Clustering High Dimension Low Sample-Size Data with Fuzzy Cluster-based Principal Component Analysis

Sato-Ilic, Mika
Faculty of Systems and Information Engineering,
University of Tsukuba,
Tennodai 1-1-1,
Tsukuba, 305-8573, Japan
E-mail: mika@risk.tsukuba.ac.jp

Introduction

We propose a method of principal component analysis (PCA) for high dimension low sample-size data based on the classification of data. The aim of PCA is to summarize the latent similarity structure of data observed in high dimensional space by projecting the data into a much smaller dimensional space. However, if the number of dimensions (variables) is much larger than the number of objects (sample size), then we can not obtain any solution of PCA since the variance-covariance matrix is singular. In order to solve this problem, we propose a variable selection criterion to reduce the number of variables based on an external criterion for a classification of data and a transformation method based on a classification of variables. Since the classical PCA is based on orthogonal projection, the metric projection defined in convex space, so the data space, is non-expansive. Therefore, a norm between two projected objects in a smaller dimensional space is inevitably smaller than the norm between the corresponding pre-projected two objects in a high dimensional space. The root cause of this problem is that PCA only focuses on minimizing the sum of square of distances from objects in a high dimensional space to a hyper plane in a lower dimensional space, and does not consider similarities among objects in a high dimensional space. In order to solve this problem, we extract the similarity structure of objects in a high dimensional space by using a fuzzy clustering method. By tacking the result to the PCA, we have proposed a new PCA considering the similarity structure of objects in a high dimensional space. [5] We apply this PCA to the transformed data.

Criterion for the variable selection

Suppose the observed data x_{ai} which are values of n objects (samples) with respect to p variables (dimensions) are denoted by the following:

$$X = (x_{ai}), \ \ i = 1, \cdots, n, \ a = 1, \cdots, p. \tag{1}$$

We discuss data when p is much larger than n, often written $p >> n$. This data is supposed to have an external criterion for classification that is data is labeled into K clusters. The labeled data are shown as follows:

$$X_k = (x_{ai_k}), \ \ i_k = 1, \cdots, n_k, \ \ a = 1, \cdots, p, \ \ k = 1, \cdots, K, \tag{2}$$

where $\sum_{k=1}^{K} n_k = n$. Objects in X is ordered according to the label's order. We propose a variable selection criterion to reduce the number of variables based on the external criterion of the classification as follows:

$$C(a) = \frac{1}{n}(\sum_{i_1=1}^{n_1} u_{i_1 1a} + \cdots + \sum_{i_K=1}^{n_K} u_{i_K Ka}), \ \ a = 1, \cdots, p, \tag{3}$$

where $u_{i_k ka}$ shows degree of belongingness of an object i_k to a cluster k with respect to a variable a. The object i_k corresponds to an object labeled to a cluster k which is represented as $\boldsymbol{x}_{i_k} = (x_{1i_k}, \cdots, x_{pi_k})^t$

in equation (2). $u_{i_k ka}$ is assumed to satisfy the following conditions:

$$u_{i_k ka} \in [0,1], \ \ \forall i_k, k, a, \quad \sum_{k=1}^{K} u_{i_k ka} = 1, \ \ \forall i_k, a. \tag{4}$$

From equation (4), the criterion shown in equation (3) can show how the obtained classification structure at each variable adjusts to the given external classification structure and $0 \leq C(a) \leq 1$. The larger value of $C(a)$ shows the larger explanation power for the external classification information. Therefore, using a threshold for $C(a)$, we can select variables capable of explaining the external classification information of data. In order to obtain the clustering results $u_{i_k ka}$, we use a fuzzy clustering. We use the u_{ika} as a general notation. Suppose d_{ija} is (i,j)-th element of a distance matrix D_a and shows dissimilarity between objects i and j with respect to a variable a. This is defined as $D_a = (d_{ija}), \ d_{ija} = \sqrt{(x_{ai} - x_{aj})^2}, \ \ i, j = 1, \cdots, n, \ \ a = 1, \cdots, p$. For the fuzzy clustering method in which the target data is dissimilarity data, the fanny method [3] is used. The objective function of this method is defined as follows:

$$J(\tilde{U}) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n}\sum_{j=1}^{n} (\tilde{u}_{ik})^m (\tilde{u}_{jk})^m d_{ij} / 2 \sum_{s=1}^{n} (\tilde{u}_{sk})^m \right). \tag{5}$$

Where, $\tilde{u}_{ik}$ shows degree of belongingness of an object i to a cluster k and satisfies the following conditions:

$$\tilde{u}_{ik} \in [0,1], \ \forall i, k, \ \ \sum_{k=1}^{K} \tilde{u}_{ik} = 1, \forall i. \tag{6}$$

$m, \ (1 < m < \infty)$ shows a control parameter which can control fuzziness of the belongingness. d_{ij} shows dissimilarity between objects i and j. The purpose of this method is to estimate $\tilde{U} = (\tilde{u}_{ik})$ which minimize equation (5). In equation (5), the objective function with respect to a variable a is redefined as follows:

$$J(U_a) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n}\sum_{j=1}^{n} (u_{ika})^m (u_{jka})^m d_{ija} / 2 \sum_{s=1}^{n} (u_{ska})^m \right), \ \ a = 1, \cdots, p. \tag{7}$$

Where $U_a, \ (a = 1, \cdots, p)$ is a matrix for a-th variable whose element u_{ika} shows degree of belongingness of an object i to a cluster k with respect to a variable a. u_{ika} can be estimated by minimizing equation (7) under the conditions shown in equation (4).

Clustering variables to categories

If $p >> n$, then the variable selection has a problem; when the threshold value for $C(a)$ is large, loss of the data information will be large, consequently the remained variables are not sufficient to explain the data structure. Likewise, when the threshold value for $C(a)$ is small, then still we have a problem of $p > n$. In order to solve this problem, we propose a method to transform the remained data after the variable selection to form data as $p < n$ without deleting any variables. Suppose the remained data after the variable selection shown in the previous section as follows:

$$\tilde{X} = (\tilde{x}_{ai}), \ \ i = 1, \cdots, n, \ a = 1, \cdots, \tilde{p}, \tag{8}$$

where $\tilde{p} < p$, however it is still $\tilde{p} > n$. First, we transform the data to include the external classification information of objects. We use interval to represent each cluster with respect to each variable as follows:

$$Y = (y_{ak}) = ([\underline{y}_{ak}, \overline{y}_{ak}]), \ \ a = 1, \cdots, \tilde{p}, \ k = 1, \cdots, K, \tag{9}$$

where $y_{ak} = [\underline{y}_{ak}, \overline{y}_{ak}]$ shows the interval-valued data of the a-th variable with respect to a cluster k which has the minimum value $\underline{y}_{ak}$ and the maximum value $\overline{y}_{ak}$. From equations (2) and (8), $\underline{y}_{ak}$ and $\overline{y}_{ak}$ are obtained as follows: $\underline{y}_{ak} = \min_{i_k} \tilde{x}_{ai_k}$, $\overline{y}_{ak} = \max_{i_k} \tilde{x}_{ai_k}$, $a = 1, \cdots, \tilde{p}$. This means that K clusters over the objects which is given as external classification information are expressed by K intervals. In order to obtain the similarity structure of variables over the K classified objects, we classify the data shown in equation (9). The dissimilarity between $\boldsymbol{y}_a = (y_{a1}, \cdots, y_{aK})$ and $\boldsymbol{y}_b = (y_{b1}, \cdots, y_{bK})$ is defined as follows:

$$d_{ab} = \sum_{k=1}^{K} \sup\{d(x, y_{bk}) | x \in y_{ak}\}, \;\; d(x, y_{bk}) = \inf\{d(x, y) | y \in y_{bk}\}, \tag{10}$$

$$d_{ba} = \sum_{k=1}^{K} \sup\{d(y_{ak}, y) | y \in y_{bk}\}, \;\; d(y_{ak}, y) = \inf\{d(x, y) | x \in y_{ak}\}. \tag{11}$$

Where, $d(x, y)$ shows distance between x and y, $\forall x \in y_{ak}$, $\forall y \in y_{bk}$. Therefore, $d_{ab} \neq d_{ba}$, $(a \neq b)$. We use the symmetric part of the dissimilarity as follows: $\tilde{d}_{ab} = (d_{ab} + d_{ba})/2$. Applying this dissimilarity $\tilde{d}_{ab}$ to the fanny method shown in equation (5), we obtain a fuzzy clustering result

$$\tilde{U} = (\tilde{u}_{ak}), \;\; a = 1, \cdots, \tilde{p}, \;\; k = 1, \cdots, \tilde{K}, \tag{12}$$

under the conditions shown in equation (6), where $\tilde{K}$ is a number of categories (clusters) satisfied $\tilde{K} < n$. Based on the result shown in equation (12), the data shown in equation (8) is categorized into $\tilde{K}$ categories as follows: $\tilde{X}_k = \{\tilde{\boldsymbol{x}}_a \mid p_{ak} = 1\}$, $\tilde{\boldsymbol{x}}_a = (\tilde{x}_{a1}, \ldots, \tilde{x}_{an})$, $k = 1, \cdots \tilde{K}$, where p_{ak} satisfy $\tilde{u}_{ak} = \max_{1 \leq k \leq \tilde{K}} \tilde{u}_{ak} \rightarrow p_{ak} = 1$, $a = 1, \ldots, \tilde{p}$, under the condition of $\sum_{k=1}^{\tilde{K}} p_{ak} = 1$. In the case that $\max_{1 \leq k \leq \tilde{K}} \tilde{u}_{ak}$ is not unique, we select the first category which appears having the maximum degree of belongingness over the categories. We rewrite the data sets $\tilde{X}_k$ as follows:

$$\tilde{X}_k = (\tilde{x}_{a_k i}), \;\; i = 1, \cdots, n, \; a_k = 1, \cdots, \tilde{p}_k, \;\; k = 1, \cdots, \tilde{K}, \tag{13}$$

where $\sum_{k=1}^{\tilde{K}} \tilde{p}_k = \tilde{p}$. In order to create the $\tilde{K} < n$ type data, variables included to the same category is summarized for a fixed object by using an interval as follows:

$$\tilde{Y} = (\tilde{y}_{ik}) = ([\underline{\tilde{y}}_{ik}, \overline{\tilde{y}}_{ik}]), \;\; i = 1, \cdots, n, \; k = 1, \cdots, \tilde{K}, \tag{14}$$

where $\tilde{y}_{ik} = [\underline{\tilde{y}}_{ik}, \overline{\tilde{y}}_{ik}]$ shows the interval-valued data of the i-th object with respect to a cluster k (a category k) which has the minimum value $\underline{\tilde{y}}_{ik}$ and the maximum value $\overline{\tilde{y}}_{ik}$. From equation (13), $\underline{\tilde{y}}_{ik}$ and $\overline{\tilde{y}}_{ik}$ are obtained as follows: $\underline{\tilde{y}}_{ik} = \min_{a_k} \tilde{x}_{a_k i}$, $\overline{\tilde{y}}_{ik} = \max_{a_k} \tilde{x}_{a_k i}$, $i = 1, \cdots, n$. Since $\tilde{K} < n$ in equation (14), we can apply the data to PCA.

Selection of number of clusters

In order to obtain a more accurate result of PCA, we use dissimilarity structure of objects in higher dimensional space in which the objects exist. For representing the dissimilarity structure, we use classification structure of objects. Since according to the change of number of clusters, the obtained classification structure is changed, obtaining an adaptable classification structure which can represent the dissimilarity structure well is closely related with how to determine an adaptable number of clusters. The criterion of selection of an appropriate number of clusters is defined as follows:

$$C(K) = \sum_{i \neq j = 1}^{n} s_{ij} \tilde{s}_{ij}^{(K)} \Bigg/ \left(\sqrt{\sum_{i \neq j = 1}^{n} s_{ij}^2} \sqrt{\sum_{i \neq j = 1}^{n} \tilde{s}_{ij}^{(K)^2}} \right), \tag{15}$$

where s_{ij} shows a similarity between objects i and j and is calculated from data shown in equation (14) as follows: $S = (s_{ij}),\ \ s_{ij} = 1 - d_{ij}/\max_{i,j}\{d_{ij}\},\ \ i, j = 1, \cdots, n$. The dissimilarity d_{ij} between $\tilde{\boldsymbol{y}}_i = (\tilde{y}_{i1}, \cdots, \tilde{y}_{i\tilde{K}})$ and $\tilde{\boldsymbol{y}}_j = (\tilde{y}_{j1}, \cdots, \tilde{y}_{j\tilde{K}})$ is obtained by using equations (10) and (11). $\tilde{s}_{ij}^{(K)}$ shows the restored similarity obtained as follows [4]:

$$\tilde{s}_{ij}^{(K)} = \sum_{k=1}^{K} \sum_{l=1}^{K} w_{kl}^{(K)} u_{ik}^{(K)} u_{jl}^{(K)}, \tag{16}$$

where $w_{kl}^{(K)}$ is considered to be a quantity which shows the asymmetric similarity between a pair of clusters when we assume the number of clusters as K. In this paper, we define the $w_{kl}^{(K)}$ as derived from an assumption of normal distribution of objects in each cluster as follows:

$$w_{kl}^{(K)} = 1-1/(1+e^{-\tilde{w}_{kl}^{(K)}}),\ \tilde{w}_{kl}^{(K)} = \frac{1}{2}\Big(\|\boldsymbol{\mu}_{(k,K)}-\boldsymbol{\mu}_{(l,K)}\|_{\Sigma_{(k,K)}^{-1}}+tr(\Sigma_{(k,K)}^{-1}\Sigma_{(l,K)}-I)+\log\frac{|\Sigma_{(k,K)}|}{|\Sigma_{(l,K)}|}\Big), \tag{17}$$

where

$$\|\boldsymbol{\mu}_{(k,K)} - \boldsymbol{\mu}_{(l,K)}\|_{\Sigma_{(k,K)}^{-1}} = (\boldsymbol{\mu}_{(k,K)} - \boldsymbol{\mu}_{(l,K)})' \Sigma_{(k,K)}^{-1} (\boldsymbol{\mu}_{(k,K)} - \boldsymbol{\mu}_{(l,K)}),\ \ \forall k, l.$$

Where $\tilde{w}_{kl}^{(K)}$ is derived from Kullback-Leibler's divergence [2]. $w_{kl}^{(K)}$ shows the similarity from a cluster k to a cluster l when we assume the number of clusters is K. I is a unit matrix. $\boldsymbol{\mu}_{(k,K)}$ and $\Sigma_{(k,K)}$ are an expected value and a variance-covariance matrix of $S_{(k,K)}$ which is shown as follows: $S_{(k,K)} = \{\tilde{\boldsymbol{y}}_i \mid p_{ik}^{(K)} = 1\},\ \ \tilde{\boldsymbol{y}}_i = (\tilde{y}_{i1}, \ldots, \tilde{y}_{ip}),\ \tilde{y}_{ia} = (\underline{\tilde{y}}_{ia} + \overline{\tilde{y}}_{ia})/2,\ \ \forall k$, where $p_{ik}^{(K)}$ satisfy $u_{ik}^{(K)} = \max_{1 \leq k \leq K} u_{ik}^{(K)} \to p_{ik}^{(K)} = 1,\ \ i = 1, \ldots, n$, under the condition of $\sum_{k=1}^{K} p_{ik}^{(K)} = 1$. $u_{ik}^{(K)}$ shows degree of belongingness of an object i to a cluster k when we assume the number of clusters is K, and satisfy the condition (6). $u_{ik}^{(K)}$ is obtained by applying calculated symmetrized dissimilarity derived from equations (10) and (11) to a fuzzy clustering shown in equation (5). From equation (17), $w_{kl}^{(K)} \neq w_{lk}^{(K)},\ \ (k \neq l),\ w_{kl}^{(K)} \in [0, 1]$ are satisfied. $C(K)$ shown in equation (15) shows the degree of alignment between s_{ij} and $\tilde{s}_{ij}^{(K)}$. Therefore, the larger value of $C(K)$ is better when compared with several cases in which we assume several numbers of clusters shown as K. In other words, selecting the best K when we obtain the largest value of $C(K)$ means selecting the best matched latent classification structure of original similarity matrix, $S = (s_{ij})$, since $s_{ij}^{(K)}$ shown in equation (16) involves the latent classification structure of s_{ij} when the number of clusters is fixed as K. The concentration around the expected value of the criterion shown in equation (15) for the different w_{kl} has been proven. [5]

PCA based on fuzzy clustering

First, we discuss single-valued PCA which is interpreted geometrically as finding a projected space spanned by vectors that show direction of the principal components. Let L be a nonempty subset of the inner product space X. Then we define a mapping P_L from X into the subsets of L called the metric projection onto L. Then $P_L(\boldsymbol{o}_1)$ is defined as follows: $P_L(\boldsymbol{o}_1) = \{\boldsymbol{o}_2 \in L \mid \parallel \boldsymbol{o}_1 - \boldsymbol{o}_2 \parallel = d(\boldsymbol{o}_1, L)\}$, where $\boldsymbol{o}_1 \in X$ and $d(\boldsymbol{o}_1, L) = \inf_{\boldsymbol{o}_2 \in L} \parallel \boldsymbol{o}_1 - \boldsymbol{o}_2 \parallel$. Let L be a convex Chebyshev set in which for each $\boldsymbol{o}_1 \in X$, there exists at least one nearest point in L. Then P_L is nonexpansive, that is,

$$\parallel P_L(\boldsymbol{o}_1) - P_L(\boldsymbol{o}_2) \parallel \leq \parallel \boldsymbol{o}_1 - \boldsymbol{o}_2 \parallel,\ \ \forall \boldsymbol{o}_1, \boldsymbol{o}_2 \in X. \tag{18}$$

The problem of the PCA is that the metric projection only satisfies equation (18) and PCA does not consider the size of values shown as follows: $C(\boldsymbol{o}_1, \boldsymbol{o}_2) = \parallel \boldsymbol{o}_1 - \boldsymbol{o}_2 \parallel - \parallel P_L(\boldsymbol{o}_1) - P_L(\boldsymbol{o}_2) \parallel$. Our obtained data is interval-valued data. The empirical joint density function for bivariate a and b for interval-valued data has been defined [1] as follows:

$$f(\tilde{y}_k, \tilde{y}_l) = \frac{1}{n} \sum_{i=1}^{n} I_i(\tilde{y}_k, \tilde{y}_l)/||Z(i)||, \tag{19}$$

where $I_i(\tilde{y}_k, \tilde{y}_l)$ is the indicator function where each element of $(\tilde{\boldsymbol{y}}_k, \tilde{\boldsymbol{y}}_l)$ is or is not in the rectangle $Z(i) = \tilde{y}_{ik} \times \tilde{y}_{il}$ consisted of two sides which are intervals $[\underline{\tilde{y}}_{ik}, \overline{\tilde{y}}_{ik}]$ and $[\underline{\tilde{y}}_{il}, \overline{\tilde{y}}_{il}]$. $\tilde{y}_k$ and $\tilde{y}_l$ are random variables. $||Z(i)||$ is the area of this rectangle. $\tilde{\boldsymbol{y}}_k$ is k-th column vector of $\tilde{Y}$ in equation (14) and is shown as follows: $\tilde{\boldsymbol{y}}_k = (\tilde{y}_{1k}, \cdots, \tilde{y}_{nk})^t = ([\underline{\tilde{y}}_{1k}, \overline{\tilde{y}}_{1k}], \cdots, [\underline{\tilde{y}}_{nk}, \overline{\tilde{y}}_{nk}])^t$. We extend the empirical joint density function shown in equation (19) as follows:

$$\tilde{f}(\tilde{y}_k, \tilde{y}_l) = \frac{1}{n}\sum_{i=1}^{n}(w_i I_i(\tilde{y}_k, \tilde{y}_l)/||Z(i)||, \quad w_i = \sum_{k=1}^{K} u_{ik}^m / \sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^m, \quad i = 1, \cdots, n, \quad m \in (1, \infty), \tag{20}$$

where $u_{ik}, \ i = 1, \cdots, n, \ k = 1, \cdots, K$ show the obtained degree of belongingness of the objects to the clusters when K is the selected appropriate number of clusters. Then fuzzy covariance for interval-valued data between variables k and l is derived as follows:

$$\hat{c}_{kl} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\tilde{y}_k - \bar{\tilde{y}}_k)(\tilde{y}_l - \bar{\tilde{y}}_l)\tilde{f}(\tilde{y}_k, \tilde{y}_l)d\tilde{y}_k d\tilde{y}_l, \quad \bar{\tilde{y}}_k = \frac{1}{2n}\sum_{i=1}^{n}(\underline{\tilde{y}}_{ik} + \overline{\tilde{y}}_{ik}). \tag{21}$$

Substituting equation (20) into equation (21), and from equation (6), we have obtained the following:

$$\begin{aligned} \hat{c}_{kl} \quad = \quad & (1/(4n))\sum_{i=1}^{n} w_i(\overline{\tilde{y}}_{ik} + \underline{\tilde{y}}_{ik})(\overline{\tilde{y}}_{il} + \underline{\tilde{y}}_{il}) - (1/n)\bar{\tilde{y}}_l \sum_{i=1}^{n}(w_i(\overline{\tilde{y}}_{ik} + \underline{\tilde{y}}_{ik}))/2 \\ & -(1/n)\bar{\tilde{y}}_k \sum_{i=1}^{n}(w_i(\overline{\tilde{y}}_{il} + \underline{\tilde{y}}_{il}))/2 + (1/n)\bar{\tilde{y}}_k \bar{\tilde{y}}_l. \end{aligned} \tag{22}$$

From equations (6) and (20), w_i satisfy the following condition:

$$w_i > 0, \quad \sum_{i=1}^{n} w_i = 1. \tag{23}$$

In a hard clustering when $u_{ik} \in \{0, 1\}$, $\sum_{k=1}^{K} u_{ik} = 1$ is satisfied, the weights w_i in equation (20) is

$$w_i = 1/n, \quad \forall i. \tag{24}$$

Since u_{ik} satisfies conditions shown in equation (6), the weight w_i in equation (20) shows how an object is clearly classified for the obtained classification structure. If an object i is clearly classified to a cluster, then the weight w_i becomes larger, and if the classification situation with respect to an object i is an uncertainty situation, then the value of w_i becomes smaller. Therefore, it can be seen that the weights shown in equation (20) show a degree of fuzziness of the clustering with respect to each object and the proposed fuzzy covariance matrix for interval-valued data, $\hat{C} = (\hat{c}_{kl})$, $k, l = 1, \cdots, \tilde{K}$ shown in equation (22) involve a classification structure over the variables which is obtained by reflecting the dissimilarity structure of objects in a higher dimensional space shown as $\| \boldsymbol{o}_1 - \boldsymbol{o}_2 \|$ in equation (18). Then based on the covariance matrix, we obtain principal components.

Numerical example

We use gene expression data for prostate cancer [6]. The data consists of 32 objects (subjects) with respect to 12626 variables (genes) shown in equation (1). As external classification information, 32 objects are labeled into two clusters of which 23 objects are from shavings of prostate tissue with cancer and 9 objects from shavings of prostate tissue are without cancer. The purpose is to obtain the classification situation of objects in a lower dimensional space by PCA. Using the variable selection criterion shown in equation (3), we selected variables which have more than 0.8 for the criterion. 90 variables remained. Based on the classification of variables, we create transformed 32 × 6 interval-valued data shown in equation (14). For clustering this interval-valued data, we check the criterion shown in equation (15) when the number of clusters are 2, 3, and 4. The value of $C(2)$ is largest when compared with other two values of $C(3)$ and $C(4)$, so we select the number of clusters as 2. From the obtained data shown in equation (14) and obtained weights from the result of fuzzy clustering shown in equation (20), we obtained the covariance shown in equation (22). Using this covariance, we obtain the principal components shown in figure 1. From this figure, we can see the subjects classified into two clusters; 1-23 are from shavings of prostate tissue with cancer and 24-32 are from shavings of prostate tissue without cancer.

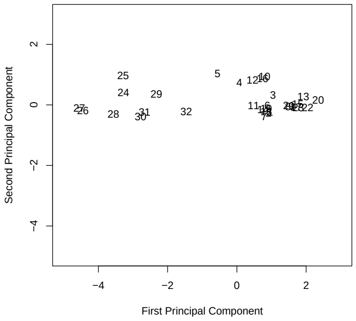

Fig. 1 Result for Proposed PCA

Table 1. Comparison of Cumulative Proportion

Proposed PCA	Centers Method
0.99	0.94

Table 1 shows a comparison of values of cumulative proportion which is the sum of the first and the second proportions corresponding to the first and the second principal components shown in the result of figures 1 and the result of the centers method [2]. This is a method of applying the data consisting of centers of intervals to the conventional PCA. This method is also identical with a method in which we use the conventional empirical joint density function shown in equation (19) and derive the covariance and then apply the obtained covariance into the conventional PCA. From equation (24), this method is the same as a case in which we use a hard clustering in a high dimensional space in our proposed PCA. From equations (23) and (24), for the fair comparison of fuzzy and hard clustering, we multiplied n to equation (22). From the result shown in table 1, we can see that the proposed PCA could obtain a better result.

Conclusion

For high dimension low sample-size data, we cannot obtain a result of conventional PCA. In order to obtain an efficient result of PCA for this type of data, we propose variable selection and data transformation methods based on classification structure. A numerical example shows a better performance of the proposed method.

REFERENCES (RÉFERENCES)

[1] L. Billard and E. Diday, Regression Analysis for Interval-Valued Data, Data Analysis, Classification, and Related Methods, Springer, pp. 369-374, 2000

[2] H.H. Bock and E. Diday (Eds.), Analysis of Symbolic Data, Springer, 2000

[3] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data, John Wiley & Sons, 1990

[4] M. Sato and Y. Sato, Extended Fuzzy Clustering Models for Asymmetric Similarity, Fuzzy Logic and Soft Computing, World Scientific, pp. 228-237, 1995

[5] M. Sato-Ilic, A Cluster-Target Similarity Based Principal Component Analysis for Interval-Valued Data, 19th International Conference on Computational Statistics, Physica-Verlag, pp. 1605-1612, 2010

[6] J.B. Welsh, L.M. Sapinoso, A.I. Su, S.G. Kern, J. Wang-Rodriguez, C.A. Moskaluk, H.F. Frierson, Jr., and G.M. Hampton, Analysis of Gene Expression Identifies Candidate Markers and Pharmacological Targets in Prostate Cancer, Cancer Research, 61, pp. 5974-5978, 2001