Simultaneous Approach to Principal Component Analysis and Fuzzy Clustering with Missing Values *

Katsuhiro Honda, Nobukazu Sugiura and Hidetomo Ichihashi
Graduate School of Engineering, Osaka Prefecture University
Sakai, Osaka, Japan
E-mail:honda@ie.osakafu-u.ac.jp

Abstract

In this paper, we propose a method for partitioning incomplete data including missing values into several fuzzy clusters using local principal components. The novel method is an extension of Fuzzy c-Varieties clustering. Numerical example shows that the method provides a tool for interpretation on the local structures of a database.

1 Introduction

Although Principal Component Analysis (PCA) is a useful technique for linear dimension reduction, it is difficult to deal with real world data including missing values. For such an incomplete data set, a simple strategy is to remove all data or attributes including missing values. The strategy, however, isn't desirable because the elimination brings a loss of information. Another technique for dealing with an incomplete data set is to impute the missing values. We can yield the maximum likelihood estimates of the missing values using EM algorithm [1] under the condition that the defects arise randomly. But the imputations are often computationally demanding. Therefore, several methods that extract principal components without elimination or imputation of data have been proposed [2] - [4]. Shibayama [3] proposed a PCA-like method to capture the structure of incomplete multivariate data without any imputations and statistical assumptions. The method is derived using the lower rank approximation of a data matrix including missing values, which accomplishes the minimization of the least square criterion.

The simultaneous approaches to the multivariate data analysis and fuzzy clustering [5] [6] have been proposed since Fuzzy c-Varieties (FCV) clustering

*Proc. of Joint 9th IFSA World Congress and 20th NAFIPS International Conference, 1810-1815 (2001)

was first proposed by Bezdek *et al.* [7] [8], which can be regarded as a simultaneous approach to PCA and the fuzzy clustering. FCV clustering partitions a data set into several linear clusters formed as linear varieties and thus we can extract local principal component vectors as the basis vectors of the prototypical linear varieties. Though it is difficult to describe characteristics of a large-scale database by only one statistical model, we often obtain a practical knowledge from local model in each cluster.

In spite of the usefulness of fuzzy clustering, it also suffers from the presence of missing values. Miyamoto *et al.* [9] proposed several methods for handling missing values in the application of Fuzzy c-Means (FCM). A basic technique is to define the distances between data points and cluster centers replacing the missing values by the weighted averages of the existent values. Another technique simply ignores the missing values and calculates the distances from the remaining coordinates.

In this paper, we propose a method for partitioning an incomplete data set including missing values into several fuzzy clusters using local principal components. First, we show that FCV is the same technique as the extraction of local principal components based on the minimization of the least square criterion, which performs the approximation of the data matrix. While the objective function of FCV is based on the minimization of the distances between data points and prototypical linear varieties, we can derive the same objective function from the least square criterion under a certain condition. Second, we propose a new technique for dealing with incomplete data sets by extending the method for the extraction of local principal components. The least square criterion is the same as that of the objective function of FCV when no missing value is involved, hence our novel technique is an extension of FCV into incomplete data sets. The first advantage of our method resides in the simplicity of calculations. We can obtain local principal components without solving eigenvalue problems. The second advantage is the flexibility of the solutions. Because the principal component vectors derived by our method have arbitrariness like factor analysis, we can also analyze the result considering some a priori knowledge.

In numerical examples, we show the characteristic properties of our method.

2 Simultaneous Approach to Principal Component Analysis and Fuzzy Clustering with Missing Values

Let $X = (x_{ij})$ denotes a $(n \times m)$ data matrix consisting of m dimensional observation of n samples and X includes missing values. Unfortunately there is no general method to deal with missing values in fuzzy clustering. Miyamoto *et al.* [9] proposed some approaches that can handle missing values in Fuzzy c-Means (FCM) [8]. A basic strategy is to replace a missing value by the weighted average of the corresponding attribution. Another simple approach is to ignore the missing values and calculate the distances from the remaining coordinates.

Timm *et al.* [10] proposed similar techniques and reported that the simple approach ignoring the missing values gave fuzzier membership assignments than the basic strategy replacing the missing values.

Ignoring the missing values, the objective function of FCM with entropy regularization [11] is written as follows:

$$\begin{aligned}\psi &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \sum_{j=1}^{m} d_{ij}(x_{ij} - b_{cj})^2 \\ &\quad +\lambda \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \log u_{ci}, \end{aligned} \tag{1}$$

where $\boldsymbol{b}_c = (b_{c1}, \cdots, b_{cm})$ is the center of the cth cluster and d_{ij} is defined by

$$d_{ij} = \begin{cases} 1 & ; x_{ij} \text{ is observed.} \\ 0 & ; x_{ij} \text{ is missing.} \end{cases} \tag{2}$$

u_{ci} is the membership with the constraint

$$\sum_{c=1}^{C} u_{ci} = 1 \quad ; i = 1, \cdots, n. \tag{3}$$

The second term of Eq.(1) is for fuzzification. The larger λ is, the fuzzier the membership assignments are.

These strategies are useful only for spherical clustering. In this paper, we enhance the method to partition an incomplete data set into several linear fuzzy clusters extracting local principal components.

2.1 Local Principal Component Analysis Using Least Square Criterion

In this section, we show that the objective function of Fuzzy c-Varieties (FCV) [7] [8] can be rewritten by using least square criterion [2] [3], which is used in Principal Component Analysis (PCA). The goal of the simultaneous approach to PCA and fuzzy clustering is to partition the data set using local principal component vectors to express local linear structures. FCV is a clustering method that partition a data set into C linear clusters. The objective function of FCV with entropy regularization consists of distances from data points to p dimensional prototypical linear varieties spanned by $\boldsymbol{a}_{cj}$'s as follows:

$$\begin{aligned}\min\ L_{fcv} &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \Big\{ (\boldsymbol{x}_i - \boldsymbol{b}_c)^{\mathrm{T}} (\boldsymbol{x}_i - \boldsymbol{b}_c) \\ &\quad - \sum_{j=1}^{p} \boldsymbol{a}_{cj}^{\mathrm{T}} R_{ci} \boldsymbol{a}_{cj} \Big\} \end{aligned}$$

$$+\lambda \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}, \tag{4}$$

$$R_{ci} \quad = \quad (\boldsymbol{x}_k - \boldsymbol{b}_c)(\boldsymbol{x}_i - \boldsymbol{b}_c)^{\mathrm{T}}, \tag{5}$$

where u_{ci} denotes the membership degree of the data point $\boldsymbol{x}_i$ to the cth cluster. Because we derive $\boldsymbol{a}_{cj}$'s as the eigenvectors of the fuzzy scatter matrix, $\boldsymbol{a}_{cj}$'s can be regarded as local principal component vectors.

In this paper, we extract the local principal components by using least square criterion. We define the least square criterion for local principal component analysis using membership u_{ci} and entropy regularization as,

$$\begin{aligned}\varphi \quad &= \quad \sum_{c=1}^{C} \mathrm{tr}\Big\{(X - Y_c)^{\mathrm{T}} U_c (X - Y_c)\Big\} \\ &\quad +\lambda \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci},\end{aligned} \tag{6}$$

where $U_c = \mathrm{diag}(u_{c1}, \cdots, u_{cn})$. $Y_c = (y_{cij})$ denotes the lower lank approximation of the data matrix X in cth cluster,

$$Y_c = F_c A_c^{\mathrm{T}} + \mathbf{1}_n \boldsymbol{b}_c^{\mathrm{T}}, \tag{7}$$

where $F_c = (\boldsymbol{f}_{c1}, \cdots, \boldsymbol{f}_{cn})^T$ is the $(n \times p)$ score matrix and $A_c = (\boldsymbol{a}_{c1}, \cdots, \boldsymbol{a}_{cp})$ is the $(m \times p)$ principal component matrix. The problem is to determine F_c, A_c and $\boldsymbol{b}_c$ so that the least square criterion is minimized.

From the necessary condition $\partial\varphi/\partial\boldsymbol{b}_c = \mathbf{0}$ for the optimality of the objective function φ, we have

$$\boldsymbol{b}_c = (\mathbf{1}_n^{\mathrm{T}} U_c \mathbf{1}_n)^{-1} X^{\mathrm{T}} U_c \mathbf{1}_n, \tag{8}$$

and Eq.(6) can be transformed into

$$\begin{aligned}\varphi \quad &= \quad \sum_{c=1}^{C}\Big\{\mathrm{tr}(X_c^{\mathrm{T}} U_c X_c) - 2\mathrm{tr}(X_c^{\mathrm{T}} U_c F_c A_c^{\mathrm{T}}) \\ &\quad +\mathrm{tr}(A_c F_c^{\mathrm{T}} U_c F_c A_c^{\mathrm{T}})\Big\} \\ &\quad +\lambda \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci},\end{aligned} \tag{9}$$

where $X_c = X - \mathbf{1}_n \boldsymbol{b}_c^{\mathrm{T}}$. From $\partial\varphi/\partial F_c = O$,

$$F_c A_c^{\mathrm{T}} A_c = X_c A_c. \tag{10}$$

Under the condition that $A_c^{\mathrm{T}} A_c = I_p$, we have $F_c = X_c A_c$ and the objective function is transformed as follows:

$$\varphi \quad = \quad \sum_{c=1}^{C}\Big\{\mathrm{tr}(X_c^{\mathrm{T}} U_c X_c) - \mathrm{tr}(A_c^{\mathrm{T}} X_c^{\mathrm{T}} U_c X_c A_c)\Big\}$$

$$
\begin{aligned}
& +\lambda \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \log u_{ci} \\
= \; & L_{fcv}. \qquad (11)
\end{aligned}
$$

Therefore it can be said that Eq.(6) is equivalent to the objective function of FCV and the minimization problem is solved by computing the p largest singular values of the fuzzy scatter matrix and their associated vectors, when the data matrix doesn't include a missing value.

2.2 Extraction of Local Principal Components from Incomplete Data Sets

When we deal with an incomplete real world data set including missing values, we cannot define the objective function of FCV composed of the distances between data points and prototypical linear varieties. Therefore we propose a new clustering method that partition an incomplete data set into several ellipsoidal fuzzy clusters using the least square criterion. Because the least square criterion with complete data set is equivalent to the objective function of FCV, the new method is an extension of FCV into incomplete data sets.

The objective function to be minimized is defined by the convex combination of Eqs.(6), (1) and the entropy term as follows:

$$
\begin{aligned}
L \; = \; & \alpha\varphi + (1-\alpha)\psi + \beta \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \log u_{ci} \\
= \; & \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \sum_{j=1}^{m} d_{ij} \Big\{ \alpha (x_{ij} - \sum_{k=1}^{p} f_{cik} a_{cjk} \\
& - b_{cj})^2 + (1-\alpha)(x_{ij} - b_{cj})^2 \Big\} \\
& + \beta \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \log u_{ci}, \qquad (12)
\end{aligned}
$$

where α is a constant which defines the tradeoff between FCM and local principal component analysis. When α is 0, Eq.(12) is equivalent to Eq.(1).

To obtain a unique solution, the objective function is minimized under the constrains that

$$
\begin{aligned}
F_c^{\mathrm{T}} U_c F_c = I_p \quad & ; \quad c = 1, \cdots, C, \qquad (13) \\
F_c^{\mathrm{T}} \mathbf{1}_n = O \quad & ; \quad c = 1, \cdots, C, \qquad (14) \\
\sum_{c=1}^{C} u_{ci} = 1 \quad & ; \quad i = 1, \cdots, n, \qquad (15)
\end{aligned}
$$

and $A_c^{\mathrm{T}} A_c$ is orthogonal.

To derive the optimal A_c and $\boldsymbol{b}_c$, we rewrite Eq.(12) as follows:

$$\begin{aligned} L &= \sum_{c=1}^{C}\sum_{j=1}^{m}\Big\{\alpha(\boldsymbol{x}_j - F_c\boldsymbol{a}_{cj} - \mathbf{1}_n b_{cj})^{\mathrm{T}} U_c D_j \\ &\quad \times(\boldsymbol{x}_j - F_c\boldsymbol{a}_{cj} - \mathbf{1}_n b_{cj}) \\ &\quad +(1-\alpha)(\boldsymbol{x}_j - \mathbf{1}_n b_{cj})^{\mathrm{T}} U_c D_j (\boldsymbol{x}_j - \mathbf{1}_n b_{cj})\Big\} \\ &\quad +\beta\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}, \end{aligned} \tag{16}$$

where

$$\begin{aligned} X &= (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_j, \cdots, \boldsymbol{x}_m), \\ D_j &= \mathrm{diag}(d_{1j}, \cdots, d_{nj}). \end{aligned}$$

From $\partial L/\partial \boldsymbol{a}_{cj} = \mathbf{0}$ and $\partial L/\partial b_{cj} = 0$, we have

$$\boldsymbol{a}_{cj} = (F_c^{\mathrm{T}} U_c D_j F_c)^{-1} F_c^{\mathrm{T}} U_c D_j (\boldsymbol{x}_j - \mathbf{1}_n b_{cj}), \tag{17}$$

$$b_{cj} = (\mathbf{1}_n^{\mathrm{T}} U_c D_j \mathbf{1}_n)^{-1} \mathbf{1}_n^{\mathrm{T}} U_c D_j (\boldsymbol{x}_j - \alpha F_c \boldsymbol{a}_{cj}). \tag{18}$$

In the same way, we can derive the optimal F_c and u_{ci}. Eq.(12) is equivalent to

$$\begin{aligned} L &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\Big\{\alpha(\boldsymbol{x}_i - A_c\boldsymbol{f}_{ci} - \boldsymbol{b}_c)^{\mathrm{T}} D_i \\ &\quad \times(\boldsymbol{x}_i - A_c\boldsymbol{f}_{ci} - \boldsymbol{b}_c) \\ &\quad +(1-\alpha)(\boldsymbol{x}_i - \boldsymbol{b}_c)^{\mathrm{T}} D_i (\boldsymbol{x}_i - \boldsymbol{b}_c)\Big\} \\ &\quad +\beta\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}, \end{aligned} \tag{19}$$

and $\partial L/\partial \boldsymbol{f}_{ci} = \mathbf{0}$ and $\partial L/\partial u_{ci} = 0$ yields

$$\boldsymbol{f}_{ci} = (A_c^{\mathrm{T}} D_i A_c)^{-1} A_c^{\mathrm{T}} D_i (\boldsymbol{x}_i - \boldsymbol{b}_c), \tag{20}$$

$$\begin{aligned} u_{ci} &= \exp\Big\{-\Big(\alpha(\boldsymbol{x}_i - A_c\boldsymbol{f}_{ci} - \boldsymbol{b}_c)^{\mathrm{T}} \\ &\quad \times D_i(\boldsymbol{x}_i - A_c\boldsymbol{f}_{ci} - \boldsymbol{b}_c) \\ &\quad +(1-\alpha)(\boldsymbol{x}_i - \boldsymbol{b}_c)^{\mathrm{T}} D_i (\boldsymbol{x}_i - \boldsymbol{b}_c)\Big)/\beta - 1\Big\}, \end{aligned} \tag{21}$$

where

$$\begin{aligned} X &= (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_n)^{\mathrm{T}}, \\ D_i &= \mathrm{diag}(d_{i1}, \cdots, d_{im}). \end{aligned}$$

The proposed algorithm can be written as follows.

Step1 Initialize $U_c, A_c, \boldsymbol{b}_c, F_c$ randomly in each cluster and normalize them so that they satisfy the constraints Eqs.(13)-(15) and $A_c^{\mathrm{T}} A_c$ is orthogonal.

Step2 Update A_c's using Eq.(17) and transform them so that each $A_c^{\mathrm{T}} A_c$ is orthogonal.

Step3 Update F_c's using Eq.(20) and normalize them so that they satisfy the constraints Eqs.(13) and (14).

Step4 Update $\boldsymbol{b}_c$'s using Eq.(18).

Step5 Update U_c's using Eq.(21) and normalize them so that Eq.(15) holds

Step6 If

$$\max_{i.c} \mid u_{ci}^{NEW} - u_{ci}^{OLD} \mid < \epsilon,$$

then stop. Otherwise, return to Step3.

Here, the score matrix F_c and the principal component matrix A_c derived by the proposed algorithm have arbitrariness like factor analysis. Assume that T is such a arbitrary orthonomal matrix as

$$T^{\mathrm{T}} T = T T^{\mathrm{T}} = I_p, \tag{22}$$

and F_c^* and A_c^* are the transformed matrices,

$$\begin{aligned} F_c^* &= F_c T, \\ A_c^* &= A_c T. \end{aligned}$$

F_c^* and A_c^* are also the solutions of the minimization problem because

$$\begin{aligned} Y_c &= F_c^* A_c^{*\mathrm{T}} + \mathbf{1}_n \boldsymbol{b}_c \\ &= F_c^{\mathrm{T}} T T^{\mathrm{T}} A_c^{\mathrm{T}} + \mathbf{1}_n \boldsymbol{b}_c \\ &= F_c A_c^{\mathrm{T}} + \mathbf{1}_n \boldsymbol{b}_c. \end{aligned} \tag{23}$$

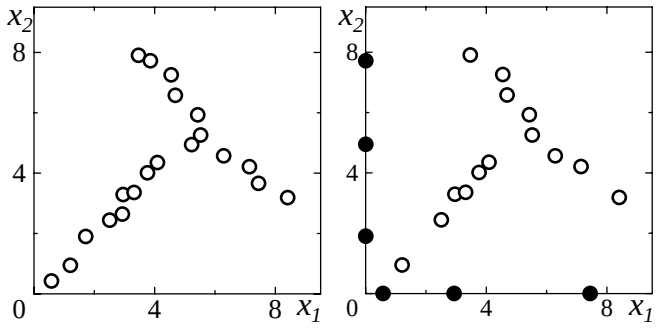
Thus the principal component vectors derived by our method have flexibilities and we can also analyze the result considering some a priori knowledge.

3 Experimental Results

3.1 Analysis of Artificial Data

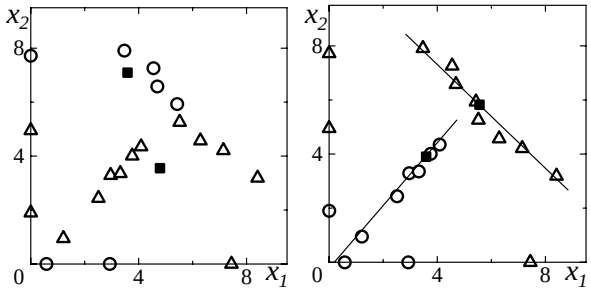
We first present a simple example to illustrate how the proposed method performs. Figure 1-a shows the original data set scattered on a plane. The data set has two local linear structures. The test set shown in Figure 1-b was made by withholding some values from the original data set. Data points including missing value are depicted on the horizontal or vertical axis, as if the missing values are 0. The results of clustering of the incomplete data set are shown in

Figure 2. Figure 2-a shows the result of FCM. The data set was partitioned into two clusters represented by $\circ$ and $\triangle$, and their centers are described by $\blacksquare$ respectively. Because FCM partitioned the data set into spherical clusters, we could not capture the local linear structures. Figure 2-b shows the result of the proposed method. The tradeoff parameter α and the weighting parameter of fuzziness β were set to 0.8 and 0.5 respectively. The data set were partitioned into two linear clusters, whose prototypical lines are represented by solid lines. In figure 2-b, the data points including missing values were also assigned to appropriate clusters.



a. original data b. incomplete data

Figure 1: scatter plots of artificial data



a. FCM b. proposed method

Figure 2: results of clustering

3.2 Analysis of "Kansei" Data

We have run more extensive analysis using with "kansei" data, which consisted of psychologically evaluated values. The data set is composed of 285 instances

in which each user evaluated 8 different graphic images used as background designs of stationery for word processor. Each user evaluated the images on a scale from 1 to 7 based on the semantic differential (SD) method as shown in Figure 3.

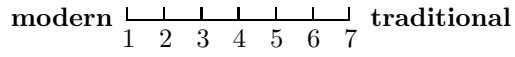


Figure 3: seven-point rating scale

Each user withheld ratings for 2 items randomly, and we extracted local principal components using the proposed clustering method. The tradeoff parameter α and the weighting parameter of fuzziness β were set to 0.8 and 0.5 respectively, and the users were partitioned into two clusters. Table 1 and 2 show the local principal component matrix A_c derived in each cluster. It seemed that the design 1 was similar to 3, 4 and 7 in cluster 1 while it was similar to 2, 5 and 6 in cluster 2.

Table 1: local principal component matrix (cluster 1)

design	A_1	
1	-2.02	12.88
2	7.37	4.03
3	-10.92	4.52
4	-0.98	11.58
5	9.46	1.30
6	7.33	-6.17
7	-10.92	-1.38
8	9.48	7.89

Table 2: local principal component matrix (cluster 2)

design	A_2	
1	3.70	11.34
2	4.12	7.40
3	8.89	4.38
4	14.74	0.87
5	3.18	10.84
6	4.48	9.02
7	17.40	-11.61
8	7.22	0.40

Then we transformed the principal component matrices A_c's by using Procrustes rotation method to test the hypothesis. Table 3 and 4 show the target matrices and the results of the rotations. The rotated matrices A_c^*'s indicated the characteristics more clearly. In this way, the rotational indeterminacy of the solution makes it possible to make in-depth analysis of the hypothesis derived by our method.

Table 3: Procrustes rotation (cluster 1)

design	target matrix		A_1^*	
1	0	1	4.16	12.36
2	1	0	8.40	0.17
3	0	1	-7.60	9.07
4	0	1	4.49	10.73
5	1	0	9.00	-3.21
6	1	0	3.65	-8.86
7	0	1	-10.33	3.82
8	1	0	12.06	2.62

Table 4: Procrustes rotation (cluster 2)

design	target matrix		A_2^*	
1	0	1	4.29	11.13
2	0	1	4.50	7.17
3	1	0	9.11	3.91
4	1	0	14.76	0.10
5	0	1	3.74	10.66
6	0	1	4.94	8.77
7	1	0	16.77	-12.50
8	1	0	7.23	0.02

4 Conclusions

In this paper, we proposed a method of handling missing values in the simultaneous application of principal component analysis and fuzzy clustering. Because the least square criterion with a complete data set is equivalent to the objective function of FCV, the new method can be regarded as an extension of FCV into incomplete data sets.

As the lower rank matrix Y_c derived in each cluster includes no missing values, we can estimate the missing values in the data matrix X using the corresponding elements of Y_c. The collaborative filtering in information system

can be seen as the problem of predicting missing values in a user-item matrix. Our future works include the application of the method to information filtering.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, Vol.39, 1977, pp 1-38

[2] T. Wiberg, "Computation of Principal Components when Data are Missing," *Proc. of 2nd Symposium on Computational Statistics*, 1976, pp 229-236

[3] T. Shibayama, "A PCA-Like Method for Multivariate Data with Missing Values," *Japanese Journal of Educational Psychology*, Vol.40, 1992, pp 257-265 (in Japanese)

[4] H. Shum, K. Ikeuchi, and R. Reddy, "Principal Component Analysis with Missing Data and its Application to Polyhedral Object Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.17, No.9, 1995, pp 854-867

[5] A. Yamakawa, K. Honda, H. Ichihashi, and T. Miyoshi, "Simultaneous Approach to Fuzzy Cluster, Principal Component and Multiple Regression Analysis," *Proc. of International Conference on Neural Networks*, 1999

[6] C.-H. Oh, H. Komatsu, K. Honda, and H. Ichihashi, "Fuzzy Clustering Algorithm Extracting Principal Components Independent of Subsidiary Variables," *Proc. of International Conference on Neural Networks*, 2000

[7] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson, "Detection and Characterization of Cluster Substructure 2. Fuzzy c-Varieties and Convex Combinations Thereof," *SIAM J. Appl. Math.*, Vol.40, No.2, 1981, pp 358-372

[8] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981

[9] S. Miyamoto, O. Takata, and K. Umayahara, "Handling Missing Values in Fuzzy c-Means," *Proc. of the Third Asian Fuzzy Systems Symposium*, 1998, pp 139-142

[10] H. Timm, and R. Kruse, "Fuzzy Cluster Analysis with Missing Values," *Proc. of 17th NAFIPS International conference*, 1998, pp 242-246

[11] S. Miyamoto, and M. Mukaidono, "Fuzzy c-Means as a Regularization and Maximum Entropy Approach," *Proc. of the 7th International Fuzzy Systems Association World Congress*, Vol.2, 1997, pp 86-92