \mathbf{y}_h) = \sum_{j=1}^{p} d(x_{ij}, y_{hj}), \tag{2}$$

where $d(\cdot, \cdot)$ represents a dissimilarity function on $\mathbb{R}$, say Euclidean distance or squared Euclidean distance. Apparently, the dissimilarity function for each pair of units has been equally weighted in Equation (2), which indicates the equivalent effectiveness of each variable for clustering. In order to find out the potentially different importance of variables for clustering, Diday and Govaert (1977) have proposed to associate a weighted factor, denoted as λ_{hj}, to dissimilarity function $d(x_{ij}, y_{hj})$ and cluster C_h, i.e.,

$$\delta(\mathbf{x}_i, \mathbf{y}_h) = \sum_{j=1}^{p} \lambda_{hj} d(x_{ij}, y_{hj}), \tag{3}$$

subjecting to $\lambda_{hj} > 0$ and $\prod_{j=1}^{p} \lambda_{hj} = 1$.

λ_{hj} is referred to as *adaptive factor*, since this weight will be updated in each iteration for each variable and for each cluster. Using $\Lambda = (\lambda_{hj})_{K \times p}$ to denote an adaptive factor matrix, we could rewrite the clustering criterion as

$$\Delta(P, L, \Lambda) = \sum_{h=1}^{K} \sum_{i \in C_h} \sum_{j=1}^{p} \lambda_{hj} d(x_{ij}, y_{hj}). \tag{4}$$

Accordingly, this adaptive dynamic clustering algorithm (ADCA) will be modified as a three-stage process as follows:

1. *Representative stage*
   Given both P and Λ fixed, L that minimizes $\Delta(P, L, \Lambda)$ can be obtained by finding for $h = 1, 2, \cdots, K$, the prototype $\mathbf{y}_h$ that minimizes $\sum_{j=1}^{p} \lambda_{hj} \sum_{i \in C_h} d(x_{ij}, y_{hj})$.
2. *Adaptive stage*
   Given both P and L fixed, Λ that minimizes $\Delta(P, L, \Lambda)$ can be obtained by finding for $h = 1, 2, \cdots, K$ and $j = 1, 2, \cdots, p$, the adaptive factor λ_{hj} that minimizes $\lambda_{hj} \sum_{i \in C_h} d(x_{ij}, y_{hj})$.
3. *Allocation stage*
   Given both L and Λ fixed, P that minimizes $\Delta(P, L, \Lambda)$ can be obtained by finding for $h = 1, 2, \cdots, K$, the cluster $C_h = \{i \in \Omega | \sum_{j=1}^{p} \sum_{i \in C_h} \lambda_{hj} \, d(x_{ij}, y_{hj}) \leq \sum_{j=1}^{p} \sum_{i \in C_m} \lambda_{mj} d(x_{ij}, y_{mj}), \forall m = 1, 2, \cdots, K\}$.

Diday and Govaert (1977) have proved that ADCA will converge once the above three stages have been properly defined.

3 A brief introduction to Wasserstein distance

In this paper, objects described by interval-valued data are concerned. More specifically, for $i = 1, 2, \cdots, n$, each object is described by an interval-valued vector $\mathbf{x}_i' = (x_{i1}, \cdots, x_{ip})$, with each unit being an interval-valued data, i.e., $x_{ij} = [\underline{x}_{ij}, \overline{x}_{ij}](1 \leq j \leq p)$. Before derivation of ADCA on interval-valued data, we shall firstly discuss the dissimilarity function between two interval-valued data.

Given any two random variables f and g, if F and G respectively represent the distribution functions, the Wasserstein L_2 distance (Gibbs and Su, 2002) is defined as

$$d_{Wass}(F, G) = \left(\int_0^1 \left(F^{-1}(t) - G^{-1}(t) \right)^2 dt \right)^{\frac{1}{2}} \tag{5}$$

where F^{-1} and G^{-1} respectively represent the quantile function of F and G.

To achieve a general computation for this definition, Irpino and Romano (2007) have provided a formulation of Wasserstein distance, i.e.,

$$d_{Wass}(F, G) = \sqrt{(\mu_f - \mu_g)^2 + (\sigma_f - \sigma_g)^2 + 2\sigma_f \sigma_g (1 - \rho_{QQ}(F, G))}, \tag{6}$$

where $\mu_f = \int_{-\infty}^{+\infty} t dF(t)$ and $\mu_g = \int_{-\infty}^{+\infty} t dG(t)$ respectively represents the expectation of the random variable f and g, $\sigma_f^2 = \int_{-\infty}^{+\infty} t^2 dF(t) - \mu_f^2$ and $\sigma_g^2 = \int_{-\infty}^{+\infty} t^2 dG(t) - \mu_g^2$ respectively denote the variance of f and g, and

$$\rho_{QQ}(F, G) = \frac{\int_0^1 \left(F^{-1}(t) - \mu_f \right) \left(G^{-1}(t) - \mu_g \right) dt}{\sigma_f \sigma_g} = \frac{\int_0^1 F^{-1}(t) G^{-1}(t) dt - \mu_f \mu_g}{\sigma_f \sigma_g} \tag{7}$$

is a measure of shape similarity between two distributions, ranging from 0 to 1. A higher value of $\rho_{QQ}(F, G)$ indicates better similarity of two distributions in terms of shape. When $\rho_{QQ}(F, G) = 1$, it can be inferred that both of the two distributions are of the same shape, i.e., same distribution function after standardization with expectation of 0 and variance of 1.

As illustrated by Irpino and Romano (2007), Equation (6) has allowed us to decompose the dissimilarity function between two distributions into three aspects. The first part indicates how the two distributions differ from each other in location, the second concentrates on difference in terms of size, and the third item reports shape difference. When F and G are of the same shape, the third part in Equation (6) will be eliminated.

According to Irpino and Verde (2008), Wasserstein distance between any two interval-valued data, i.e., $x_i = [\underline{x}_i, \overline{x}_i]$ and $x_j = [\underline{x}_j, \overline{x}_j]$, can be expressed as

$$d_{Wass}(x_i, x_j) = \sqrt{\left(\frac{\overline{x}_i + \underline{x}_i}{2} - \frac{\overline{x}_j + \underline{x}_j}{2} \right)^2 + \frac{1}{3} \left(\frac{\overline{x}_i - \underline{x}_i}{2} - \frac{\overline{x}_j - \underline{x}_j}{2} \right)^2}. \tag{8}$$

It is easy to verify that Equation (8) holds due to the assumption that points in the concerning two intervals follow uniform distribution (Bock and Diday, 2000; Billard and Diday, 2003; Diday and Noirhomme-Fraiture, 2008), which leads to $\mu_{x_i} = \frac{1}{2}(\overline{x}_i + \underline{x}_i)$, $\mu_{x_j} = \frac{1}{2}(\overline{x}_j + \underline{x}_j)$, $\sigma_{x_i}^2 = \frac{1}{12}(\overline{x}_i - \underline{x}_i)^2$, and $\sigma_{x_j}^2 = \frac{1}{12}(\overline{x}_j - \underline{x}_j)^2$.

It is interesting to notice that there is another expression of interval-valued data, i.e., $x_i = (x_i^c, x_i^r)$ and $x_j = (x_j^c, x_j^r)$, where x_i^c and x_j^c respectively represent midpoints of x_i and x_j,

while x_i^r and x_j^r correspond to radius, i.e., $x_i^r = \frac{1}{2}(\overline{x}_i - \underline{x}_i)$ and $x_j^r = \frac{1}{2}(\overline{x}_j - \underline{x}_j)$. Therefore, we could rewrite Equation (8) as follows:

$$d_{Wass}(x_i, x_j) = \sqrt{(x_i^c - x_j^c)^2 + \frac{1}{3}(x_i^r - x_j^r)^2}. \tag{9}$$

It is easy to prove that Wasserstein distance is a metric. For any three interval-valued data x_i, x_j and x_k, Wasserstein distance $d_{Wass}(\cdot, \cdot)$ satisfies the following three properties, i.e.,

- Non-negativity, i.e., $d_{Wass}(x_i, x_j) \geq 0$, and $d_{Wass}(x_i, x_j) = 0$ if and only if $x_i = x_j$.
- Symmetry, i.e., $d_{Wass}(x_i, x_j) = d_{Wass}(x_j, x_i)$.
- Triangle inequality, i.e., $d_{Wass}(x_i, x_j) + d_{Wass}(x_j, x_k) \geq d_{Wass}(x_i, x_k)$.

Notably, Wasserstein distance could also deal with single-valued numeric data. For instance, given x_i being a single-valued numeric data, Equation (9) proceeds by setting $x_i^r = 0$. Furthermore, the equation will reduce to Euclidean distance if two single-valued numeric data are involved in computation.

4 Adaptive Dynamic Clustering Algorithm using Squared-Wasserstein Distance

Equipped with Wasserstein distance, we are now able to accomplish the derivation of ADCA on interval-valued data.

We still assume that there is a set of objects as $\Omega = \{1, 2, \cdots, n\}$, however, each object in this paper is described by p interval-valued variables, i.e., $\mathbf{x}_i' = (x_{i1}, x_{i2}, \cdots, x_{ip})$, with each unit being an interval-valued data $x_{ij} = [\underline{x}_{ij}, \overline{x}_{ij}]$ $(1 \leq i \leq n,\ 1 \leq j \leq p)$. Let us denote the prototype of the hth $(1 \leq h \leq K)$ cluster C_h as $\mathbf{y}_h' = (y_{h1}, y_{h2}, \cdots, y_{hp})$, with $y_{hj} = [\underline{y}_{hj}, \overline{y}_{hj}]$ $(1 \leq j \leq p)$. Using squared-Wasserstein distance, the clustering criterion can be expressed as

$$\Delta(P, L, \Lambda) = \sum_{h=1}^{K} \sum_{i \in C_h} \sum_{j=1}^{p} \lambda_{hj} d_{Wass}^2(x_{ij}, y_{hj}), \tag{10}$$

subjecting to $\lambda_{hj} > 0$ and $\prod_{j=1}^{p} \lambda_{hj} = 1$. If we define $\Psi_{hj} = \sum_{i \in C_h} d_{Wass}^2(x_{ij}, y_{hj})$, the criterion in Equation (10) can also be given by

$$\Delta(P, L, \Lambda) = \sum_{h=1}^{K} \sum_{j=1}^{p} \lambda_{hj} \Psi_{hj}. \tag{11}$$

In order to achieve the minimization of the criterion in Equation (11), as mentioned in previous section, a three-stage algorithm shall be employed, which updates the prototype set L in *Representative stage*, the adaptive factor set Λ in *Adaptive stage*, and the partition P in *Location stage* for each iteration.

(i) Definition of the best prototypes

In the first step of *Representative stage*, we aim to find out the best prototype $\hat{\mathbf{y}}_h' = ([\hat{\underline{y}}_{h1}, \hat{\overline{y}}_{h1}], \cdots, [\hat{\underline{y}}_{hp}, \hat{\overline{y}}_{hp}])$ that minimizes Ψ_{hj}, when the adaptive factor set Λ and the partition P are fixed. For $j = 1, 2, \cdots, p$, using distance definition in Equation (9), we take partial

derivatives of Ψ_{hj} with respect to y_{hj}^c and y_{hj}^r respectively and let them be zero, i.e.,

$$\frac{\partial \Psi_{hj}}{\partial y_{hj}^c} = \frac{\partial}{\partial y_{hj}^c}\left(\sum_{i\in C_h}(x_{ij}^c - y_{hj}^c)^2 + \frac{1}{3}(x_{ij}^r - y_{hj}^r)^2\right) = 0, \tag{12}$$

$$\frac{\partial \Psi_{hj}}{\partial y_{hj}^r} = \frac{\partial}{\partial y_{hj}^r}\left(\sum_{i\in C_h}(x_{ij}^c - y_{hj}^c)^2 + \frac{1}{3}(x_{ij}^r - y_{hj}^r)^2\right) = 0. \tag{13}$$

The solutions to Equation (12) and Equation (13) are

$$\hat{y}_{hj}^c = \frac{1}{|C_h|}\sum_{i\in C_h} x_{ij}^c, \quad \hat{y}_{hj}^r = \frac{1}{|C_h|}\sum_{i\in C_h} x_{ij}^r, \tag{14}$$

where $|C_h|$ represents the object number of C_h.

Accordingly, we could have

$$\hat{\underline{y}}_{hj} = \hat{y}_{hj}^c - \hat{y}_{hj}^r = \frac{1}{|C_h|}\sum_{i\in C_h}\underline{x}_{ij}, \quad \hat{\overline{y}}_{hj} = \hat{y}_{hj}^c + \hat{y}_{hj}^r = \frac{1}{|C_h|}\sum_{i\in C_h}\overline{x}_{ij}. \tag{15}$$

And the best prototype is $\hat{\mathbf{y}}_h' = ([\hat{\underline{y}}_{h1}, \hat{\overline{y}}_{h1}], \cdots, [\hat{\underline{y}}_{hp}, \hat{\overline{y}}_{hp}])$.

(ii)Definition of the best adaptive factors

Once the best prototypes have been determined, the next problem is to look for the best adaptive factors $\hat{\lambda}_{hj}$ for $j = 1, 2, \cdots, p$ that achieves minimization of $\sum_{j=1}^p \lambda_{hj}\Psi_{hj}$. Considering that λ_{hj} is subject to the constraint $\prod_{j=1}^p \lambda_{hj} = 1$, we can deduce the solutions by the Lagrange multiplier method (Arfken, 1985), i.e.,

$$\frac{\partial}{\partial \lambda_{hj}}\left[\sum_{j=1}^p \lambda_{hj}\Psi_{hj} - \mu\left(\prod_{j=1}^p \lambda_{hj} - 1\right)\right] = 0, \quad \text{for} \quad j = 1, 2, \cdots, p, \tag{16}$$

where $\mu \neq 0$ is an unknown Lagrange multiplier. From Equation (16) we can easily obtain the following result, i.e.,

$$\hat{\lambda}_{hj} = \frac{\mu}{\Psi_{hj}}. \tag{17}$$

Due to $\prod_{j=1}^p \lambda_{hj} = 1$, the Lagrange multiplier can be expressed as

$$\mu = \left(\prod_{j=1}^p \Psi_{hj}\right)^{\frac{1}{p}}. \tag{18}$$

Thus, we finally derive the solutions $\hat{\lambda}_{hj}$ $(j = 1, 2, \cdots, p)$ as follows:

$$\hat{\lambda}_{hj} = \frac{\mu}{\Psi_{hj}} = \frac{\left\{\prod_{k=1}^p \left[\sum_{i\in C_h}(x_{ik}^c - \hat{y}_{hk}^c)^2 + \frac{1}{3}(x_{ik}^r - \hat{y}_{hk}^r)^2\right]\right\}^{\frac{1}{p}}}{\sum_{i\in C_h}(x_{ij}^c - \hat{y}_{hj}^c)^2 + \frac{1}{3}(x_{ij}^r - \hat{y}_{hj}^r)^2}. \tag{19}$$

A close look at Equation (19) will help us understand the meaning of adaptive factors. For the hth cluster, apparently, the denominator Ψ_{hj} varies with variables while the numerator is fixed. Consequently, $\hat{\lambda}_{hj}$ tends to get lower value when Ψ_{hj} is larger, while on the contrary, the value of $\hat{\lambda}_{hj}$ will be higher. Since Ψ_{hj} indicates within-cluster sum of dissimilarity in the jth variable, we may infer that adaptive factors give higher weight to variables, on which the concerning cluster appears rather compact, while assigns lower value of weight to variables that makes the cluster loose. Indeed, this contributes to recognizing the importance of variables in clustering.

Noticeable, both Equation (14) and Equation (19) allow mixture of data type, i.e., single real-valued variables together with interval-valued variables in the concerning data set. This is because single real-valued data can be rewritten as a special form of interval-valued data, i.e., the lower bound equals to the upper bound.

Equipped with the above definitions, in the following we summarize ADCA on interval-valued data with squared-Wasserstein distance.

1. *Initialization*
   Randomly select K prototypes and a partition $\{C_1, C_2, \cdots, C_K\}$ of Ω.
2. *Allocation stage*
   $test \leftarrow 0$
   for $i = 1$ to n **do**
   define the winning cluster C_{k^*} such that
   $k^* = \arg\min_{h=1,\ldots,K} \Psi_h(\mathbf{x}_i, \mathbf{y}_h)$
   if $i \in C_k$ and $k \neq k^*$ **then**
   $test \leftarrow 1$
   $C_{k^*} \leftarrow C_{k^*} \cup \{i\}$
   $C_k \leftarrow C_k \setminus \{i\}$
   end if
   end for
3. *Representative stage*
   For $h = 1, 2, \cdots, K$ compute the prototype $\hat{\mathbf{y}}_h = ([\hat{\underline{y}}_{h1}, \hat{\overline{y}}_{h1}], [\hat{\underline{y}}_{h2}, \hat{\overline{y}}_{h2}], \cdots, [\hat{\underline{y}}_{hp}, \hat{\overline{y}}_{hp}])'$.
4. *Adaptive stage*
   For $j = 1, 2, \cdots, p$ and $h = 1, 2, \cdots, K$, compute $\hat{\lambda}_{hj}$.
5. *Stopping criterion*
   If $test = 0$ then STOP, otherwise go to *step 2*.

The convergence of this algorithm is achieved due to the decrease of the partitioning criterion in Equation (10) at each iteration, which is on account of the optimization of the adequacy criterion at each *Representative stage* and each *Adaptive stage*.

5 Experiment on Synthetic Data Sets

In order to demonstrate the merits of the proposed algorithm, experiment based on Monte Carlo simulation will be carried out in this section.

The comparison is expected to validate (1) the merits of squared-Wasserstein distance in ADCA, compared with L_1 distance, L_2 distance and Hausdorff distance; (2) the superiority

of ADCA to DCA when using squared-Wasserstein distance. Consequently, ADCA considering four different distances, i.e., L_1 distance, L_2 distance, Hausdorff distance and squared-Wasserstein distance as well as DCA based on squared-Wasserstein distance will be performed on synthetic data sets with clusters of different shapes and sizes. We will use an external indicator to measure the similarity between different partitions obtained from different clustering algorithms and a priori partition labeled to the synthetic data.

5.1 Data

We will consider synthetic data sets with the same configuration proposed by De Souza and De Carvalho (2004). Each data set, with 350 observations described by two interval-valued variables, includes three clusters of different sizes and shapes: two clusters with an ellipsoidal shape and size of 150 each and one cluster with a spherical shape and size of 150. And this 3-cluster partition will work as a priori partition.

Since an interval-valued data can be constructed by the combination of center and radius, this paper will generate interval-valued observations as follows:

$$\big([x_{i1}^c - x_{i1}^r, x_{i1}^c + x_{i1}^r], [x_{i2}^c - x_{i2}^r, x_{i2}^c + x_{i2}^r]\big), \tag{20}$$

where x_{ij}^c and x_{ij}^r $(i = 1, 2, \cdots, 350, j = 1, 2)$ represent the midpoint and radius respectively. The parameter vector (x_{i1}^c, x_{i2}^c) obeys a bi-variate normal distribution, say $N_2(\mu, \mathbf{\Sigma})$, where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad and \quad \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}. \tag{21}$$

As shown in Table 1, two types of data set will be constructed according to different configurations of μ. The predefined clusters are either well-separated or overlapping in the two data sets, since the parameter of x_{ij}^c decides positions of interval-valued observations.

Data sets	Clusters	μ_1	μ_2	σ_1^2	σ_2^2
well-separated	C_1	28	22	100	9
	C_2	60	30	9	144
	C_3	45	38	9	9
overlapping	C_1	45	22	100	9
	C_2	60	30	9	144
	C_3	52	38	9	9

TAB. 1 – *Two kinds of synthetic data sets.*

With different values of x_{ij}^r, we will obtain interval-valued observations in different shapes and sizes. There are five configurations of x_{ij}^r, namely, $[1, 4]$, $[1, 8]$, $[1, 12]$, $[1, 16]$ and $[1, 20]$, involved in our experiments. For each configuration, we will randomly select 350 sample points of x_{ij}^r that uniformly distributed within each of the five given intervals.

5.2 Indicator

The indicator for evaluating effectiveness of different clustering algorithms is the Corrected Rand (CR) index. Given two partitions of the same data set, $U = \{u_1, u_2, \cdots, u_R\}$ and $V = \{v_1, v_2, \cdots, v_C\}$, CR is defined as

$$CR = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i\cdot}}{2}\sum_{j=1}^{C}\binom{n_{\cdot j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{R}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{C}\binom{n_{\cdot j}}{2}\right] - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i\cdot}}{2}\sum_{j=1}^{C}\binom{n_{\cdot j}}{2}}, \tag{22}$$

where $\binom{n}{2} = \frac{n(n-1)}{2}$ and n_{ij} represents the number of objects that are in clusters of both u_i and v_j, $n_{i\cdot}$ and $n_{\cdot j}$ count the number of object respectively in cluster u_i and v_j, and n is the total number of objects in the data set.

Ranging within the interval of $[-1, 1]$, CR helps to show how well the partition obtained by the clustering algorithm matches the priori partition. The higher the absolute value of the indicator, the better the obtained partition agrees to the priori partition.

5.3 Procedures

The experiment will be conducted in the framework of Monte Carlo simulations as follows:

1. For each type of data set (well-separated or overlapping) and each configuration of x_{ij}^r, the experiment will be repeated 50 times.
2. In each time of experiment, randomly select 350 pairs of (x_{i1}^c, x_{i2}^c) and (x_{i1}^r, x_{i2}^r) from the corresponding distributions as described above to build 350 2-dimensional interval-valued observations.
3. Adopt ADCA based on L_1 distance, L_2 distance, Hausdorff distance and squared Wasserstein distance and DCA using squared Wasserstein distance to obtain 3-cluster partitions respectively. For each algorithm, the best partition with the lowest value of the clustering criterion will be selected among 50 replications.
4. In each time of experiment, record CR index for each of the five algorithms. Calculate the average value of CR index for each algorithm in 50 times.
5. For each type of data set and each configuration of x_{ij}^r, perform t-test with a 5% level of significance for 50 paired samples by algorithms A and B under the following hypothesis (null and alternative):
   H_0: $CR(A) - CR(B) \leq 0$ $\quad$ H_1: $CR(A) - CR(B) > 0$,
   where algorithm A represents ADCA based on squared-Wasserstein distance, and algorithm B can be replaced by ADCA using L_1 distance, L_2 distance, Hausdorff distance, or DCA using squared-Wasserstein distance. Record the t-statistics values for each type of data set and each configuration.

5.4 Results

Comparative results of simulation experiments have been listed in Table 2 and Table 3.

Table 2 displays the result of averaged CR index value of each algorithm in 50 times of experiment for each data set and for each configuration. Apparently, most cases demonstrate

Data sets	Configurations	L_1 (ADCA)	L_2 (ADCA)	*Hausd.* (ADCA)	*sq-Wass.* (ADCA)	*sq-Wass.* (DCA)
well-separated	$[1, 4]$	0.9214	0.9689	0.9150	0.9684	0.6098
	$[1, 8]$	0.9277	0.9493	0.8690	0.9567	0.6063
	$[1, 12]$	0.8647	0.8717	0.8589	0.9406	0.6045
	$[1, 16]$	0.7163	0.7820	0.6771	0.9146	0.6035
	$[1, 20]$	0.6606	0.7237	0.5892	0.8932	0.5970
overlapping	$[1, 4]$	0.3374	0.3407	0.2847	0.4588	0.3488
	$[1, 8]$	0.4861	0.5012	0.4680	0.5509	0.3495
	$[1, 12]$	0.4206	0.4611	0.3884	0.5225	0.3446
	$[1, 16]$	0.3918	0.4191	0.3419	0.5021	0.3524
	$[1, 20]$	0.3375	0.3407	0.2847	0.4588	0.3488

TAB. 2 – *Average values of CR index for ADCA based on L_1 distance, L_2 distance, Hausdorff distance and squared-Wasserstein distance and for DCA using squared-Wasserstein distance.*

the merits of using squared-Wasserstein distance in ADCA. Concerning the well-separated data sets, squared-Wasserstein distance is superior to L_1 distance and Hausdorff distance regardless of configurations of x_{ij}^r. Compared with L_2 distance, squared-Wasserstein distance shows its advantage except for the configuration of $[1, 4]$. For the overlapping data sets, squared-Wasserstein distance also outstands from the other three distances with higher averaged CR values.

Concerning the comparison of ADCA and DCA when using squared-Wasserstein distance, as expected, the averaged CR values of ADCA based on squared-Wasserstein distance are much higher than average values of CR index for DCA using squared-Wasserstein distance, regardless of configurations and data set types. We could therefore conclude that ADCA outperforms DCA when using squared-Wasserstein distance.

To further support the above conclusions, we perform paired sample t-test for CR index values between squared-Wasserstein-distance-based ADCA and L_1-distance-based ADCA (or L_2-distance-based ADCA, or Hausdorff-distance-based ADCA, or squared-Wasserstein-distance-based DCA). The results have been displayed in Table 3.

Data sets	Configurations	L_1 (ADCA)	L_2 (ADCA)	$Hausd.$ (ADCA)	$Wass.$ (DCA)
well-separated	$[1, 4]$	2.368^*	-1.418	2.933^*	43.076^*
	$[1, 8]$	2.063^*	1.783	4.186^*	27.707^*
	$[1, 12]$	2.969^*	2.914^*	2.948^*	17.841^*
	$[1, 16]$	8.862^*	6.328^*	8.297^*	19.705^*
	$[1, 20]$	11.307^*	7.206^*	11.543^*	17.879^*
overlapping	$[1, 4]$	2.981^*	1.286	7.003^*	14.500^*
	$[1, 8]$	4.605^*	3.393^*	6.439^*	13.571^*
	$[1, 12]$	6.637^*	4.393^*	7.026^*	10.958^*
	$[1, 16]$	10.983^*	7.713^*	10.895^*	16.229^*
	$[1, 20]$	11.964^*	10.139^*	13.893^*	12.318^*

TAB. 3 – *T-statistics values in paired sample t-test (values with a superscript of $*$ indicates rejection of H_0).*

As expected, the result shows the superiority of ADCA to DCA when using squared-Wasserstein distance. The null hypothesis can be rejected regardless of data configurations. This is mainly accounted by the fact that DCA treats all clusters equally, yet the synthetic data set is constructed with different cluster types.

Focusing on ADCA, squared-Wasserstein distance outperforms L_1 distance and Hausdorff distance in 100% of the configurations for both datasets, which is consistent of results displayed in Table 2. In comparison with L_2 distance, however, the percentages of rejecting hypothesis fall to 60% and 80% respectively for the well-separated data set and the overlapping data set. This suggests that ADCA based on squared-Wasserstein distance performs better in recognizing intervals with larger radius. The superiority of squared-Wasserstein distance to other distances is probably due to the fact that it employs all possible points in the concerning intervals, yet L_1 distance (L_2 distance, or Hausdorff distance) uses only boundary information of intervals. The emphasize on complete information within intervals contributes to better recognization of intervals in different size and shape.

Indeed, the merits of this proposed algorithm have been well supported by comparative results in this section. To sum up, in most cases, squared-Wasserstein-distance-based ADCA has achieved decent performance in discovering underlying structure and recognizing clusters of different shapes and sizes in a data set.

6 Real data

In this section, two cases from real world will be considered to further validate the usefulness of the proposed algorithm. In the first case, comparison on values of CR index by different algorithms will be conducted again, since a prior partition has been given. Temperature changing rules in 60 stations of China is concerned in the second case.

6.1 Car data set

We firstly consider the car data set (De Carvalho et al., 2006a), concerning 33 cars featured by 8 interval-valued variables, i.e., *Price*, *Engine Capacity*, *Top Speed*, *Acceleration*, *Step*, *Length*, *Width* and *Height*. The priori partition, defined by a nominal variable *Car Category*, is shown as follows:

- *Utility*
  1-Alfa 145/U 5-Audi A3/U 12-Punto/U 13-Fiesta/U 17-Lancia Y/U
  24-Nissan Micra/U 25-Corsa/U 28-Twingo/U 29-Rover 25/U 31-Skoda Fabia/U
- *Sedan*
  2-Alfa 156/B 6-Audi A6/B 8-BMW serie 3/B 14-Focus/B
  21-Mercedes Classe C/B 26-Vectra/B 30-Rover 75/B 32 Skoda Octavia/B
- *Sports*
  4-Aston Martin/S 11-Ferrari /S 15-Honda NSK/S 16-Lamborghini/S
  19-Maserati GT/S 20-Mercedes SL/S 27-Porsche/S
- *Luxury*
  3-Alfa 166/L 7-Audi A8/L 9-BMW serie 5/L 10-BMW serie 7/L
  18-Lancia K/L 22-Mercedes Classe E/L 23-Mercedes Classe S/L 33-Passat/L

ADCA based on different distances, i.e., L_1 distance, L_2 distance, Hausdorff distance, squared-Wasserstein distance, and DCA using squared-Wasserstein distance, will be applied to this data set. Each algorithm will run 100 times and the best result with the lowest value of clustering criterion will be selected. CR index will be calculated to find out which algorithm achieves a partition best matching the prior partition (see Table 4).

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	CR index
L_1 (ADCA)	4/S 11/S 15/S 16/S 19/S 20/S 27/S	6/B 7/L 9/L 10/L 22/L 23/L	12/U 13/U 17/U 24/U 25/U 28/U 29/U 31/U	1/U 2/B 3/L 5/U 8/B 14/B 18/L 21/B 26/B 30/B 32/B 33/L	0.5623
L_2 (ADCA)	4/S 11/S 15/S 16/S 19/S 20/S 27/S	6/B 7/L 9/L 10/L 22/L 23/L	12/U 13/U 17/U 24/U 25/U 28/U 29/U 31/U	1/U 2/B 3/L 5/U 8/B 14/B 18/L 21/B 26/B 30/B 32/B 33/L	0.5623
Hausd. (ADCA)	4/S 11/S 15/S 16/S 19/S 20/S 27/S	6/B 7/L 9/L 10/L 22/L 23/L	12/U 13/U 17/U 24/U 25/U 28/U 29/U 31/U	1/U 2/B 3/L 5/U 8/B 14/B 18/L 21/B 26/B 30/B 32/B 33/L	0.5623
sq-Wass. (DCA)	4/S 11/S 16/S	7/L 9/L 10/L 15/S 19/S 20/S 22/L 23/L 27/S	1/U 12/U 13/U 14/B 17/U 24/U 25/U 26/B 28/U 29/U 31/U 32/B	2/B 3/L 5/U 6/B 8/B 18/L 21/B 30/B 33/L	0.3884
sq-Wass. (ADCA)	4/S 11/S 15/S 16/S 19/S 20/S 27/S	6/B 7/L 9/L 10/L 22/L 23/L	12/U 13/U 17/U 24/U 25/U 28/U 29/U 31/U	1/U 2/B 3/L 5/U 8/B 14/B 18/L 21/B 26/B 30/B 32/B 33/L	0.5623

TAB. 4 – *Clustering results for the Car data set.*

Apparently, ADCA has obtained the same partition regardless of distance. But the result is quite different from the partition by DCA based on squared-Wasserstein distance. As shown in the column on the right side, ADCA gains a higher value of CR index (0.5623) than DCA (0.3884). Consequently, the superiority of ADCA using squared-Wasserstein distance has been again demonstrated in this case, compared with squared-Wasserstein-distance-based DCA.

6.2 China temperature data set

The second application concerns with monthly temperature of 60 meteorological stations of China in 1988 (Tao et al., 1997). Table 5 has displayed an outline of the data set. Rather than single-valued data, interval-valued data is used to record temperature in each month for each station, with lower/upper bounds corresponding to minimum/maximum values respectively. In such a way, analysts can learn not only average value but also variation of the temperature.

Stations	Jan.	Feb.	...	Dec.
AnQing	$[1.8, 7.1]$	$[2.1, 7.2]$	...	$[4.3, 11.8]$
BaoDing	$[-7.1, 1.7]$	$[-5.3, 4.8]$	...	$[-3.9, 5.2]$
BeiJing	$[-7.2, 2.1]$	$[-5.9, 3.8]$	...	$[-4.4, 4.7]$
BoKeTu	$[-23.4, -15.5]$	$[-24, -14]$	...	$[-21.1, -13.1]$
ChangChun	$[-16.9, -6.7]$	$[-17.6, -6.8]$	...	$[-15.9, -7.2]$
...	...	...	...	...
ZhiJiang	$[2.7, 8.4]$	$[2.7, 8.7]$	...	$[5.1, 13.3]$

TAB. 5 – *Outline of China temperature data set.*

In this case, we will perform ADCA based on squared-Wasserstein distance to partition the 60 stations according to their monthly temperatures. We fix the cluster number to 5, since there are five major climate types in China, i.e., namely *Severe cold*, *Cold*, *Hot summer and cold winter*, *Mild*, and *Hot summer and warm winter* (Domrös and Peng, 1988). And we will

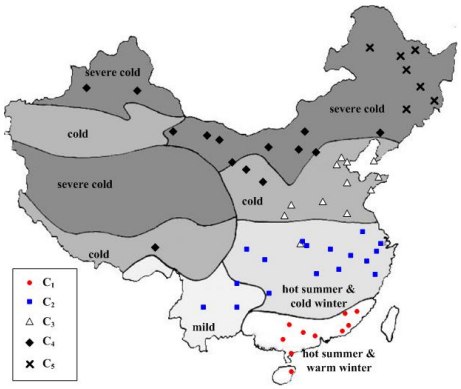

FIG. 1 – *Clustering result of China temperature data set ($K = 5$)*

examine the agreement between this prior partition and the partition obtained by ADCA based on squared-Wasserstein distance.

Figure 1 has shown the clustering result, along with geographic location of the 60 stations and the five climate zones in a map of China. Remarkably, stations of each cluster locate in one climate zone (Cluster 1 and 5) or neighboring climate zones (Cluster 2, 3 and 4). That is, the obtained partition provides a good match for the climate zones. In fact, this could be mainly accounted by latitude of each station. For instance, stations of Cluster 1 lie in south China, while stations in the northeast are grouped into Cluster 5. Nevertheless, there still exist one exception. In the third cluster, it is station of Qingjiang that mix up in the second cluster. Due to locating in river basin, this station experiences lower temperature than stations within the surrounding area. Consequently, the topography has contributed to allocating Qingjiang station to Cluster 3 (*Cold* zone), rather than Cluster 2 (*Hot summer and cold winter* zone).

For further interpretation, the corresponding adaptive factors have been listed in Table 6. For each cluster, we could make a comparison on adaptive factor values between different variables, which unveils the importance of the corresponding variable on clustering results. Take Cluster 1 as an example. The adaptive factors from June to October are greater than 1, while others less than 1, which indicates that the five months from summer to autumn have performed rather vital role on allocating this cluster. In fact, Cluster 2, 4 and 5 also share similar characteristics of high adaptive factor values from June to October, and Cluster 2 and 5 even extend such feature to spring. However, it is quite different that Cluster 3 shows lower values of adaptive factors in July and August, which may declare that stations in this cluster differ from each other in temperature during summer time.

To demonstrate the merits of ADCA based on squared-Wasserstein distance, we could further evaluate the clustering results by different methods in terms of CR index (see Table 7), given that the aforementioned five climate zones as a prior classification. As expected, ADCA based on squared-Wasserstein distance outperforms most of other methods, i.e., ADCA based on L_1 distance and Hausdorff distance, and squared-Wasserstein-distance-based DCA. Yet, ADCA using L_1 distance also performs equally well in this case.

Variables	Vectors of adaptive facotrs				
	λ^1	λ^2	λ^3	λ^4	λ^5
Jan.	0.609	0.704	0.785	0.670	0.598
Feb.	0.498	0.564	0.628	0.745	0.663
Mar.	0.719	0.998	0.782	1.083	0.619
Apr.	0.912	1.218	1.665	0.790	1.532
May	0.703	1.255	1.491	0.950	1.277
Jun.	1.736	1.212	1.139	0.701	1.428
Jul.	1.810	1.061	0.547	1.011	1.390
Aug.	2.003	1.056	0.697	2.229	1.121
Sept.	1.213	1.115	1.053	1.601	1.172
Oct.	1.379	1.205	1.497	1.221	1.180
Nov.	0.947	1.120	1.440	0.894	0.914
Dec.	0.715	0.808	1.060	0.894	0.741

TAB. 6 – *Vectors of adaptive factors for China temperature data set.*

Methods	L_1 (ADCA)	L_2 (ADCA)	*Hausd.* (ADCA)	*sq-Wass.* (DCA)	*sq-Wass.* (ADCA)
CR index	0.2491	0.4054	0.2491	0.2491	0.4054

TAB. 7 – *Comparison in CR index for China temperature data set.*

7 Conclusions

Based on squared-Wasserstein distance, this paper has contributed to presenting an adaptive dynamic clustering algorithm (ADCA) on interval-valued data. Comparative analysis on both synthetic data sets and real-life cases have revealed that (1) DCA based on squared-Wasserstein distance has been improved in performance by using adaptive factors; (2) squared-Wasserstein distance shows superiority to L_1 distance, L_2 distance and Hausdorff distance in ADCA. Consequently, the proposed algorithm could be considered as a good choice for clustering interval-valued data.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 71031001, 70771004).

References

Arfken, G. (1985). Lagrange multipliers. In J. Mathews and R. Walker (Eds.), *Mathematical Methods for Physicists,3rd ed.*, pp. 945–950. Orlando: Academic Press.

Billard, L. and E. Diday (2003). From the statistics of data to the statistics of knowledge. *Journal of the American Statistical Association 98*(462), 470–487.

Bock, H.-H. (2002). Clustering algorithms and kohonen maps for symbolic data. *J. Japanese Soc. Comput. Statist. 15*, 1–13.

Bock, H. H. and E. Diday (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin: Springer-Verlag.

Chavent, M., F. De Carvalho, Y. Lechevallier, and R. Verde (2006). New clustering methods for interval data. *Computational Statistics 21*(2), 211–229.

Chavent, M. and Y. Lechevallier (2002). Dynamical clustering algorithm of interval data: optimization of an adequacy criterion based on Hausdorff distance. In K. Jaguja, A. Sokolowsky, and H. Bock (Eds.), *Classification, Clustering and Data Analysis (IFCS2002)*, pp. 53–59. Berlin: Springer.

Costa, A., B. Pimentel, and R. Souza (2010). K-means clustering for symbolic interval data based on aggregated kernel functions. *Tools with Artificial Intelligence, IEEE International Conference on 2*, 375–376.

De Carvalho, F. (2007). Fuzzy c-means clustering methods for symbolic interval data. *Pattern Recognition Letters 28*, 423–437.

De Carvalho, F., P. Brito, and H. Bock (2006a). Dynamic clustering for interval data based on L2 distance. *Computational Statistics 21*(2), 231–250.

De Carvalho, F. and Y. Lechevallier (2009a). Dynamic clustering of interval-valued data based on adaptive quadratic distances. *IEEE transactions on systems, man, and cybernetics-Part A: Systems and Humans 39*(6), 1295–1306.

De Carvalho, F. and Y. Lechevallier (2009b). Partitional clustering algorithms for symbolic interval data based on single adaptive distances. *Pattern Recognition 42*(7), 1223 – 1236.

De Carvalho, F., R. Souza, M. Chavent, and Y. Lechevallier (2006b). Adaptive Hausdorff distances and dynamic clustering of symbolic data. *Pattern Recognition Letters 27*(3), 167–179.

De Souza, R. M. and F. De Carvalho (2004). Clustering of interval data based on city-block distances. *Pattern Recognition Letters 25*, 353–365.

De Souza, R. M., F. De Carvalho, C. Tenório, and Y. Lechevallier (2004). Dynamic cluster methods for interval data based on mahalanobis distances. In D. Banks, L. House, F. R. McMorris, P. Arabie, W. Gaul, H.-H. Bock, W. Gaul, and M. Vichi (Eds.), *Classification, Clustering, and Data Mining Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 351–360. Springer Berlin Heidelberg.

Diday, E. (1989). Introduction à l'approche symbolique en analyse des donnés. *Revue Francaise d'automatique, d'informatique et de Recherche Opérationnelle: Recherche Opérationnelle 23*(2), 193–236.

Diday, E. and M. Brito (1989). Symbolic cluster analysis. In O. Opitz (Ed.), *Conceptual and Numerical Analysis of Data*, pp. 45–84. Heidelberg: Springer.

Diday, E. and G. Govaert (1977). Classification automatique avec distances adaptatives. *RAIRO Inform. Computer Sci. 11*(4), 329–349.

Diday, E. and M. Noirhomme-Fraiture (2008). *Symbolic Data Analysis and the SODAS Software*. Chichester: Wiley-Interscience.

Diday, E. and J. Simon (1976). Clustering analysis. In K. Fu (Ed.), *Digital Pattern Recognition*, pp. 47–94. Heidelberg: Springer.

Domrös, M. and G. Peng (1988). *The climate of China*. Berlin, Germany: Springer-Verlag.

Gibbs, A. and F. Su (2002). On choosing and bounding probability metrics. *Internat. Statist. Rev. 70*, 419.

Irpino, A. and E. Romano (2007). Optimal histogram representation of large data sets: Fisher vs piecewise linear approximations. *RNTI 9*, 99–110.

Irpino, A. and R. Verde (2008). Dynamic clustering of interval data using a Wasserstein-based distance. *Pattern Recognition Letters 29*, 1648–1658.

Tao, S., C. Fu, Z. Zeng, and Q. Zhang (1997). Two long-term instrumental climatic data bases of the people's republic of china. Technical report 4699, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing. ftp://cdiac.ornl.gov/pub/ndp039/ (accessed June 2011).