AN APPLICATION OF FUZZY c-MEANS CLUSTERING TO PCA-LIKE METHOD FOR MISSING VALUE ESTIMATION

K.Honda[1], A.Yamakawa[2], A.Kanda[1] and H.Ichihashi[1]

1 - Graduate School of Engineering, Osaka Prefecture University
2 - Information Business Communication, Fukushima College

1 - 1-1 Gakuencho, Sakai City, Osaka 599-8531 Japan,
Phone: +81-722-54-9355, Fax: +81-722-54-9915
2 - 1-1 Miyashiro Chigoike, Fukushima City, Fukushima 960-0181 Japan,
Phone: +81-24-553-7115, Fax: +81-24-553-3222
honda@ie.osakafu-u.ac.jp, asuka@fukushima-g.ac.jp, kanda@ie.osakafu-u.ac.jp,
ichi@ie.osakafu-u.ac.jp

Key Words: principal component analysis, fuzzy clustering, missing value

ABSTRACT

In many real world applications data sets with missing values are quite common. Missing values occur for several reasons and in various situations. In this paper, we propose a new approach which extracts local principal components for missing value estimation. The new method is based on a simultaneous approach to principal component analysis and fuzzy clustering with an incomplete data set including missing values. In the simultaneous approach, we extract local principal components by using eigenvectors of the data matrix. In numerical experiment, we apply the proposed technique to the recommendation system of background designs of stationery for word processor.

1. INTRODUCTION

Missing values have frequently been encountered in data analysis in real applications. There are many approaches to handle data sets including missing values. Several methods that extract principal components without elimination or imputation of data have been proposed [1] [2] [3]. Shibayama [2] proposed a PCA(Principal Component Analysis)-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 data matrix including missing values, which accomplishes the minimization of the least square criterion.

In this paper, we propose a new approach which extracts local principal components for missing value estimation. The new method is based on the simultaneous application of PCA and fuzzy clustering, which is a technique for partitioning an incomplete data set including missing values into some fuzzy clusters by using local principal components. The simultaneous approaches [4] [5] to the multivariate data analysis and fuzzy clustering have been proposed since Fuzzy c-Varieties (FCV) clustering was first proposed by Bezdek *et al.* [6] [7], which can be regarded as a simultaneous approach to PCA and fuzzy clustering. FCV clustering partitions a data set into several linear clusters which form 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 the characteristics of a large-scale database by a single statistical model, we often obtain a practical knowledge from local model in each cluster.

In numerical experiment, we apply the proposed technique to the recommendation system of background designs of stationery for word processor and extract local principal components from the incomplete data set.

2. LOCAL PRINCIPAL COMPORNENTS FROM FUZZY SUBSET OF DATA WITH MISSING VALUES

Let Z denotes a $(N \times n)$ data matrix consisting of n dimensional observation of N samples and D_{zi} be a $n \times n$ diagonal matrix whose diagonal elements are the ith row vector of matrix Z. Partitioning N samples into C clusters, we define local linear models $Y_{ci}(n \times t)$ as

$$Y_{ci} = D_{zi}V_c + V_{0c} \quad (c = 1, \cdots, C; i = 1, \cdots, N) \tag{1}$$

where $V_c(n \times t)$ and $V_{0c}(n \times t)$ are weight matrices (t is the number of dimensions). If the following constraints(Eqs.(2) and (3)) are satisfied, the parameters are fixed unique.

$$\mathbf{1}_n^T V_{0c} = \mathbf{0}_t^T \tag{2}$$

$$V_c^T S_c V_c = I \tag{3}$$

Where $\mathbf{1}_n = (1, \cdots, 1)^T$ and $\mathbf{0}_t = (0, \cdots, 0)^T$ are n and t dimensional vectors respectively. S_c is a diagonal matrix whose diagonal elements are defined as follows:

$$s_{cj} = \sum_{i=1}^{N} u_{ci}(x_{ij} - v_{cj})^2 \tag{4}$$

where u_{ci}, which takes the value from interval $[0, 1]$, is the membership of the sample data i and $\boldsymbol{v}_c = (v_{c1}, \cdots, v_{cn})^T$ is the cluster center in the cluster c.

We suppose that D_{wi} is a diagonal matrix, and the jth diagonal element of D_{wi} takes "1" when the jth diagonal element of D_{zi} is observed, otherwise it takes "0". The objective function of PCA for an incomplete data set with missing values is defined as follows:

$$\min\ J_c = \sum_{i=1}^{N} u_{ci} \mathrm{tr}\Big\{(Y_{ci} - \mathbf{1}_n \boldsymbol{w}_{ci}^T)^T D_{wi}(Y_{ci} - \mathbf{1}_n \boldsymbol{w}_{ci}^T)\Big\} \tag{5}$$

When we suppose V_c and V_{0c} are fixed, $\hat{\boldsymbol{w}}_{ci}^T$ is derived by minimization of $\boldsymbol{w}_{ci}$.

$$\hat{\boldsymbol{w}}_{ci}^T = \mathbf{1}_n^T D_{wi} Y_{ci} / n_i \tag{6}$$

$$n_i = \mathbf{1}_n^T D_{wi} \mathbf{1}_n \tag{7}$$

Consequently, Eq.(5) is transformed into

$$g^*(V_c, V_{0c}) = \sum_i u_{ci} \mathrm{tr}(Y_{ci}^T C_i Y_{ci}) \tag{8}$$

where

$$C_i = (Q_{n/Dwi})^T D_{wi} Q_{n/Dwi} \tag{9}$$

$$Q_{n/Dwi} = I - \mathbf{1}_n \mathbf{1}_n^T D_{wi} / n_i \tag{10}$$

Eq.(8) is rewritten by substituting Eq.(1) as

$$g^*(V_c, V_{0c}) = \mathrm{tr}(V_c^T A_{c1} V_c) + 2\mathrm{tr}(V_c^T A_{c2} V_{0c}) + \mathrm{tr}(V_{0c}^T A_{c3} V_{0c}) \tag{11}$$

A_1, A_2 and A_3 are denoted by

$$A_{c1} = \sum_i u_{ci} D_{zi} C_i D_{zi} \tag{12}$$

$$A_{c2} = \sum_i u_{ci} D_{zi} C_i \tag{13}$$

$$A_{c3} = \sum_i u_{ci} C_i \tag{14}$$

From $\partial g^*/\partial V_{0c} = 0$, we have

$$A_{c3}V_{0c} = -A_{c2}^T V_c \tag{15}$$

We can obtain the optimal solution $\hat{V}_{0c}$ from Eq.(15). But the general inverse matrix does not exist when the data set includes missing values, because of depression of order of A_{c3}. Therefore we calculate $\hat{V}_{c0}$ from Eq.(15) with A_{c3}^+ which is Moore and Penrose generalized inverse of A_{c3}.

$$\hat{V}_{0c} = -A_{c3}^+ A_{c2}^T V_c \tag{16}$$

$\hat{V}_{0c}$ is a solution, because Eq.(16) satisfies Eq.(2). Accordingly let Eq.(11) be

$$\begin{aligned} g^{**}(V_c) &= \mathrm{tr}\{V_c^T(A_{c1} - A_{c2}A_{c3}^+A_{c2}^T)V_c\} \\ &= \mathrm{tr}(V_c^T A_c V_c) \end{aligned} \tag{17}$$

Where

$$A_c = A_{c1} - A_{c2}A_{c3}^+A_{c2}^T \tag{18}$$

Eq.(3) can be transformed into

$$V_c^T S_c V_c = V_c^T (S_c^{1/2})^T (S_c^{1/2}) V_c = I \tag{19}$$

Let $\tilde{V}_c = S_c^{1/2} V_c$ then Eqs.(3) and (17) are rewritten as follows:

$$\tilde{V}_c^T \tilde{V}_c = I \tag{20}$$

$$g^{**}(V_c) = \mathrm{tr}(\tilde{V}_c^T (S_c^{-1/2})^T A_c (S_c^{-1/2}) \tilde{V}_c) \tag{21}$$

The characteristic equation (Eq.(22)) is derived by minimizing g^{**} under the constraint of Eq.(3).

$$S_c^{-1/2} A_c S_c^{-1/2} \tilde{V}_c = \tilde{V}_c \Delta_c \tag{22}$$

Therefore we obtain the principal components from the eigenvectors corresponding to the least eigenvalue of Eq.(22).

3. SIMULTANEOUS APPROACH OF FUZZY CLUSTERING AND PRINCIPAL COMPONENT ANALYSYS

Miyamoto *et al.* [8] proposed an approach that can handle missing values in Fuzzy c-Means (FCM) clustering [7]. FCM is a fuzzy clustering method that partitions a data set into C spherical fuzzy clusters. The objective function of FCM for an incomplete data set with missing values is written as follows:

$$\psi = \sum_{c=1}^{C}\sum_{i=1}^{N} u_{ci} \sum_{j=1}^{n} d_{ij}(x_{ij} - v_{cj})^2 + \beta \sum_{c=1}^{C}\sum_{i=1}^{N} u_{ci} \log u_{ci} \tag{23}$$

where d_{ij} is defined by

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

and the entropy term is added for fuzzification [9]. β is a weighting parameter to specify degree of fuzziness of fuzzy clusters. This strategy is useful only for spherical clustering but can be enhanced by using another prototypical cluster centers.

We propose the simultaneous approach of PCA and FCM for an incomplete data set. Let the objective function of the proposed method be

$$\min\ L = \sum_{c=1}^{C}\sum_{i=1}^{N} u_{ci} \sum_{j=1}^{n} d_{ij}\Big\{\alpha \sum_{k=1}^{t}(y_{cijk} - w_{cik})^2 + (1-\alpha)(x_{ij} - v_{cj})^2\Big\}$$

$$+\beta\sum_{c=1}^{C}\sum_{i=1}^{N}u_{ci}\log\frac{u_{ci}}{\pi_c}+\sum_{i=1}^{N}\gamma_i(\sum_{c=1}^{C}u_{ci}-1)-\tau(\sum_{c=1}^{C}\pi_c-1) \quad (24)$$

We can obtain the principal components from the minimization of $(y_{cijk}-w_{cik})^2$ in the 1st term. $(x_{ij}-v_{cj})^2$ is the objective function for FCM. α is a constant which defines the tradeoff between FCM and PCA. When α is 1, Eq.(24) equals the objective function of PCA without clustering. When all of d_{ij} take 1, the algorithm results in a regular PCA for complete data set without missing values. On the other hand, when $\alpha = 0$ the algorithm is equivalent to the one in FCM. t is the number of principal components. We introduced K-L information term [10] in Eq.(24) instead of the entropy term in Eq.(23). γ_i and τ are the Lagrangean multiplier. From $\partial L/\partial u_{ci} = 0$ and $\partial L/\partial v_{cj} = 0$, we have

$$u_{ci} = \frac{\pi_c \exp(B_{ci})}{\sum_{a=1}^{C}\pi_c \exp(B_{ai})} \quad (25)$$

$$v_{cj} = \frac{\sum_{i=1}^{N}u_{ci}d_{ij}x_{ij}}{\sum_{i=1}^{N}u_{ci}d_{ij}} \quad (26)$$

where

$$B_{ai} = -\frac{1}{\beta}\sum_{j=1}^{n}d_{ij}\Big\{(\alpha\sum_{k=1}^{t}(y_{aijk}-w_{aik})^2+(1-\alpha)(x_{ij}-v_{aj})^2\Big\} \quad (27)$$

The algorithm is as follows:

Step1 Randomly choose membership u_{ck} $(c = 1, \cdots, C;\ k = 1, \cdots, n)$ from unit interval [0,1] and compute the cluster center vector $\boldsymbol{v}_c$ by Eq.(26)

Step2 Compute $\boldsymbol{y}_{cij}$ and $\boldsymbol{w}_{ci}$ by using eigenvalues of the eigenvalue problem (Eq.(22)).

Step3 Update $\boldsymbol{v}_c$ and u_{ck} by Eqs.(26) and (25).

Step4 If $\max_{c,i}|u_{ci}^{NEW}-u_{ci}^{OLD}| < \varepsilon$ then stop. Otherwise, go to Step2.

4. EXPERIMENTAL RESULTS

We implemented the novel technique presented in the previous section and tested with "kansei" data set which is a set of data of 8 different images used as background designs of stationery for word processor and psychologically evaluated by 285 users.

Each user evaluated the images on a scale from 1 to 7 based on the semantic differential (SD) method as shown in Fig.1.

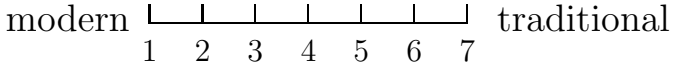

Figure 1: Seven-point Rating Scale

We compare the results of PCA, FCV and our proposed method.

Table 1: Comparison of Results

	PCA	FCV for Complete Data		FCV for Incomplete Data	
Design	——	1st Cluster	2nd Cluster	1st Cluster	2nd Cluster
1	0.265	0.365	-0.468	0.306	-0.478
2	0.501	-0.277	0.398	-0.703	0.516
3	0.170	0.421	-0.289	0.305	-0.308
4	0.301	0.125	-0.292	0.415	-0.266
5	0.405	-0.272	0.051	-0.032	0.108
6	0.451	-0.286	0.134	-0.259	0.222
7	0.096	0.512	-0.385	0.221	-0.316
8	0.461	-0.414	0.534	-0.174	0.423

The 2nd column in Table.1 shows the principal component vector with PCA. From this result, it is difficult to obtain clearly the features of background images of the stationary.

The principal component vectors with FCV are shown in the 3rd and 4th columns in Table.1. The tradeoff parameter α and the weighting parameter of fuzziness β were set to 0.5 and 0.3 respectively, and the users of the stationary were partitioned into two clusters. From the cluster centers, we can conclude that the 1st cluster consists of users who feel that the stationary is modern and the 2nd cluster consists of users who have traditional images. From the result of the 1st cluster, it is shown that design No.1,3,4 and 7 belong to a group which have similar images and No.2,5,6 and 8 are the other group. Particularly No.7 and No.8 have negative correlation. From the 2nd cluster, we can obtain the similar result.

The 5th and 6th columns in Table.1 show the results by our proposed method. The incomplete data consist of 6 items which is randomly removed 2 items from the complete data set. Therefore 25% of data have missing values in the incomplete data set. The results of clustering are similar to that of complete data set. Consequently we can obtain the approximately the same features with our proposed method even when the data include 25% missing values.

5. CONCLUSION

In this paper, we have proposed a new approach to extract local principal components from an incomplete data set. For extracting useful features, a set of data is partitioned into some linear clusters. Experimental results showed that our proposed method is efficient to obtain features from incomplete data set including missing values. Our future work is to determine the ability of the proposed method by using some other benchmarks.

REFERENCES

1. Wiberg, T.: Computation of Principal Components when Data are Missing. Proc. of the 2nd Symposium on computational Statistics (1976) 229–236

2. Shibayama, T.: A PCA-Like Method for Multivariate Data with Missing Values. Japanese Journal of Educational Psychology, Vol.40 (1992) 257–265 (in Japanese)

3. Shum, H., Ikeuchi, K., Reddy,R.: 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) 854–867

4. Yamakawa, A., Honda, K., Ichihashi, H., Miyoshi, T.: Simultaneous Approach to Fuzzy Cluster, Principal Component and Multiple Regression Analysis. Proc. of International Conference on Neural Networks (1999)

5. Oh, C.-H., Komatsu, H., Honda, K., Ichihashi, H.: Fuzzy Clustering Algorithm Extracting Principal Components Independent of Subsidiary Variables. Proc. of International Conference on Neural Networks (2000)

6. Bezdek, J. C., Coray, C., Gunderson, R., Watson, J.: Detection and Characterization of Cluster Substructure 2. Fuzzy c-Varieties and Convex Combinations Thereof. SIAM J. Appl. Math., Vol.40, No.2 (1981) 358–372

7. Bezdek, J. C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)

8. Miyamoto, S., Takata, O., Umayahara, K.: Handling Missing Values in Fuzzy c-Means. Proc. of the Third Asian Fuzzy Systems Symposium (1998) 139–142

9. Miyamoto, S., Mukaidono, M.:Fuzzy c-Means as a Regularization and Maximum Entropy Approach. Proc. of the 7th International Fuzzy Systems Association World Congress, Vol.2 (1997) 86–92

10. Ichihashi, H., Honda, K. and Tani, N.: Gaussian Mixture PDF Approximation and Fuzzy c-Means Clustering with Entropy Regularization, Proc. of the 4th Asian Fuzzy Systems Symposium, Vol.1 (2000) 217–221