Double monothetic clustering for histogram-valued data

Jaejik Kim^{1,a}, L. Billard^b

^aDepartment of Statistics, Sungkyunkwan University, Korea;

^bDepartment of Statistics, University of Georgia, USA

Abstract

One of the common issues in large dataset analyses is to detect and construct homogeneous groups of objects in those datasets. This is typically done by some form of clustering technique. In this study, we present a divisive hierarchical clustering method for two monothetic characteristics of histogram data. Unlike classical data points, a histogram has internal variation of itself as well as location information. However, to find the optimal bipartition, existing divisive monothetic clustering methods for histogram data consider only location information as a monothetic characteristic and they cannot distinguish histograms with the same location but different internal variations. Thus, a divisive clustering method considering both location and internal variation of histograms is proposed in this study. The method has an advantage in interpreting clustering outcomes by providing binary questions for each split. The proposed clustering method is verified through a simulation study and applied to a large U.S. house property value dataset.

Keywords: divisive clustering, histogram data, symbolic data

1. Introduction

One of the common issues in large dataset analyses is to detect and construct homogeneous groups of objects in those datasets. This is typically done by some form of clustering technique. In this work, a monothetic divisive clustering procedure is introduced for histogram-valued (or, simply, histogram) datasets.

Histogram-valued data are a class of symbolic data (introduced by Diday, 1987). Suppose that we are interested in classes or groups, rather than individual observations. In that case, we usually aggregate observations by each class or group, and then summary statistics for each class such as mean and median, etc. are used to analyse the data. However, the use of summary statistics might lose some valuable information. For example, credit card companies have a massive credit card transaction data for customers and they are usually interested in customers rather than individual transactions. Thus, to analyse customers, they might aggregate transaction informations by each customer and use the average amount of usage for each customer. In this case, even if two customers have the same average amount, their consumption patterns might be different and such summary statistics cannot capture customer's behaviors. To overcome this problem, if histograms of transaction amount for each customer are used for the analysis, instead of summary statistics, the loss of information can be reduced and consumption patterns of customers with the same average amount could be distinguished by histograms. In addition, if large classical datasets are transformed into histogram datasets, we can have more manageable size of data and computing can be faster by using histograms.

¹ Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-Gu, Seoul 03063, Korea. E-mail: jaejik@skku.edu

The histogram-valued data are also called multivalued quantitative data in some of the literature such as Irpino and Verde (2008). Unlike classical data which are single points in p -dimensional space $\mathbb{R}^p$, symbolic data are hypercubes (broadly defined) and/or Cartesian products of distributions in $\mathbb{R}^p$ (Billard, 2006). Detailed descriptions of such data can be found in, e.g., Bock and Diday (2000), Billard and Diday (2003, 2007), Billard (2011). In particular, a symbolic variable X whose observed value is a histogram takes the form $\xi = \{[a_k, a_{k+1}), \pi_k; k = 1, \dots, v\}$ (where $[a_k, a_{k+1})$ are the histogram based subintervals with support (e.g., relative frequency π_k). It is assumed that values within a histogram subinterval $[a_k, a_{k+1})$ are uniformly spread across the subinterval. Interval-valued data are a special case where $v = 1$ and hence $\pi_k = 1$.

There are several different divisive hierarchical clustering methods for classical data. Edwards and Cavalli-Sforza (1965) introduced a divisive clustering method that finds the optimal bipartition among the $2^{n-1} - 1$ possible bipartitions for a cluster with n objects using the intra-cluster sum of squares as a clustering criterion. Since this method considers all possible bipartitions in a cluster, it always gives a global optimal bipartition but is not computationally efficient. MacNaughton-Smith *et al.* (1964) proposed a method that iteratively uses an average distance between an object and a group of objects. Also, Har-even and Brailovsky (1995) introduced a probabilistic validation approach for divisive clustering. Williams and Lambert (1959) and Lance and Williams (1968) proposed monothetic divisive clustering methods for classical binary data.

Chavent (2000) developed divisive hierarchical clustering methods for interval and ordinal categorical data using a hierarchy of a set of objects and a monothetic characteristic of each cluster of the hierarchy. Brito and Chavent (2012) and Kim (2009) applied the monothetic method to interval and histogram data and Kim and Billard (2012) introduced a monothetic clustering method for multimodal data. Those monothetic methods reduce the amount of calculation needed to detect an optimal bipartition, so that only $p(n-1)$ bipartitions need to be examined to find the optimal bipartition rather than all the $2^{n-1} - 1$ possible bipartitions. When the number of objects is large, the number of possible bipartitions is even further reduced by the monothetic methods. Also, they provide binary questions which help to interpret clustering structures.

For histogram data, the existing monothetic methods depend only on location information for individual histograms to find the optimal bipartition. For example, the method in Brito and Chavent (2012) classifies histograms by whether cumulative relative frequencies at a certain subinterval are larger than 0.5, and Kim (2009) uses internal mean values of each histogram. However, since, unlike classical data, histograms have not only location but also variation information, clustering algorithms for histogram data should consider both the location and dispersion informations. For example, suppose that histograms have the same mean or median and different variances. In this case, the methods using only location information cannot distinguish them because they have the same location information. Thus, in this study, we propose a divisive monothetic clustering algorithm using both these characteristics of histograms.

Section 2 provides some background material needed for the clustering method. Thus, in Section 2.1, basic statistics are presented. Then, in Section 2.2, clustering criteria are defined. The algorithm is presented in Section 3 and it is verified through the well known Fisher (1936) iris dataset and a simulation study in Section 4. A large dataset is analysed in Section 5.

Table 1: An example of histogram observations with the same subintervals

Observation	[0, 2)	[2, 4)	[4, 6)	[6, 8)	[8, 10)	[10, 12)
$\mathbf{y}_1$	0.0	0.2	0.25	0.25	0.3	0.0
$\mathbf{y}_2$	0.7	0.2	0.10	0.00	0.0	0.0
$\mathbf{y}_3$	0.0	0.0	0.00	0.20	0.2	0.6

2. Background

2.1. Basic statistics

Let $\mathbf{Y} = (Y_1, \dots, Y_p)$ be a p -dimensional variable. Then, the variable Y_j , $j = 1, \dots, p$, takes histogram-valued (i.e., histogram) realizations y_{ij} , $i = 1, \dots, n$, for symbolic objects $\omega_i \in \Omega$, where $\Omega = \{\omega_i, i = 1, \dots, n\}$ is a set of individual objects ω_i , defined by

$$Y_j(\omega_i) = y_{ij} = \left\{ [a_{jk}, a_{j,k+1}), \pi_{ijk}; k = 1, \dots, v_j \right\}, \quad i = 1, \dots, n, j = 1, \dots, p,$$

where $a_{jk} \leq a_{j,k+1} \in \mathbb{R}$ and where $0 \leq \pi_{ijk} \leq 1$ and $\sum_{k=1}^{v_j} \pi_{ijk} = 1$ (Diday, 1995; Bock and Diday, 2000; Billard and Diday, 2007). In general, it is assumed that values on each subinterval are uniformly distributed. Note that we assume that the k^{th} subinterval of histograms for the variable Y_j , $[a_{jk}, a_{j,k+1})$, depends only on the subscript j . This means that all histogram observations for the variable Y_j have the same subintervals. In the symbolic data context, individual histograms for each variable can have different numbers and lengths of subintervals. While theoretically this is not a problem, computationally this can be quite burdensome. To overcome this difficulty, Kim and Billard (2013), De Carvalho and De Souza (2010) and Brito and Chavent (2012) introduced a method to make all histograms for each variable have the same subintervals. That method is based on the uniform assumption for each subinterval. However, since the shape of histograms depends on the length of subintervals, the length of subintervals, $(a_{j,k+1} - a_{jk})$, for each variable should be small enough to avoid the change of histogram shapes.

For example, suppose that we have three observed histograms for a variable Y_1 as follows: $\mathbf{y}_1 = y_{11} = \{[2, 4), 0.2; [4, 8), 0.5; [8, 10), 0.3\}$, $\mathbf{y}_2 = y_{21} = \{[0, 2), 0.7; [2, 5), 0.3\}$, $\mathbf{y}_3 = y_{31} = \{[6, 10), 0.4; [10, 12), 0.6\}$. From the given histograms, we see that each histogram has different numbers and lengths of subintervals. In order to make the three histograms have the same subintervals, the domain of Y_1 is first obtained from observed histograms. Since the minimum and maximum values of all subintervals are 0 and 12 in this example, the interval $[0, 12)$ is divided by the minimum length of all subintervals. Thus, $[0, 12)$ is divided by the minimum length 2. Thus, the identical subintervals for all histograms for Y_1 are $\{[0, 2), [2, 4), [4, 6), [6, 8), [8, 10), [10, 12)\}$. Under the uniform assumption, we can calculate supports for the identical subintervals using the overlapped proportion between the observed and identical subintervals. Also, zero is assigned to the identical subintervals that do not overlap with the observed subintervals. The histogram observations with the identical subintervals are shown in Table 1. Thus, henceforth, for computational efficiency, we assume that all histogram observations for each variable have been suitably transformed so as to have the same subintervals.

Under the assumption that values on each subinterval are uniformly distributed, Billard and Diday (2003) defined the internal mean and standard deviation for a histogram-valued observation as follows:

Definition 1. Let y_{ij} be the i^{th} histogram observation for a variable Y_j . Then, the internal mean for y_{ij} is defined by

$$M_{ij} = \sum_{k=1}^{v_j} \left(\frac{a_{jk} + a_{j,k+1}}{2} \right) \pi_{ijk}, \quad (2.1)$$

and the internal standard deviation for y_{ij} is given by

$$S_{ij} = \left[\sum_{k=1}^{v_j} \left\{ \frac{(a_{jk} - M_{ij})^2 + (a_{jk} - M_{ij})(a_{j,k+1} - M_{ij}) + (a_{j,k+1} - M_{ij})^2}{3} \right\} \pi_{ijk} \right]^{\frac{1}{2}}. \quad (2.2)$$

The internal mean of (2.1) can be a representative value for location information for a histogram and the internal standard deviation of (2.2) can be considered as its dispersion information. In our proposed clustering algorithm, both internal mean and standard deviation values of histogram observations are used to find the optimal bipartition at each stage of clustering.

2.2. Clustering criteria

Suppose that there are histogram observations $\mathbf{y}_i \in \Lambda = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ of the p -dimensional variable $\{Y_j, j = 1, \dots, p\}$ where $\mathbf{y}_i = \{y_{ij}, j = 1, \dots, p\}$, and suppose P_r is a partition of Λ at the r^{th} stage of clustering. Then, $P_r = \{C_u, u = 1, \dots, r\}$, where $C_u = \{\mathbf{y}_1, \dots, \mathbf{y}_{n_u}\}$ is a cluster with n_u observations, with $\sum_{u=1}^r n_u = n$. At the $(r+1)^{\text{th}}$ stage, a single cluster C_u in P_r is split into C_u^1 and C_u^2 . Thus, a new partition can be written as

$$P_{r+1} = (P_r \cup \{C_u^1, C_u^2\}) - \{C_u\}. \quad (2.3)$$

Now our interest is which cluster C_u is split and how to find the optimal C_u^1 and C_u^2 in C_u . If we have a partition P_r , there are $\sum_{u=1}^r (2^{n_u-1} - 1) = z$ (say) possible bipartitions for P_{r+1} . Since the number of possible bipartitions exponentially increases as n increases, for a large n , it is computationally infeasible to explore all z bipartitions. In order to solve this problem, we need a clustering criterion and strategy which can find the optimal C_u^1 and C_u^2 without having to consider all z possibilities. The criterion we adopt is minimizing the intra-cluster size that evaluates the compactness of each cluster.

Definition 2. For a cluster $C_u = \{\mathbf{y}_1, \dots, \mathbf{y}_{n_u}\}$, the *intra-cluster size* $I(C_u)$ is, for $u = 1, \dots, r$,

$$I(C_u) = \frac{1}{n \cdot n_u} \sum_{i_1=2}^{n_u} \sum_{i_2=1, i_2 < i_1}^{n_u} D^2(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}), \quad i_1, i_2 = 1, \dots, n_u, \quad (2.4)$$

where $D^2(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$ is a distance or dissimilarity measure between the observations $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$ in C_u and $n = \sum_{u=1}^r n_u$.

Various distance/dissimilarity measures between two histograms have been proposed so far. Kim and Billard (2013) extended several dissimilarity/distance measures for interval data into histogram data. Also, Verde and Irpino (2008), Arroyo *et al.* (2011), and Dias and Brito (2011) developed distance measures and applied them to statistical models for histogram data such as regression and time-series. However, in this study, we consider the squared Euclidean distance (Brito and Chavent, 2012) or the Mallows distance (Irpino and Verde, 2006) as $D^2(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$ in (2.4).

Definition 3. For a partition $P_r = (C_1, \dots, C_r)$ at the r^{th} stage, the *total intra-cluster size* is given by, for $r = 1, \dots, n-1$,

$$T(P_r) = \sum_{u=1}^r I(C_u). \quad (2.5)$$

From (2.3) and (2.5), the total intra-cluster size for P_{r+1} can be written as

$$T(P_{r+1}) = T(P_r) - \{I(C_u) - I(C_u^1) - I(C_u^2)\}. \quad (2.6)$$

Thus, minimizing $T(P_{r+1})$ is equivalent to maximizing $T(P_r) - T(P_{r+1})$ (note that $T(P_{r+1}) \leq T(P_r)$ for all r). Thus, let $\Delta_u = T(P_r) - T(P_{r+1})$ when C_u is split into (C_u^1, C_u^2) . Then,

$$\Delta_u = \{I(C_u) - I(C_u^1) - I(C_u^2)\}. \quad (2.7)$$

Therefore, clustering algorithms should efficiently find the bipartition (C_u^1, C_u^2) maximizing Δ_u .

3. Clustering algorithm

In this section, we describe a divisive clustering for histogram data with a double monothetic process at each stage. Clustering methods introduced by Chavent (2000) and Brito and Chavent (2012) use a monothetic searching algorithm to find the optimal bipartition at each stage, but they depend only on location information of histograms. However, a histogram has both location and dispersion information. That is, even when histograms have the same mean or median, they should be distinguishable if they have different dispersions. Since the clustering algorithms in Chavent (2000) and Brito and Chavent (2012) do not consider the dispersion information, they cannot provide desirable partitions for clusters with the same location but different dispersions. Thus, we propose a divisive hierarchical clustering algorithm using both location and dispersion of histograms.

In contrast to a polythetic algorithm which uses all p variables simultaneously, a monothetic algorithm basically uses one variable at a time to find the bipartition in an optimal criterion. It can also find a binary question for each variable which shows a monothetic characteristic at each stage. In general, the form of a binary question is ‘Is $Y_j \leq c_r$?’ where c_r is the cut point at the r^{th} stage.

Suppose that there are histogram observations $\mathbf{y}_i, i = 1, \dots, n$, of the p -dimensional variable $\{Y_j, j = 1, \dots, p\}$. From these histogram-valued observations, we can compute distances among all observations such as the squared Euclidean distance introduced in Brito and Chavent (2012) and the Mallows distance proposed in Irpino and Verde (2006). By using these distances, the proposed divisive hierarchical clustering method with a double monothetic process can be constructed. The double monothetic process consists of a process by the internal mean of histograms and a process by the internal standard deviation. The details of the proposed clustering algorithm are as follows:

Algorithm 1. *Double monothetic algorithm:*

1. For each variable Y_j , compute the internal mean M_{ij} and the internal standard deviation S_{ij} for all histogram observations $y_{ij}, i = 1, \dots, n, j = 1, \dots, p$, using (2.1) and (2.2), respectively.
2. Start with $P_1 \equiv \Lambda = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$. Then, set $r = 1$.
3. At the r^{th} stage, we have a partition $P_r = \{C_u, u = 1, \dots, r\}$, and each cluster contains histogram-valued observations (i.e., $C_u = \{\mathbf{y}_{ui}, i = 1, \dots, n_u\}$).
4. For each cluster $C_u, u = 1, \dots, r$, and each variable $Y_j, j = 1, \dots, p$,
 - 4-1. Sort the observations in C_u in ascending order using the internal mean values of M_{ij} . Let $\{\mathbf{y}_{u(i)}^{M_j}, i = 1, \dots, n_u\}$ be the observations in C_u sorted by the internal mean values for the variable Y_j , and let $C_{u,l}^{M_j,1} = \{\mathbf{y}_{u(1)}^{M_j}, \dots, \mathbf{y}_{u(l)}^{M_j}\}$ and $C_{u,l}^{M_j,2} = \{\mathbf{y}_{u(l+1)}^{M_j}, \dots, \mathbf{y}_{u(n_u)}^{M_j}\}, l = 1, \dots, n_u - 1$,

i.e., the first l observations of the sorted observations go into $C_{u,l}^{M_j,1}$, and the remaining $(n_u - l)$ observations go into $C_{u,l}^{M_j,2}$.

4-2. For $l = 1, \dots, n_u - 1$,

4-2-1. Compute Δ_u of (2.7) for the bipartition $(C_{u,l}^{M_j,1}, C_{u,l}^{M_j,2})$ as follows:

$$\Delta_{u,l}^{M_j} = I(C_u) - I(C_{u,l}^{M_j,1}) - I(C_{u,l}^{M_j,2}),$$

where $I(\cdot)$ is the intra-cluster size of (2.4).

4-3. Similarly to step 4-1, sort the observations in C_u in ascending order using the internal standard deviation values of S_{ij} . Let $\{\mathbf{y}_{u(i)}^{S_j}, i = 1, \dots, n_u\}$ be the observations in C_u sorted by the internal standard deviation values for the variable Y_j , and let $C_{u,l}^{S_j,1} = \{\mathbf{y}_{u(1)}^{S_j}, \dots, \mathbf{y}_{u(l)}^{S_j}\}$ and $C_{u,l}^{S_j,2} = \{\mathbf{y}_{u(l+1)}^{S_j}, \dots, \mathbf{y}_{u(n_u)}^{S_j}\}$, $l = 1, \dots, n_u - 1$.

4-4. For $l = 1, \dots, n_u - 1$,

4-4-1. Compute Δ_u of (2.7) for the bipartition $(C_{u,l}^{S_j,1}, C_{u,l}^{S_j,2})$ as follows:

$$\Delta_{u,l}^{S_j} = I(C_u) - I(C_{u,l}^{S_j,1}) - I(C_{u,l}^{S_j,2}).$$

5. Find a bipartition corresponding to Δ , where

$$\Delta = \max \{ \Delta_{u,l}^{M_j}, \Delta_{u,l}^{S_j}; u = 1, \dots, r, j = 1, \dots, p, l = 1, \dots, n_u - 1 \},$$

i.e., identify u , l , j , and $(M_j$ or $S_j)$ corresponding to Δ .

6. If $(C_{u,l}^{M_j,1}, C_{u,l}^{M_j,2})$ is identified in step 5, the cut point at the r^{th} stage, c_r , is $(M_{(l)j} + M_{(l+1)j})/2$, where $M_{(l)j}$ and $M_{(l+1)j}$ are the internal means of the histograms for the variable Y_j of $\mathbf{y}_{u(l)}^{M_j}$ and $\mathbf{y}_{u(l+1)}^{M_j}$, respectively, and the binary question is 'Is $M(Y_j) \leq c_r$?'.

If $(C_{u,l}^{S_j,1}, C_{u,l}^{S_j,2})$ is identified, the cut point $c_r = (S_{(l)j} + S_{(l+1)j})/2$, where $S_{(l)j}$ and $S_{(l+1)j}$ are the internal standard deviations of the histograms for the variable Y_j of $\mathbf{y}_{u(l)}^{S_j}$ and $\mathbf{y}_{u(l+1)}^{S_j}$, respectively, and the binary question is 'Is $S(Y_j) \leq c_r$?'.

7. Set $r = r + 1$ and repeat steps 3–6 until $r = R$ or $r = n$, where R is a prespecified value and n is the total number of observations.

Algorithm 1 investigates two aspects for each observation to find the optimal bipartition. For each cluster, it firstly explores $(n_u - 1)$ bipartitions constructed by the internal mean values of histogram observations for a variable Y_j , and then it considers another $(n_u - 1)$ bipartitions obtained by the internal standard deviation values of histograms for a variable Y_j . Thus, it includes a double monothetic process considering both location and dispersion informations of histograms and finally, at the r^{th} stage, it finds a bipartitions giving the largest reduction of the total intra-cluster size among $\sum_{u=1}^r 2p(n_u - 1)$ bipartitions. As mentioned in Section 1, there exist $\sum_{u=1}^r (2^{n_u-1} - 1)$ possible bipartitions at the r^{th} stage. In contrast, since Algorithm 1 investigates $\sum_{u=1}^r 2p(n_u - 1)$ bipartitions, the order of complexity $O(2^n)$ is reduced into $O(n)$ by the proposed algorithm. However, since the proposed algorithm does not explore all possible bipartitions at each stage, it does not guarantee the optimal clustering outcome as with other clustering algorithms.

At every stage, the proposed method provides a binary question by either the internal mean or standard deviation as a cut point and it describes the clustering structure of histogram data. If the binary question comes from the bipartition $(C_{u,l}^{M_{j,1}}, C_{u,l}^{M_{j,2}})$ constructed by internal mean values, it can be interpreted that two clusters of the bipartition have different locations for the variable Y_j and observations in $C_{u,l}^{M_{j,1}}$ have internal mean values for Y_j which are less than c_r while internal mean values for Y_j of observations in $C_{u,l}^{M_{j,2}}$ are larger than c_r . If the binary question is obtained from the bipartition $(C_{u,l}^{S_{j,1}}, C_{u,l}^{S_{j,2}})$ by internal standard deviations, it means that two clusters have different degrees of dispersion for the variable Y_j and $C_{u,l}^{S_{j,2}}$ is more dispersed than $C_{u,l}^{S_{j,1}}$. The cut point from Y_j is either the average of internal mean or standard deviation values of the histograms for two observations around the cut point with respect to Y_j .

4. Simulation study

In this section, the proposed clustering algorithm is verified through a well known dataset and a simulated dataset. To perform the the proposed clustering method, it firstly requires a distance measure for histogram data. In this study, we use the squared Euclidean distance used in Brito and Chavent (2012). The squared Euclidean distance between two histogram observations, $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$, is defined by

$$D^2(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \sum_{j=1}^p D^2(y_{i_1,j}, y_{i_2,j}) = \sum_{j=1}^p \left[\sum_{k=1}^{v_j} (\pi_{i_1,jk} - \pi_{i_2,jk})^2 \right]. \quad (4.1)$$

Consider the Iris dataset used in Fisher (1936). Originally, Fisher's iris dataset has 150 classical observations for three species (setosa, versicolor, virginica) and four variables (Y_1 = 'Sepal Length', Y_2 = 'Sepal Width', Y_3 = 'Petal Length', Y_4 = 'Petal Width'), and each species has 50 classical observations. By aggregating successive ten classical observations in the iris dataset for each histogram object, a histogram-valued dataset with 15 observations can be generated. Histogram-valued observations $\mathbf{y}_1, \dots, \mathbf{y}_5$ belong to the species 'setosa', $\mathbf{y}_6, \dots, \mathbf{y}_{10}$ are included in the species 'versicolor', and $\mathbf{y}_{11}, \dots, \mathbf{y}_{15}$ are in the species 'virginica'. Thus, each iris species is represented by five histogram-valued observations.

A dendrogram in Figure 1(a) was obtained by applying the proposed clustering algorithm to the Iris histogram dataset. The dendrogram shows that the proposed algorithm correctly classified observations into the three iris species. Also, from the binary questions in the dendrogram, the three clusters were distinguished by internal mean values of histograms. For example, the first bipartition $(C_1 = \{\mathbf{y}_1, \dots, \mathbf{y}_5\}, C_2 = \{\mathbf{y}_6, \dots, \mathbf{y}_{15}\})$ has a binary question 'Is $M(Y_4) \leq 0.74$?'. Here, C_1 corresponds to the answer 'Yes' of the binary question and C_2 is 'No'. This means that all observations in C_1 have internal mean values of 'Petal Width' less than 0.74. Conversely, internal mean values of observations in C_2 are greater than 0.74 in Y_4 . Note that, the vertical axis in the dendrogram represents the total intra-cluster size values of (2.5) for each partition, and the answer 'Yes' for each binary question corresponds to the left hand side of a split and 'No' is the right hand side.

In the Iris histogram dataset, we saw an example of clusters classified by locations of histograms. Now, we consider a case that clusters with different dispersions are overlapping. Therefore, consider the histogram observations generated from the three different bivariate normal distributions of (4.2). Each histogram-valued observation $\mathbf{y}_i$ is constructed from 100 classical observations generated from a bivariate normal distribution and each cluster has five histogram observations. Figure 1(b) illustrates

Figure 1: (a) Dendrogram from Iris histogram data; (b) A simulated dataset from three bivariate normal distributions.

Table 2: The simulation result for bivariate normal distributions

	Brito and Chavent (2012) algorithm	The proposed algorithm
# of sets that the algorithm correctly classifies all observation into 3 clusters	29	1,000

such a dataset generated from (4.2). As shown in Figure 1(b), clusters C_1 and C_2 have the same internal mean but different variance values. In contrast, C_1 and C_3 have the same variance but different locations.

$$\begin{aligned}
 C_1 &= \{y_1, \dots, y_5\} \sim \text{BVN}\left(\begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right), \\
 C_2 &= \{y_6, \dots, y_{10}\} \sim \text{BVN}\left(\begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 5 & 0.8 \\ 0.8 & 5 \end{pmatrix}\right), \\
 C_3 &= \{y_{11}, \dots, y_{15}\} \sim \text{BVN}\left(\begin{pmatrix} 10 \\ 5 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).
 \end{aligned} \tag{4.2}$$

Both the algorithm of Brito and Chavent (2012) and the proposed algorithm were applied to 1,000 sets of histogram data simulated from (4.2), and we counted the number of sets that each algorithm correctly classifies all 15 observations into the three clusters. As shown in Table 2, the proposed algorithm perfectly classified for all 1,000 simulated datasets while Brito and Chavent (2012) algorithm did not work properly. Since Brito and Chavent (2012) algorithm depends only on location information of histograms to find the optimal bipartition, it cannot distinguish clusters C_1 and C_2 which have the same location but different dispersions. In contrast, the proposed algorithm considers both the internal mean and standard deviation of histograms as location and dispersion informations; therefore, it can distinguish those clusters.

Table 3: Clusters in the U.S. house property value dataset

P_2	State
C_1	AL, AR, AZ, FL, GA, IA, ID, IL, IN, KS, KY, LA, ME, MI, MN, MO, MS, NC, ND, NE, NM, NV, OH, OK, PA, SC, SD, TN, TX, UT, WI, WV
C_2	AK, CA, CO, CT, DC, DE, HI, MA, MD, MT, NH, NJ, NY, OR, RI, VA, VT, WA, WY

5. Data analysis

To demonstrate the divisive double monothetic clustering method for histogram-valued data proposed in this study, we analyse a large dataset for house property values for 51 states of U.S. obtained from the American Community Survey Public Use Microdata Sample (ACS PUMS) 2012 (available in the US Census Bureau website; http://www2.census.gov/acs2012_1yr/pums/csv_hus.zip). The original classical data (ACS PUMS) have various and many records for each sampled household. In this analysis, we are interested in classifying the 51 U.S. states by house property values. That is, the goal of this analysis is to investigate which states have the most similar distributions of house price for various house sizes. Note that only single family houses are considered in this analysis.

To analyse house property values for each state by house sizes, we consider house sizes as the number of bedrooms for each house and aggregate the data by each state and each house size. Since the number of houses with one bedroom for each state in the original dataset is not enough to be meaningful in the present context, one bedroom houses are merged with two bedroom houses. Similarly, since each sample size of houses with five bedrooms or more is not enough, houses with five bedrooms or more are pooled as one variable. Note that all missing values for the house property value are removed. Thus, we consider four variables; Y_1 is the house property value for a house with two bedrooms or less, Y_2 is the value for a three bedroom house, Y_3 is for a four bedroom house, and Y_4 is the value for a house with five bedrooms or more. From 753,917 individual households, house property values for these 4 variables are aggregated by each state, and a histogram-valued dataset with 51 observations (states) and 4 variables ($Y_1, \dots, Y_4$) is generated from these aggregated data.

To apply our method proposed in this study to this histogram-valued dataset, we first make the histogram-valued data have the same subintervals for each variable and calculate the squared Euclidean distance values using (4.1). Then, the proposed clustering method classifies the 51 U.S. states, and the clustering result is shown in Figure 2(a).

The vertical axis of Figure 2(a) is the total intra-cluster size of (2.5) and the first five binary questions are provided in the dendrogram. The first three bipartitions (P_2, P_3, P_4) were detected by the mean values of the house price with three bedrooms for U.S. states. The fourth bipartition had a cut point as the mean value of the house price with one or two bedrooms, and fifth bipartition was obtained by the standard deviation of four bedroom houses for U.S. states.

From Figure 2(a), it seems that there are two major clusters in the dataset. To investigate the optimal number of clusters, cluster validity indexes were computed. Cluster validity indexes evaluate clustering outcomes and provide the optimal number of clusters. In this analysis, we used the Dunn and Davis-Bouldin indexes developed in Kim and Billard (2011). Figure 2(b) shows both Dunn and Davis-Bouldin index values for the clustering result in Figure 2(a). As shown in Figure 2(b), the first bipartition ($P_2 = \{C_1, C_2\}$) has the largest Dunn index and the smallest Davis-Bouldin index values. Thus, we can conclude that there are two major clusters in the dataset. Table 3 shows a list of U.S. states for the two clusters.

This bipartition was detected by the binary question ‘Is $M(Y_2) \leq \$205,595$ ’. That is, all U.S. states in C_1 have internal mean values for the three bedroom house lower than \$205,595.46, and

Figure 2: (a) Clustering result for house property values of each U.S. state.; (b) Cluster validity indexes for the clustering result.

internal mean values of states in C_2 are higher than that price. Most states in C_2 are mainly located in the north eastern region of U.S.

6. Conclusion and discussion

In this study, a divisive double monothetic clustering method for histogram data is introduced. Chavent (2000) proposed a divisive monothetic clustering method for interval data and ordered categorical data, and Brito and Chavent (2012) extended the method to histogram data. Both monothetic methods find the optimal bipartition by whether the cumulative probability is larger than 0.5. Kim (2009) also extended Chavent (2000) to histogram data by using the internal mean values to find the optimal bipartition. That is, they use only location information of histograms. However, a histogram has both location and internal variation. Thus, when two histograms have the same location but different dispersions, they should be distinguishable. To distinguish histograms in such situations, we propose a

double monothetic clustering algorithm using both internal mean and standard deviation as location and dispersion information, respectively. Also, the proposed clustering algorithm has the same order of complexity, $O(n)$, as the existing monothetic methods, and similarly to the existing methods, it provides binary questions to interpret clustering outcomes.

However, since the proposed clustering algorithm depends on the mean and standard deviation of individual histograms, it might not distinguish histograms that have the same first two moments but different higher moments. For example, there might be right skewed and left skewed histograms with the same mean and standard deviation and the proposed algorithm cannot separate these histograms. In that case, a monothetic clustering algorithm exploring the third and higher moments of histograms can be considered.

Limam *et al.* (2004) introduced a divisive monothetic clustering method for interval data using both intra-class homogeneity and inter-class discrimination criteria. Their method employs a stepwise strategy to find the best partition at each step and uses a probabilistic approach to assign objects to clusters. In contrast, the method proposed in this study uses a forward strategy and focuses on maximizing the homogeneity of clusters.

Although divisive clustering is slower than agglomerative clustering, divisive clustering usually yields more accurate hierarchies because it starts with complete information for the global distribution. In contrast, since agglomerative algorithms merge clusters based on local information, it might not capture global structures. Also, agglomerative clustering do not provide variables related to clustering outcomes while the proposed algorithm gives binary questions with splitting variables and cut points for more interpretable results.

Acknowledgement

This paper was supported by Sungkyun Research Fund, Sungkyunkwan University, 2015 (No. S-2015-1295-0000).

References

- Arroyo J, González Rivera G, Maté C, and Muñoz San Roque A (2011). Smoothing methods for histogram-valued time series: an application to value-at-risk, *Statistical Analysis and Data Mining*, **4**, 216–228.
- Billard L (2006). Symbolic data analysis: what is it? In Rizzi, A. and Vichi, M., editors, *Proceedings in Computational Statistics 2006*, pages 261–269, Rome, Italy.
- Billard L (2011). Brief overview of symbolic data and analytic issues, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **4**, 149–156.
- Billard L and Diday E (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis, *Journal of the American Statistical Association*, **98**, 470–487.
- Billard L and Diday E (2007). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, John Wiley & Sons, England.
- Bock HH and Diday E (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin.
- Brito P and Chavent M (2012). Divisive monothetic clustering for interval and histogram-valued data, In *Proceedings ICPRAM 2012-1st International Conference on Pattern Recognition Applications and Methods*, Vilamoura, Portugal.
- Chavent M (2000). Criterion-based divisive clustering for symbolic data. In Bock, H.-H. and Diday, E., editors, *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Informa-*

- tion from Complex Data, pages 299–311. Springer-Verlag, Berlin.
- De Carvalho FAT and De Souza RMCR (2010). Unsupervised pattern recognition models for mixed feature-type symbolic data, *Pattern Recognition Letters*, **31**, 430–443.
- Dias S and Brito P (2011). A new linear regression model for histogram-valued variables. In *Proceedings of the 58th ISI World Statistics Congress*, Dublin, Ireland.
- Diday E (1987). Introduction à l'approche symbolique en analyse des données. *Premières Journées Symbolique-Numérique*, CEREMADE, Université Paris IX, 21–56.
- Diday E (1995). Probabilist, possibilist and belief objects for knowledge analysis, *Annals of Operations Research*, **55**, 225–276.
- Edwards AWF and Cavalli-Sforza EL (1965). A method for cluster analysis, *Biometrics*, **21**, 362–375.
- Fisher RA (1936). The use of multiple measurements in taxonomic problems, *Annals of Human Genetics*, **7**, 179–188.
- Har-even M and Brailovsky VL (1995). Probabilistic validation approach for clustering, *Pattern Recognition Letters*, **16**, 1189–1196.
- Irpino A and Verde R (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In Batagelj V (eds), *Proceeding IFCS 2006* (pp. 185–192), Springer, Heidelberg.
- Irpino A and Verde R (2008). Dynamic clustering of interval data using a Wasserstein-based distance, *Pattern Recognition Letters*, **29**, 1648–1658.
- Kim J (2009). *Dissimilarity Measures for Histogram-valued Data and Divisive Clustering of Symbolic Objects* (PhD thesis), University of Georgia.
- Kim J and Billard L (2011). A polythetic clustering process and cluster validity indexes for histogram-valued objects, *Computational Statistics & Data Analysis*, **55**, 2250–2262.
- Kim J and Billard L (2012). Dissimilarity measures and divisive clustering for symbolic multimodal-valued data, *Computational Statistics & Data Analysis*, **56**, 2795–2808.
- Kim J and Billard L (2013). Dissimilarity measures for histogram-valued observations, *Communications in Statistics-Theory and Methods*, **42**, 283–303.
- Lance GN and Williams WT (1968). Note on a new information-statistic classificatory program, *The Computer Journal*, **11**, 195.
- Limam MM, Diday E, and Winsberg S (2004). Probabilist allocation of aggregated statistiscal units in classification trees for symbolic class description. In Banks D, House L, McMorris FR, Arabie P, and Gaul W (Eds), *Classification, Clustering and Data Mining Applications* (pp. 371–379), Springer, Heidelberg.
- MacNaughton-Smith P, Williams WT, Dale MB, and Mockett LG (1964). Dissimilarity analysis: a new technique of hierarchical sub-division, *Nature*, **202**, 1034–1035.
- Verde R and Irpino A (2008). Comparing histogram data using a Mahalanobis-Wasserstein distance. In *COMPSTAT 2008* (pp. 77–89), Physica-Verlag HD.
- Williams WT and Lambert JM (1959). Multivariate methods in plant ecology: I. Association-Analysis in Plant Communities, *Journal of Ecology*, **47**, 83–101.