Knowledge Discovery from Local Principal Components Independent of Arbitrary Factors *

Chi-Hyon Oh, Katsuhiro Honda, and Hidetomo Ichihashi
Graduate School of Engineering, Osaka Prefecture University,
Gakuen-cho 1-1, Sakai, Osaka, 599-8531 Japan

Abstract

In this paper, we propose a technique of extracting local principal components independent of arbitrary factors chosen. The proposed method takes advantage of Fuzzy c-Regression Models (FCRM) to estimate the parameters of regression models for fuzzy clusters. We decompose the fuzzy scatter matrix of each cluster into two matrices by using the partial regression coefficient matrix obtained by the FCRM. One is closely related to the arbitrary factors and the other is independent of them. Solving the eigen-value problem of the decomposed matrix enables us to extract the local principal components in which influences of arbitrary factors are neutralized. We apply our method to a POS transaction data set in order to discover useful knowledge from it.

1 Introduction

Principal component analysis (PCA) is a technique to compress multidimensional variables into a few indices. Through the observation of those indices, correlations among variables are extracted and we are able to discover some knowledge from the data set. Though we can gain general knowledge from the data set by the PCA, sometimes it could be trivial or valueless because the PCA extract the indices, *i.e.* principal components, so as to have them include the dominant feature of the data set. It might be necessary to analyze data sets from various points of view if one wants to accumulate several characteristics of them. Yanai [1] proposed a technique which has capability of neutralizing influences of arbitrary factors chosen. In [1], Yanai extracted principal components independent of the factors by exploiting the regression analysis technique.

Nonetheless, it is not sufficient to only get rid of influences of arbitrary factors considering lots of real data sets have some substructures. Respective examination of each substructure could lead to appropriate analyses of data sets. Fuzzy clustering is an effective vehicle to partition data sets into some substructures and has a lot of varieties. Fuzzy c-Regression Models (FCRM) [2] proposed by Hathaway *et al.* is one of those algorithms in which parameters of regression models for clusters, *i.e.* partial regression coefficients, are estimated.

In this paper, we propose a technique of extracting local principal components independent of arbitrary factors. Firstly, we implement the FCRM to derive the parameters of c-regression models in our method. The fuzzy scatter matrix is decomposed into two matrices then in the same manner as Yanai's approach. We can obtain local principal components independent of arbitrary factors by solving eigen-value problem of the decomposed fuzzy scatter matrix. We apply our method to a POS transaction data set in which each data point is composed of 20 attributes such as meteorological element and sales of two different supermarkets to gain diverse knowledge from it.

2 Extraction of Local Principal Components Independent of Arbitrary Factors

Yanai's approach [1] to extract principal components independent of arbitrary factors takes advantage of the regression analysis technique. It decomposes the variance covariance matrix derived in principal component analysis into two parts by using the obtained partial regression coefficient. One is closely related to external

*Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies, ed. by N. Baba, L. C. Jain and R. J. Howlett, IOS Press, 17-21 (2001)

criteria and the other is independent of them. In Yanai's approach, the factors one intends to neutralize are regarded as the external criteria. Taking substructures of data sets into consideration, the partial regression coefficient for each of them is estimated in our method. We employ Hathaway's FCRM [2] for the purpose.

In Yanai's approach, it is assumed that the mean value of the data set is zero. We, therefore, introduce the concept of cluster center to the original FCRM. Let Y_c and X_c be sets of response variables, *i.e.* external criteria, and explanatory variables for each cluster c. They can be written as follows:

$$Y_c = (\boldsymbol{y}_{c1}^T, \boldsymbol{y}_{c2}^T, \cdots, \boldsymbol{y}_{cn}^T)^T = \{y_{cki}\}, \qquad y_{cki} = y_{ki} - v_{ci}^y, \quad i = 1, \cdots, t, \tag{1}$$

$$X_c = (\boldsymbol{x}_{c1}^T, \boldsymbol{x}_{c2}^T, \cdots, \boldsymbol{x}_{cn}^T)^T = \{x_{ckj}\}, \qquad x_{ckj} = x_{kj} - v_{cj}^x, \quad j = 1, \cdots, s, \tag{2}$$

where T denotes transposition. t and s are the dimensionalities of response and explanatory variables and n is the number of data points. Note that $\boldsymbol{y}_{ck}^T$ and $\boldsymbol{x}_{ck}^T$ are represented row vectors for convenience sake of notation of the following equations. v_{ci}^y and v_{cj}^x are cluster centers of response and explanatory variables.

Suppose we identify the following C linear models to extract local linear structures.

$$\boldsymbol{y}_c = \boldsymbol{x}_c B_c + \boldsymbol{e}_c, \quad c = 1, \cdots, C, \tag{3}$$

where B_c is the partial regression coefficient matrix for each cluster c and can be written as follows:

$$B_c = \begin{pmatrix} b_{c11} & b_{c12} & \dots & b_{c1t} \\ b_{c21} & b_{c22} & \dots & b_{c2t} \\ \vdots & \vdots & \ddots & \vdots \\ b_{cs1} & b_{cs2} & \dots & b_{cst} \end{pmatrix}$$

B_c is calculated by least squares method in the general linear regression analysis.

The FCRM can be driven by optimization of an objective function. We use Lagrange's method of indeterminate multiplier to derive the objective function for the FCRM. We have the following objective function to be minimized.

$$L = \sum_{c=1}^{C}\sum_{k=1}^{n} u_{ck}\|\boldsymbol{y}_{ck} - \boldsymbol{x}_{ck}B_c\|^2 + T_0\sum_{c=1}^{C}\sum_{k=1}^{n} u_{ck}\log u_{ck} + \sum_{k=1}^{n} T_k\Big(\sum_{c=1}^{C} u_{ck} - 1\Big). \tag{4}$$

u_{ck} is membership of the data point k to the cluster c which represents degree of belonging to clusters. u_{ck} satisfies the following condition.

$$\boldsymbol{u}_c \in \{(u_{ck}) | \sum_{c=1}^{C} u_{ck} = 1, u_{ck} \in [0,1], c = 1, \cdots, C\}. \tag{5}$$

In (4), the first term represents residual sum of squares. The second term represents entropy maximization as regularization which was introduced in Fuzzy c-Means by Miyamoto *et al.* [3] for the first time. It enables us to obtain fuzzy clusters. T_0 is the weighting parameter which specifies degree of fuzziness. The remaining terms describe the constraint of the membership u_{ck}, *i.e.* (5). T_k is the Lagrangian multiplier.

From the necessary condition $\partial L/\partial b_{cji} = 0$ for the optimality of the objective function L, B_c can be derived as follows:

$$B_c = (X_c^T D_c X_c)^{-1} X_c^T D_c Y_c. \tag{6}$$

In (6), D_c is the diagonal matrix, diagonal entries of which are memberships of data points.

$$D_c = \begin{pmatrix} u_{c1} & 0 & \dots & 0 \\ 0 & u_{c2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & u_{cn} \end{pmatrix}$$

From the necessary conditions $\partial L/\partial u_{ck} = 0$, $\partial L/\partial v_{cj}^x = 0$ and $\partial L/\partial v_{ci}^y = 0$ for L, we also have the following equations for the membership u_{ck}, the cluster center of the explanatory variable v_{cj}^x and the cluster center of the response variable v_{cj}^y.

$$u_{ck} = \exp(A_{ck})/\sum_{a=1}^{c}\exp(A_{ak}), \quad A_{ak} = -\frac{1}{T_0}\sum_{i=1}^{t}\Big(y_{aki} - \sum_{j=1}^{s} x_{akj}b_{aij}\Big)^2, \tag{7}$$

$$v_{cj}^{x} = \sum_{k=1}^{n} u_{ck} x_{kj} / \sum_{k=1}^{n} u_{ck}, \tag{8}$$

$$v_{cj}^{y} = \sum_{k=1}^{n} u_{ck} y_{kj} / \sum_{k=1}^{n} u_{ck}. \tag{9}$$

The solution algorithms for the FCRM are based on an iterative procedure. We remark them below.

FCRM Algorithms
Step 1: Set the number of clusters C, the coefficient of the entropy term T_0 and the terminal condition ϵ. Initialize the membership $u_c k$ randomly so as to have it follow the condition in (12).
Step 2: Update the cluster center v_{cj}^{x} and v_{cj}^{y} using (10) and (11).
Step 3: Calculate the partial regression coefficient matrix B_c using (7).
Step 4: Update the membership u_{ck} using (9).
Step 5: If $\max |u_{ck}^{NEW} - u_{ck}^{OLD}| < \epsilon$, then stop. Otherwise, return to Step 2.

We can represent the set of response variables Y_c as in (10) by using the partial regression coefficient matrix B_c.

$$\begin{aligned} Y_c &= X_c B_c + \{Y_c - X_c B_c\} \\ &= Y_{Xc} + Y_{Xc}^{-}. \end{aligned} \tag{10}$$

In (10), Y_{Xc} is the predicted value by X_c. On the one hand, Y_{Xc}^{-} is the unaccountable part according to X_c. It is, therefore, able to be said that Y_{Xc}^{-} is independent of X_c. We can divide fuzzy scatter matrix S_{fc} into two parts in the same manner as (10).

$$\begin{aligned} S_{fc} &= Y_c^T D_c Y_c \\ &= Y_c^T D_c (Y_{X_c} + Y_{X_c}^{-}) \\ &= Y_c^T D_c X_c B_c + \{Y_c^T D_c Y_c - Y_c^T D_c X_c B_c\} \\ &= S_{fc}^{X} + S_{fc}^{X-}, \end{aligned} \tag{11}$$

wehre S_{fc}^{X} is the factor matrix accountable by X_c and S_{fc}^{X-} is independent of X_c. Using (6), we can denote S_{fc}^{X} and S_{fc}^{X-} as follows:

$$S_{fc}^{X} = Y_c^T D_c X_c (X_c^T D_c X_c)^{-1} X_c^T D_c Y_c, \tag{12}$$

$$S_{fc}^{X-} = Y_c^T D_c Y_c - Y_c^T D_c X_c (X_c^T D_c X_c)^{-1} X_c^T D_c Y_c. \tag{13}$$

These factor matrices are both symmetric. We can extract principal components independent of explanatory variables by solving the eigen-value problem of the factor matrix S_{fc}^{X-}. We can neutralize the influence of arbitrary factors by setting them as explanatory variables.

3 Knowledge Discovery from a POS Transaction Data Set

We applied the proposed method to a POS (Point of Sales) transaction data set of two different supermarkets to discover useful knowledge from it. The data set consists of 333 data, which have 20 items, the number of customers having come to the store, meteorological element and so on. We enumerate each item and item number below.

Items of the POS transaction data set

> 1: Holiday, 2: Friday, 3: Saturday, 4: Sunday, 5: Average temperature of the day, 6, 7, 8 and 9: Temperature at 6, 12, 15 and 18 o'clock, 10: Humidity, 11 and 12: Weather category during day and night, 13: Precipitation, 14, 15 and 16: Precipitation during 9-12, 12-15 and 15-18 o'clock, 17 and 18: the number of customers of supermarket A and B, 19 and 20: Sales of perishables of supermarket A and B

Table 1: Fuzzy Factor Loading.

Item #	1	2	3	4	5	6	7	8	9	10
Cold season	0.01	-0.48	0.35	0.39	-0.12	-0.12	-0.09	-0.06	-0.10	-0.28
Hot season	-0.01	-0.47	0.30	0.53	-0.46	-0.45	-0.43	-0.46	-0.46	-0.11

Item #	11	12	14	15	16	17	18	19	20
Cold season	-0.33	-0.32	-0.30	-0.22	-0.19	0.78	0.87	0.92	0.88
Hot season	-0.02	0.11	-0.14	-0.15	-0.02	0.75	0.86	0.90	0.92

The items of days of week, Holyday, Friday, Saturday, Sunday, are dummy variables and weather categories are represented by integer values.

We applied Fuzzy c-Varieties (FCV) [4], which is one of the fuzzy clustering algorithm proposed by Bezdek *et al.*, to the POS data set before applying our method to it. In the FCV, prototypes of clusters are multi dimensional linear varieties. Since the linear varieties are spanned by some local principal component vectors, the FCV can be regarded as a simultaneous algorithm of fuzzy clustering and the PCA. Through the analysis of the obtained local principal components, we can discover some knowledge of each cluster. We set the number of clusters to two. The data set was divided into two seasonal clusters, the hot and cold seasons. After analyzing the local principal components obtained by the FCV, we gain the knowledge which says precipitation and the sales of perishables have a close correlation. The less it rains, the more the sales increase.

We applied our method to the POS transaction data set on the basis of the results of the FCV. Since we have obtained the relation with precipitation and the sales of perishables, we chose precipitation as the factor to be neutralized its influence. We show the value of fuzzy factor loading [5] of each item in Table 1. The fuzzy factor loading quantifies correlations between local principal components and each item. When a certain item has the same sign of the fuzzy factor loading as the other items, it has positive correlation with them. Therefore, if the value of an item increases, that of the other item which has the same sign of the fuzzy factor loading also increases and vice versa. We also specified the number of clusters as two in our method. The data set was divided into the hot and cold seasons. In Table 1, we underlined noteworthy values. Observing the values of fuzzy factor loading, we can mention that the numbers of customers of both supermarkets increase on Saturday and Sunday not Friday and the sales grow in both seasons since Friday has the value of fuzzy factor loading which has the opposite sign of the sales of perishables. Furthermore, when the temperature comes down, the sales go up. In this manner, we can gain diverse knowledge from the data set by using our proposed method.

4 Conclusion

In this paper, we proposed a technique of extracting local principal components independent of arbitrary factors. Firstly, we estimate the partial regression coefficient of each cluster by using the FCRM and decompose it into two matrices then. We can extract the local principal components in which influences of arbitrary factors are neutralized from the decomposed matrix. In the numerical example, we applied the proposed method to a POS transaction data set. Our method has the capability of discovering diverse knowledge from the data set.

References

[1] H. Yanai, Factor Analysis with External Criteria, Japanese Psychological Research 12, 4 (1970) 143-153.

[2] R. J. Hathaway and J. C. Bezdek, Switching Regression Models and Fuzzy Clustering, IEEE Trans. on Fuzzy Systems 1, 3 (1993) 195-204.

[3] S. Miyamoto and M. Mukaidono, Fuzzy c-Means as a Regularization and Maximum Entropy Approach, Proc. IFSA'97 2 (1997) 86-92.

[4] 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. 40, 2 (1981) 358-372.

[5] Y.Yabuuchi and J.Watada, Fuzzy Principal Compoment Analysis and Its Application, Biomedical Fuzzy and Human Sciences 3, 1 (1997) 83-92.