Communications for Statistical Applications and Methods
2015, Vol. 22, No. 3, 265–276

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.3.265
Print ISSN 2287-7843 / Online ISSN 2383-4757

Exploratory Methods for Joint Distribution Valued Data and Their Application

Kazuto Igarashi[1,a], Hiroyuki Minami[b], Masahiro Mizuta[b]

[a]Graduate School of Information Science and Technology, Hokkaido University, Japan
[b]Information Initiative Center, Hokkaido University, Japan

Abstract

In this paper, we propose hierarchical cluster analysis and multidimensional scaling for joint distribution valued data. Information technology is increasing the necessity of statistical methods for large and complex data. Symbolic Data Analysis (SDA) is an attractive framework for the data. In SDA, target objects are typically represented by aggregated data. Most methods on SDA deal with objects represented as intervals and histograms. However, those methods cannot consider information among variables including correlation. In addition, objects represented as a joint distribution can contain information among variables. Therefore, we focus on methods for joint distribution valued data. We expanded the two well-known exploratory methods using the dissimilarities adopted Hall Type relative projection index among joint distribution valued data. We show a simulation study and an actual example of proposed methods.

Keywords: Symbolic Data Analysis (SDA), cluster analysis, multidimensional scaling, projection index, kernel density estimation

1. Introduction

These days, the development of information technology enables us to collect and store data easily and has subsequently resulted in large and complex data. A conventional analytic framework is not effective when analyzing these kinds of data.

Symbolic Data Analysis (SDA) is one of the attractive frameworks to analyze large and complex data. SDA was proposed by Diday in 1980's. Many methods have been extended in the framework of SDA (Billard and Diday, 2006; Bock and Diday, 2000; Diday and Noirhomme-Fraiture, 2008). In SDA, target objects are typically represented by aggregated data. Those are called symbolic objects; consequently, how to represent them is important to keep information in the original data. Most methods on SDA analyze objects represented as intervals or histograms. However, objects represented as intervals do not contain information of distributions because they focus on only a pair of minimum and maximum values. In the case of objects represented as histograms, they can contain information of one dimensional distributions, but they lose information among variables including correlation. In addition, objects represented as a joint distribution can contain information among variables. When we analyze data in detail, methods for objects represented as a joint distribution are effective. However, there are few methods for them on SDA.

Therefore, we propose methods for objects represented as a joint distribution. Especially, we focus on two well-known statistical methods, hierarchical cluster analysis and multidimensional scaling. In

[1] Corresponding author: Information Initiative Center, Hokkaido University, N11-W5, Kita-ku, Sapporo, Hokkaido, 060-0811, Japan. E-mail: igarashi@iic.hokudai.ac.jp

Published 31 May 2015 / journal homepage: http://csam.or.kr

these two methods, how to define dissimilarities is a common problem to grasp the characteristic structure of joint distributions. We adopt Hall Type relative projection index as dissimilarities among joint distributions to the problem. The index is used for an extension of projection pursuit which is one of the methods of dimension reduction.

This paper is organized as follows. We introduced the background and the outline of this study in this section. In section 2, we explain basic concept and the terms of SDA and introduce preceding studies. In section 3, we provide details of the proposed approaches. Section 4 is about simulation study. We show the effectiveness of the proposed methods by comparing them with existing methods. We also apply the proposed methods to telemonitoring data on Parkinson's disease patients as an actual example in section 5. Section 6 provides a conclusion.

2. Symbolic Data Analysis

In this section, we explain the basic concept and terms of SDA, and preceding studies of cluster analysis and multidimensional scaling on SDA.

2.1. Target objects

In conventional analysis, the object is described with a single value or a vector. Therefore, when multiple observation values are provided to each variable for objects, we would lose most information because we aggregate them into a summarized value like an average. In SDA, objects on conventional analysis are called individuals, and we analyze what individuals are aggregated. They are called concepts. Concepts are typically described with intervals, histograms and distributions. We especially focus on multidimensional distributions i.e., joint distributions. We call them joint distribution valued data.

2.2. Preceding studies

A lot of methods in terms of cluster analysis on SDA were proposed such as Gowda and Diday (1991) and Chavent and Lechevallier (2002). For cluster analysis which represents objects as distributions, Katayama *et al.* (2009) proposed hierarchical cluster analysis using Symmetric Kullback-Leibler divergence as dissimilarities. In addition, Terada and Yadohisa (2010) proposed non-hierarchical cluster analysis using cumulative distribution function as dissimilarities. These methods realize cluster analysis in consideration of information among variables and distributions of the original data. However, the method by Katayama *et al.* (2009) has a limitation to the available case because it assumes that target distributions are normal distributions. The method by Terada and Yadohisa (2010) must keep all probabilities of intervals because it uses cumulative distribution functions. Therefore, there is a problem that computational complexity becomes enormous in higher dimensions.

Multidimensional scaling on SDA are proposed by Groenen *et al.* (2006). The dissimilarities of the method are represented as interval valued data. Mizuta and Minami (2012) proposed multidimensional scaling in which dissimilarities are distributions. However, there are few methods in terms of multidimensional scaling in which target objects are represented as multidimensional distributions.

3. Proposed Method

In this section, we give notations and explain how to calculate the dissimilarities among joint distributions. We also show methods for hierarchical cluster analysis and multidimensional scaling using dissimilarities.

3.1. Notations

We assume that there are m concepts and i^{th} concept consists of $n_i \times p$ matrix $\mathbf{X}_i$. n_i is the number of individuals included in i^{th} concept. Each individual is described as p variables. We denote m concepts as $\mathbf{X}$.

$$\mathbf{X}_i = \begin{pmatrix} x_{i,1,1} & x_{i,1,2} & \cdots & x_{i,1,p} \\ x_{i,2,1} & x_{i,2,2} & \cdots & x_{i,2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i,n_i,1} & x_{i,n_i,2} & \cdots & x_{i,n_i,p} \end{pmatrix}, \tag{3.1}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{pmatrix}. \tag{3.2}$$

The i^{th} symbolic object is built by aggregating n_i individuals. We represent i^{th} symbolic object $\xi_i(\mathbf{z})$ as the density function on p variables. We approximate $\xi_i(\mathbf{z})$ as a joint distribution in the form of a density function from $\mathbf{X}_i$. Then, we estimate the probability density function by kernel density estimation.

3.2. Dissimilarities

It is important to define a dissimilarity among joint distributions. In the study, we adopt Hall Type relative projection index which is used for relative projection pursuit (Hiro *et al.*, 2004). Projection pursuit is a method for dimension reduction to search for low dimensional space where we found an interesting structure. The original projection pursuit assumes that the normal distribution is the most uninteresting structure. If projected data have the structure which is most different from normal distribution, we regard it as the most interesting structure. Projection pursuit uses projection index to measure the distance between distributions. There are some indices including Hall index, area index and moment index. We focus on Hall index. Hall index defines dissimilarities with the difference of density functions. One dimensional Hall index is defined as

$$J \equiv \int_{-\infty}^{\infty} \{f_\alpha(u) - \phi(u)\}^2 \, du. \tag{3.3}$$

$f_\alpha(u)$ is a probability density function of the samples projected in one dimensional space by projection vector $\boldsymbol{\alpha}$. $\phi(\cdot)$ is the probability density function of the standard normal distribution.

Projection pursuit was extended to use the standard normal distribution as well as any distributions. The method is called relative projection pursuit. Hiro *et al.* (2004) proposed a Hall Type relative projection index that provides dissimilarities between a distribution of object's samples and a distribution of referred samples;

$$\begin{aligned} I(\boldsymbol{\alpha}) &= \int_{-\infty}^{\infty} \{f_\alpha(u) - g_\alpha(u)\}^2 \, du \\ &= \int_{-\infty}^{\infty} f_\alpha(u)^2 du + \int_{-\infty}^{\infty} g_\alpha(u)^2 du - 2\int_{-\infty}^{\infty} f_\alpha(u) g_\alpha(u) du, \end{aligned} \tag{3.4}$$

where $f_\alpha(u)$ is a density function of object's samples and $g_\alpha(u)$ is that of referred samples. They are projected in one dimensional space by projection vector $\boldsymbol{\alpha}$.

We use Hall Type relative projection index to calculate dissimilarities among distribution valued data. $f_i(\boldsymbol{z})$ represents a density function of i^{th} joint distribution valued data which have p variables, where $\boldsymbol{z} = (z_1, z_2, \ldots, z_p)$. In the same way, $f_j(\boldsymbol{z})$ represents a density function of j^{th} joint distribution valued data. The dissimilarity s_{ij} is represented as

$$\begin{aligned} s_{ij} &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left\{ f_i(\boldsymbol{z}) - f_j(\boldsymbol{z}) \right\}^2 d\boldsymbol{z} \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_i(\boldsymbol{z})^2 d\boldsymbol{z} + \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_j(\boldsymbol{z})^2 d\boldsymbol{z} - 2 \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_i(\boldsymbol{z}) f_j(\boldsymbol{z}) d\boldsymbol{z}. \end{aligned} \tag{3.5}$$

We adopt kernel density estimation using normal distribution as kernel function. The distribution valued data is represented as

$$\xi_i(\boldsymbol{z}) \approx f_i(\boldsymbol{z}) = \frac{1}{(2\pi)^{\frac{p}{2}} n_i h_{i,1} h_{i,2} \cdots h_{i,p}} \sum_{k=1}^{n_i} \left\{ \prod_{r=1}^{p} \exp\left(-\frac{(z_r - x_{i,k,r})^2}{2{h_{i,r}}^2} \right) \right\}, \tag{3.6}$$

where $h_{i,1}, h_{i,2}, \ldots, h_{i,p}$ are optimal band widths of the density function $f_i(\boldsymbol{z})$ by Scott (1992) represented as

$$h_{i,r} = \left(\frac{4}{p+2} \right)^{\frac{1}{p+4}} \sigma_r n_i^{-\frac{1}{p+4}} \quad (r = 1, 2, \ldots, p), \tag{3.7}$$

where σ_r is the standard deviation of the r^{th} variable of whole object's samples.

Using the estimated density function, we transform the items;

$$\begin{aligned} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_i(\boldsymbol{z})^2 d\boldsymbol{z} &= \frac{1}{2^p \pi^{\frac{p}{2}} n_i^2 h_{i,1} h_{i,2} \cdots h_{i,p}} \sum_{k=1}^{n_i} \sum_{l=1}^{n_i} \left\{ \prod_{r=1}^{p} \exp\left(-\frac{(x_{i,k,r} - x_{i,l,r})^2}{4{h_{i,r}}^2} \right) \right\}, \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_j(\boldsymbol{z})^2 d\boldsymbol{z} &= \frac{1}{2^p \pi^{\frac{p}{2}} n_j^2 h_{j,1} h_{j,2} \cdots h_{j,p}} \sum_{k=1}^{n_j} \sum_{l=1}^{n_j} \left\{ \prod_{r=1}^{p} \exp\left(-\frac{\left(x_{j,k,r} - x_{j,l,r}\right)^2}{4{h_{j,r}}^2} \right) \right\}, \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_i(\boldsymbol{z}) f_j(\boldsymbol{z}) d\boldsymbol{z} &= \frac{1}{2^{\frac{p}{2}} \pi^{\frac{p}{2}} n_i n_j \prod_{r=1}^{p} \left(h_{i,r}^2 + h_{j,r}^2 \right)^{\frac{1}{2}}} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \left\{ \prod_{r=1}^{p} \exp\left(-\frac{\left(x_{i,k,r} - x_{j,l,r}\right)^2}{2\left({h_{i,r}}^2 + {h_{j,r}}^2\right)} \right) \right\}. \end{aligned}$$

Then, the dissimilarity s_{ij} is represented as

$$\begin{aligned} s_{ij} = {} & \frac{1}{2^p \pi^{\frac{p}{2}} n_i^2 h_{i,1} h_{i,2} \cdots h_{i,p}} \sum_{k=1}^{n_i} \sum_{l=1}^{n_i} \left\{ \prod_{r=1}^{p} \exp\left(-\frac{(x_{i,k,r} - x_{i,l,r})^2}{4{h_{i,r}}^2} \right) \right\} \\ & + \frac{1}{2^p \pi^{\frac{p}{2}} n_j^2 h_{j,1} h_{j,2} \cdots h_{j,p}} \sum_{k=1}^{n_j} \sum_{l=1}^{n_j} \left\{ \prod_{r=1}^{p} \exp\left(-\frac{\left(x_{j,k,r} - x_{j,l,r}\right)^2}{4{h_{j,r}}^2} \right) \right\} \\ & - \frac{1}{2^{\frac{p}{2}-1} \pi^{\frac{p}{2}} n_i n_j \prod_{r=1}^{p} \left(h_{i,r}^2 + h_{j,r}^2 \right)^{\frac{1}{2}}} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \left\{ \prod_{r=1}^{p} \exp\left(-\frac{\left(x_{i,k,r} - x_{j,l,r}\right)^2}{2\left({h_{i,r}}^2 + {h_{j,r}}^2\right)} \right) \right\}. \end{aligned} \tag{3.8}$$

3.3. Cluster analysis for joint distribution valued data

Cluster analysis classifies objects into some groups. The method merges similar object sequentially. The algorithm of the proposed hierarchical cluster analysis for joint distribution valued data is as follows.

Step 1: As initial state, regard m symbolic objects as m clusters.

Step 2: Calculate dissimilarities using the expression (3.8).

Step 3: Merge the most similar two objects as one new cluster.

Step 4: Calculate dissimilarities among the new cluster generated by Step 3 and the others.

Step 5: Repeat Step 3 and Step 4 until the number of clusters is one.

Here, we explain the analysis procedure. First, we build symbolic objects. Next, we generate clusters using the above algorithm. Then, we decide how to generate new clusters such as single linkage method, complete linkage method and Ward method (Haltigan, 1975). Hereafter, we adopt the Ward method. Then, we visualize the result by a dendrogram. Finally, we interpret the result.

3.4. Multidimensional scaling for joint distribution valued data

Multidimensional scaling is to visualize relationships among objects by configurations in low dimensional space. It is based on dissimilarities. In the proposed method, we use dissimilarity matrix $\mathbf{S} = \{s_{ij}\}$ as input data of Torgerson's method (Torgerson, 1958). The analysis procedure is as follows. First, we build symbolic objects and calculate dissimilarities matrix using the expression (3.8). Next, we apply Torgerson's method to dissimilarities matrix $\mathbf{S}$. Torgerson's method is based on the theorem of Young-Householder. We transform the dissimilarities matrix for the non-negative matrix $\mathbf{B} = \{b_{ij}\}$,

$$b_{ij} = -\frac{1}{2}\left(s_{ij}^2 - s_{i\cdot}^2 - s_{\cdot j}^2 + s_{\cdot\cdot}^2\right), \tag{3.9}$$

where

$$s_{i\cdot}^2 = \frac{1}{m}\sum_{j=1}^{m} s_{ij}^2, \quad s_{\cdot j} = \frac{1}{m}\sum_{i=1}^{m} s_{ij}^2, \quad s_{\cdot\cdot}^2 = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} s_{ij}^2.$$

We apply eigenvalue decomposition to $\mathbf{B}$. After that, we construct configurations using relatively large eigenvalues and eigenvectors. Finally, we visualize the result by configurations in low dimensional space and interpret it.

4. Simulation Study

We adopt the proposed methods to artificial dataset generated by copulas. Copula is a function which indicates the relationship of marginal distribution functions. In addition, we compare the results to those by three existing approaches.

We use two types of copulas, Gumbel copula and Clayton copula (Nelsen, 1999), with various parameter values. When θ indicates a parameter, Gumbel copula is represented as

$$C(\mu, \nu) = \exp\left(-\left[(-\ln\mu)^\theta + (-\ln\nu)^\theta\right]^{\frac{1}{\theta}}\right) \quad (1 \leq \theta).$$

Table 1: Copulas in the simulation

Concept	Copula	Parameter θ	Kendall's τ
1 – 5	Gumbel copula	2.5	0.6
6 – 10	Gumbel copula	5.0	0.8
11 – 15	Clayton copula	3.0	0.6
16 – 20	Clayton copula	8.0	0.8

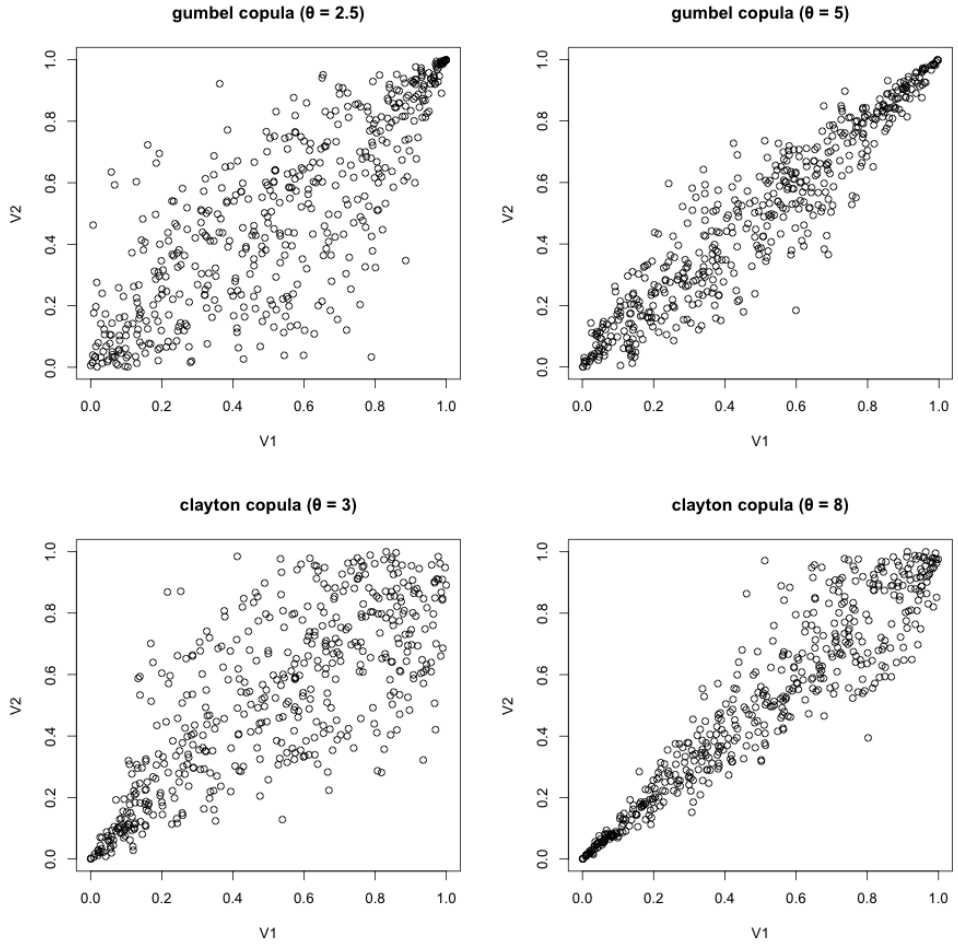

Figure 1: *Examples of copulas.*

Clayton copula is represented as

$$C(\mu, \nu) = \left[\max\left(\mu^{-\theta} + \nu^{-\theta} - 1, 0\right)\right]^{-\frac{1}{\theta}} \quad (-1 \leq \theta < 0 \text{ or } 0 \leq \theta).$$

Table 1 shows details of the copulas. There are four kinds of copula and each copula generates five concepts. Thus, there are 20 concepts in total. Each concept consists of 500 individuals. We adjust the parameters so that the values of their Kendall's τ are same. Figure 1 shows the examples of the copulas in the simulation.

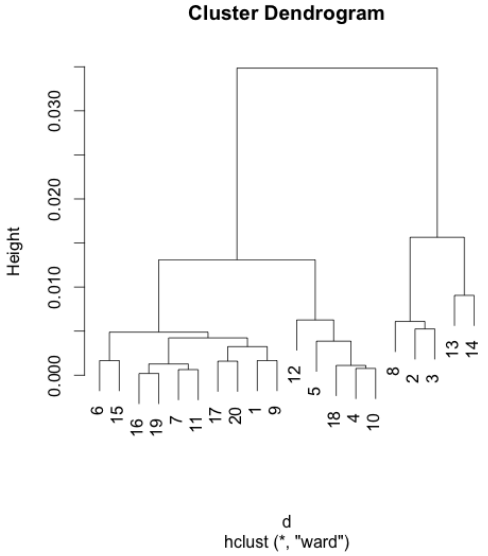

Figure 2: *Result of interval based clustering.*

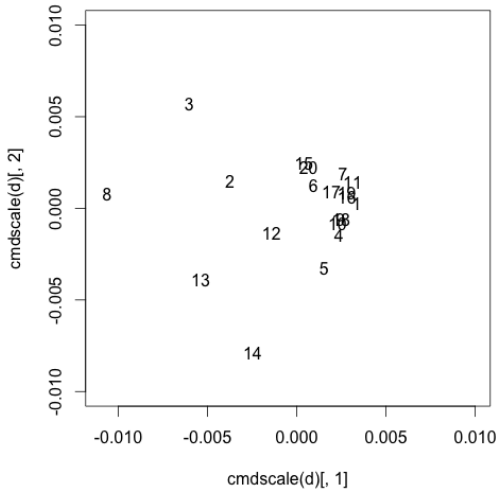

Figure 3: *Result of interval based multidimensional scaling.*

We introduce three existing methods. The 1^{st} method is interval based approach (Chavent and Lechevallier, 2002). The 2^{nd} method is histogram based approach (Diday and Noirhomme-Fraiture, 2008). The last method is distribution based approach (Katayama *et al.*, 2010). The dissimilarities on these methods are important to compare with the proposed methods.

Chavent and Lechevallier (2002) uses Hausdorff distance as dissimilarities for interval valued data. When i^{th} concept is represented as the interval $[\min_{ik}, \max_{ik}]\,(k = 1, \ldots, p)$, Hausdorff distance between concept i and j is represented as

$$s_{ij} = \sum_{k=1}^{p} \max\left[\left|\min_{ik} - \min_{jk}\right|, \left|\max_{ik} - \max_{jk}\right|\right]. \tag{4.1}$$

L_2 distance is adopted as dissimilarities for histogram valued data (Diday and Noirhomme-Fraiture, 2008). When i^{th} concept is represented as $(q_{i,k,1}, q_{i,k,2}, \ldots, q_{i,k,b_k})$; $k = 1, \ldots, p$ where $\sum_{l=1}^{b_k} q_{i,k,l} = 1$, b_k is the number of bins in the histogram for k^{th} variable. L_2 distance between concept i and j is represented as

$$s_{ij} = \sum_{k=1}^{p} \sum_{l=1}^{b_k} \left(q_{i,k,l} - q_{j,k,l}\right)^2. \tag{4.2}$$

Katayama *et al.* (2010) uses symmetric Kullback-Leibler divergence as dissimilarities for distribution valued data. When i^{th} normal distribution is represented as $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, symmetric Kullback-Leibler divergence between concept i and j is represented as

$$s_{ij} = \mathrm{tr}\left(\boldsymbol{\Sigma}_i {\boldsymbol{\Sigma}_j}^{-1}\right) + \mathrm{tr}\left(\boldsymbol{\Sigma}_j {\boldsymbol{\Sigma}_i}^{-1}\right) + \mathrm{tr}\left(\left({\boldsymbol{\Sigma}_i}^{-1} + {\boldsymbol{\Sigma}_j}^{-1}\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^{\mathrm{T}}\right) - 2p. \tag{4.3}$$

Figures 2 and 3 are the results of interval based approach, and 4 and 5 are a histogram based approach. The figures show that these existing approaches do not provide clusters based on kinds of

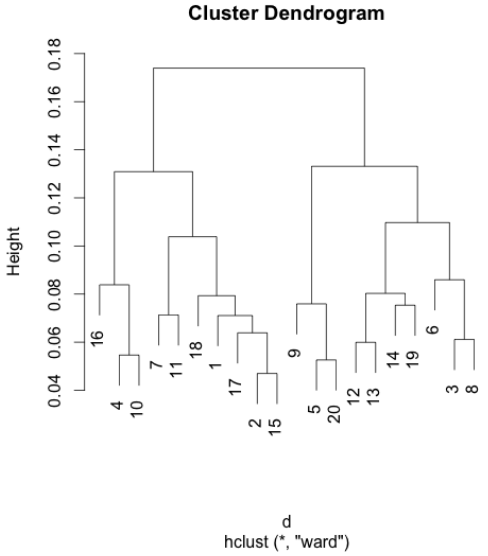

Figure 4: *Result of histogram based clustering.*

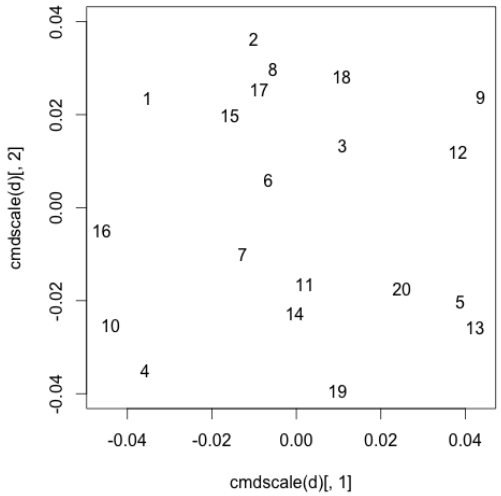

Figure 5: *Result of histogram based multidimensional scaling.*

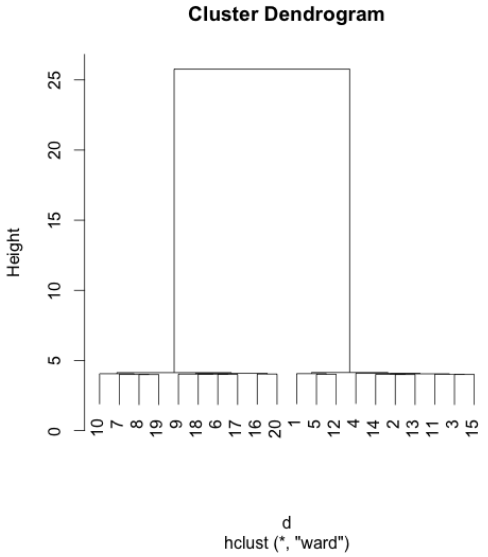

Figure 6: *Result of distribution based clustering.*

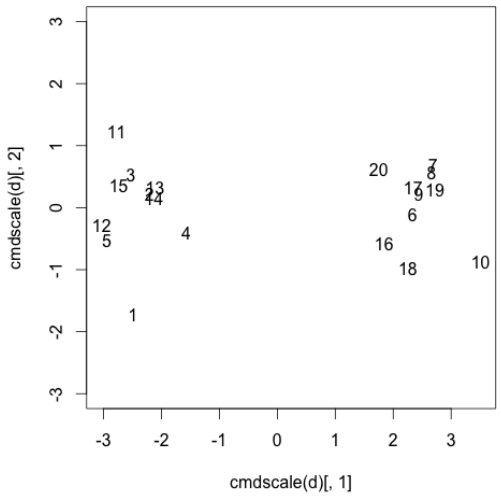

Figure 7: *Result of distribution based multidimensional scaling.*

copulas. Figures 6 and 7 are a distribution based approach. Objects are classified into two clusters based on correlation; however, this approach does not properly grasp the structure of copulas.

Figure 8 is the result of the proposed hierarchical cluster analysis. Symbolic objects are classified into four clusters by dividing at a height of 0.1. Figure 8 shows that the proposed method can grasp the structure of different correlation and copulas. Figure 9 is the result of multidimensional scaling. The vertical axis shows the differences of copulas. The horizontal axis shows those of correlation. We

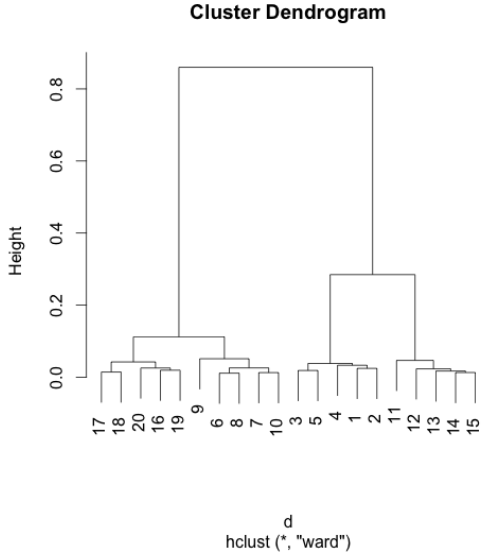

Figure 8: *Result of the proposed hierarchical cluster analysis*

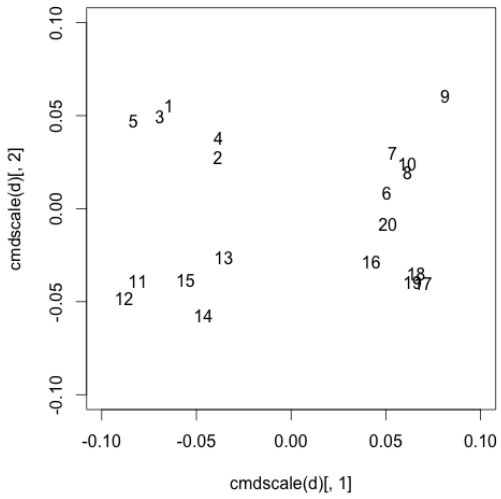

Figure 9: *Result of the proposed multidimensional scaling*

Table 2: Details of variables

Variable		Description
y_1	MDVP: Jitter(abs)	KP-MDVP absolute jitter in microseconds
y_2	MDVP: Shimmer	KP-MDVP local shimmer
y_3	NHR	Noise-to-Harmonics ratio
y_4	HNR	Harmonics-to-Noise ratio
y_5	DFA	Detrended fluctuation analysis
y_6	PPE	Pitch period entropy

can also grasp the relationships among variables using the proposed multidimensional scaling.

5. Application

In this section, we show an actual example of the proposed methods. We introduce the dataset at first; subsequently, we explain an application of the proposed methods for the dataset and interpret the results.

5.1. Dataset

We use telemonitoring data on Parkinson's disease patients. The dataset is open to the public in a Web site called UCI Machine Learning Repository. This is about voice measure in terms of the noise in patient's phonation (Tsanas *et al.*, 2010). It consists of 5,875 individuals and 20 variables, including patient's ID, sex, age and voice measures. They also contain the two evaluations diagnosed by a doctor, called UPDRS. We focus on six voice measures (Table 5.1). In the dataset, there are 42 patients (male 28, female 14) in the initial stage of Parkinson's disease. They have collected their own phonation by telemonitoring system called Intel AHTD for six months. Tsanas *et al.* (2010) tried to predict UPDRS values by regression methods using variables on voice measures.

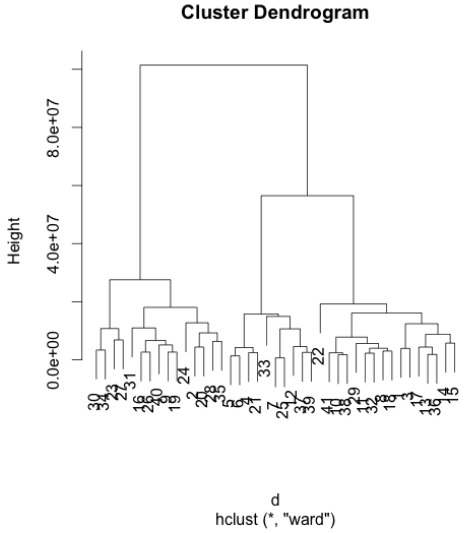

Figure 10: *Result of hierarchical cluster analysis.*

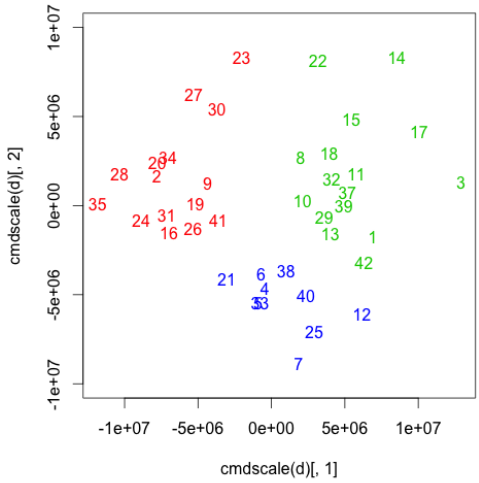

Figure 11: *Result of multidimensional scaling.*

5.2. How to build symbolic objects

We explain how to build symbolic objects to apply the proposed methods to the dataset. We regard each patient as a concept. Then, the individual is each measurement. We build joint distribution valued data by aggregated six voice measures. We excluded the 36^{th} patient from the dataset because she had too large of value and was regarded as an outlier.

Jitter (y_1) and Shimmer (y_2) are parameters to evaluate periodic disorder of the vocal-fold vibration. Jitter quantifies the fluctuation of the periodicity in every basic period. Similarly, Shimmer quantifies a fluctuation of the amplitude. NHR (y_3) and HNR (y_4) measure the power ratio of harmonics wave ingredient and noise wave ingredient from the dividing sound wave. DFA (y_5) measures the extent of turbulent noise in the speech signal, quantifying the stochastic self-similarity of the noise caused by turbulent airflow in the vocal tract (Little *et al.*, 2007). Incomplete vocal-fold closure causes a rise of DFA. PPE (y_6) measures the impaired control of stable pitch during sustained phonation (Little *et al.*, 2009). PPE is robust to confounding factors, such as smooth vibrate, which is present in healthy voices as well as dysphonia voices. Thus, this measure contributes significant information separating healthy control and Parkinson's disease patients (Tsanas *et al.*, 2010).

5.3. Results

Figure 10 shows the result of the proposed hierarchical clustering method. Patients are classified into three clusters by dividing at height of 4.0e+07. We assume that cluster 1, cluster 2 and cluster 3 sequentially from the left.

Figure 11 shows the configuration of the result of the proposed multidimensional scaling. We consider that the configuration in the two dimensional space contain significant information because the cumulative contribution ratio of first and second eigenvalues is over 90%. Figure 11 also shows the structure of three clusters with Figure 10. In the figure, the patient number is plotted with the respective cluster's color (cluster 1: red, cluster 2: blue, cluster 3: green).

We interpret the results. Figure 12 is a matrix of scatterplot of all individuals. DFA values of cluster 1 are high. This means that phonation of the patients in cluster 1 shows a self-similarity of the

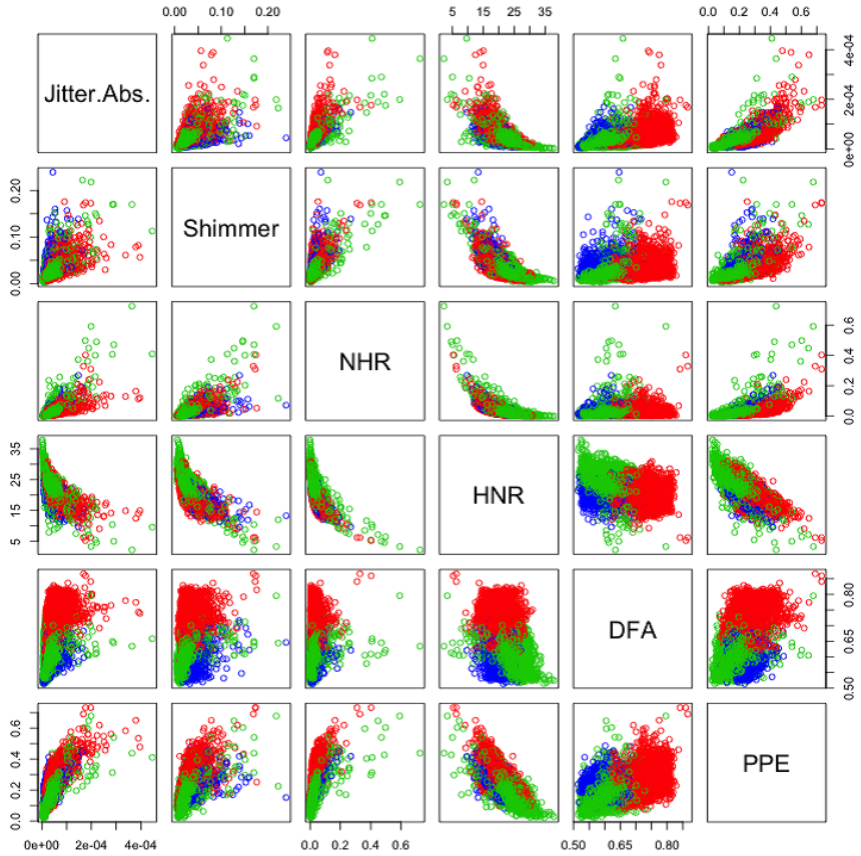

Figure 12: *Pairs plot of individuals.*

Table 3: Interpretation of clusters

Cluster	Interpretation
1	Patients who have self-similarity of the noise in their phonation.
2	Patients who have large amount of the noise in their phonation.
3	Patients who have relatively few symptoms of Parkinson's disease.

noise. However, DFA values and HNR values of cluster 2 are low. This means that phonation of the patients in cluster 2 does not show self-similarity of the noise. But, much noise is contained in their phonation. Most values of the variables in cluster 3 are relatively low except HNR.

Now, we summarize the interpretation in Table 5.2. The configuration has two directions which show voice features. One shows the degree of self-similarity of the noise in their phonation, the other shows the amount of the noise.

6. Concluding Remarks

In this paper, we proposed hierarchical cluster analysis and multidimensional scaling for joint distribution valued data in the framework of SDA. We can keep information among variables including

correlation by representing objects as joint distribution valued data. In addition, we expanded existing methods using the dissimilarities adopted Hall Type relative projection index among joint distribution valued data. We investigated the effectiveness of the proposed methods by simulation study using copulas. We compared the proposed methods and existing approaches; in addition, we also applied the proposed methods to telemonitoring data on Parkinson's disease patients. We confirmed that the clusters and the configuration grasped the voice characteristics of Parkinson's disease patients.

References

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, John Wiley & Sons, Chichester.

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

Chavent, M. and Lechevallier, Y. (2002). Dynamical clustering of interval data: Optimization of an adequacy criterion based on Hausdorff distance. In K. Jajuga, A. Sokoowski, and H. H. Bock (Eds.), *Classification, Clustering and Data Analysis*, Springer, 53–59.

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

Gowda, K. C. and Diday, E. (1991). Symbolic clustering using a new dissimilarity measure, *Pattern Recognition*, **24**, 567–578.

Groenen, P. J. F., Winsberg, S., Rodriguez, O. and Diday, E. (2006). I-Scal: Multidimensional scaling of interval dissimilarities, *Computational Statistics and Data Analysis*, **51**, 360–378.

Hartigan, J. A. (1975). *Clustering Algorithms*, John Wiley & Sons, New York.

Hiro, S., Komiya, Y., Minami, H. and Mizuta, M. (2004). Multidimensional relative projection pursuit, *Japanese Journal of Applied Statistics*, **33**, 225-241.

Katayama, K., Minami, H. and Mizuta, M. (2010). Hierarchical symbolic clustering for distribution valued data, *Journal of the Japanese Society of Computational Statistics*, **22**, 83–89.

Little, M., McSharry, P. E., Hunter, E. J., Spielman, J. and Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease, *IEEE Transactions on BioMedical Engneering*, **56**, 1-19.

Little, M., McSharry, P. E., Roberts, S. J., Castello, D. and Moroz, I. M. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection, *Biomedical Engineering OnLine*, **6**, 1015–1022.

Mizuta, M. and Minami, H. (2012). Analysis of distribution valued dissimilarity data. In W. A. Gaul, A. Geyer-Schulz, L. Schmidt-Thieme, and J. Kunze (Eds.), *Challenges at the Interface of Data Analysis, Computer Science, and Optimization* (pp. 23–28), Springer, Berlin.

Nelsen, R. B. (1999). *An Introduction to Copulas*, Springer, New York.

Scott, D. W. (1992). *Multivariate Density Estimation*, John Wiley & Sons, New York.

Terada, Y. and Yadohisa, H. (2010). Non-hierarchical clustering for distribution-valued data, In *Proceedings of COMPSTAT 2010* (pp. 1653–1660), Physical-Verlag, Berlin.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*, Wile, New York.

Tsanas, A. Little, M. A. McSharry, P. E. and Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests, *IEEE Transactions on Biomedical Engineering*, **57**, 884–893.

UCI Machine Learning Repository (2015). Available from: http://archive.ics.uci.edu/ml/

Received January 23, 2015; Revised March 12, 2015; Accepted March 31, 2015