Fuzzy Clustering Based Correlation and Its Application to Principal Component Analysis

Mika Sato-Ilic
University of Tsukuba, Tsukuba, Ibaraki, JAPAN
e-mail: mika@risk.tsukuba.ac.jp

Abstract

We propose a new correlation based on a fuzzy clustering result and a new principal component analysis (PCA) based on eigenvalues of this correlation. The merit of the use of fuzzy clustering based correlation to the PCA is that we can obtain eigenvalues of covariance matrix of variables for high-dimension and low-sample size (HDLSS) data in which the number of variables (dimensions) is much larger than the number of objects. Therefore, ordinary PCA can apply to HDLSS data by the use of the fuzzy clustering based correlation. In addition, we show a numerical example using a microarray data which is a typical HDLSS data to show a better performance of the proposed PCA with the fuzzy clustering based correlation when compared with the ordinary PCA.

Keywords: Fuzzy cluster, high-dimension and low-sample size (HDLSS) data, self-organized dissimilarity

1. Introduction

Recently the analysis of HDLSS data has gained tremendous interest. HDLSS data have been obtained in various areas, such as genomics, bioinformatics, chemometrics, brain sciences, and functional data analysis. However, it is well known that the analysis of HDLSS data has difficulty since this type of data has irrelevant and redundant variables (dimensions) related with the curse of dimensionality. Therefore, it is important to reduce the number of dimensions in this type of data. PCA is a well known method to reduce the number of variables (dimensions) and to obtain the latent structure of data as the similarity of objects in lower dimensional space spanned by the obtained principal components.(Jolliffe (2002)) However, we cannot use ordinary PCA based on eigenvalues of the covariance matrix of variables for HDLSS data, since we cannot obtain correct solutions as the eigenvalues of the covariance matrix of variables when dealing with HDLSS data. (Ahn et al. (2007), Baik et al. (2005), Hall et al. (2005), Johnstone (2001)) The new correlation proposed in this paper can overcome this problem by measuring the similarity of two kinds of correlations, the correlation of variables and the correlation of classification structures of the variables, and induce the fuzzy clustering based dissimilarity of objects including the feature of similarity of classification structures of objects. This paper consists of the following: Section 2 describes fuzzy clustering and fuzzy clustering based dissimilarity. Based on the idea of the fuzzy clustering based dissimilarity, section 3 proposes a fuzzy clustering based correlation of variables and section 4 presents a new PCA exploiting the fuzzy clustering based correlation of variables. Section 5 describes a numerical example and section 6 contains conclusions.

2. Fuzzy Clustering and Fuzzy Clustering based Dissimilarity

We observe data

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

where x_{ia} shows data of an object i with respect to a variable a. n shows number of objects and p is number of variables. The state of fuzzy clustering is represented by a partition matrix $U = (u_{ik})$ whose elements show the degree of belongingness of the objects to the clusters, $u_{ik},\ i = 1, \cdots, n,\ k = 1, \cdots, K$, where K is the number of clusters. In general, u_{ik} satisfies the following conditions:

$$u_{ik} \in [0,1], \forall i, k; \ \sum_{k=1}^{K} u_{ik} = 1, \forall i. \tag{2}$$

Fuzzy c-means (FCM) (Bezdek (1981)) is one of the methods of fuzzy clustering. FCM is the method which minimizes the weighted within-class sum of squares:

$$J(U, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_K) = \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^m d^2(\boldsymbol{x}_i, \boldsymbol{v}_k), \tag{3}$$

where $\boldsymbol{v}_k = (v_{ka}),\ k = 1, \cdots, K,\ a = 1, \cdots, p$ denotes the center of a cluster k, $\boldsymbol{x}_i = (x_{ia}),\ i = 1, \cdots, n,\ a = 1, \cdots, p$ is i-th object, and $d^2(\boldsymbol{x}_i, \boldsymbol{v}_k)$ is the square Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{v}_k$. The exponent m which determines the degree of fuzziness of the clustering is chosen from $[1, \infty)$ in advance. The purpose is to obtain the solutions U and $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K$ which minimize equation (3). From conditions shown in equation (2), the local extrema of equation (3) can be obtained as follows:

$$u_{ik} = 1 / \sum_{l=1}^{K} \{d(\boldsymbol{x}_i, \boldsymbol{v}_k) / d(\boldsymbol{x}_i, \boldsymbol{v}_l)\}^{\frac{2}{m-1}}, \ \ \boldsymbol{v}_k = \sum_{i=1}^{n} (u_{ik})^m \boldsymbol{x}_i / \sum_{i=1}^{n} (u_{ik})^m, \ \forall i, k. \tag{4}$$

If we assume $\boldsymbol{v}_k$ in equation (4), then the minimizer of equation (3) is shown as:

$$J(U) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} u_{ik}^m u_{jk}^m d_{ij} / (2 \sum_{l=1}^{n} u_{lk}^m) \right), \tag{5}$$

where $d_{ij} = d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)$. When $m = 2$, equation (5) is the objective function of the FANNY algorithm (Kaufman and Rousseeuw (1990)) for any dissimilarity d_{ij}. We define dissimilarity of objects i and j can be explained by observed dissimilarity of the objects and latent dissimilarity of classification structures of objects i and j. That is,

$$\tilde{d}_{ij} = g(f_{ij}, d_{ij}),\ i, j = 1, \cdots, n, \tag{6}$$

where d_{ij} shows the observed dissimilarity of objects i and j, and f_{ij} shows dissimilarity between classification structures of objects i and j, and $\tilde{d}_{ij}$ shows the real dissimilarity of objects i and j. g shows any real function satisfying $[0, c_1] \times [0, c_2] \to [0, c_3]$, where $c_1, c_2, c_3 \in (0, \infty)$ are constants. Dissimilarities satisfy following conditions:

$$d_{ij} \geq d_{ii} \geq 0, \ \ \tilde{d}_{ij} \geq \tilde{d}_{ii} \geq 0, \ \ f_{ij} \geq f_{ii} \geq 0, \ \ \forall i, j. \tag{7}$$

$$d_{ij} = d_{ji}, \ \ \tilde{d}_{ij} = \tilde{d}_{ji}, \ \ f_{ij} = f_{ji}, \ \ \forall i, j, \ \ i \neq j. \tag{8}$$

As a special case of equation (6), we define $\tilde{d}_{ij}$ as follows:

$$\tilde{d}_{ij} = f_{ij} d_{ij},\ i, j = 1, \cdots, n. \tag{9}$$

Equation (9) shows that $\tilde{d}_{ij}$ is defined by observed dissimilarity d_{ij} effected by f_{ij}. That is, from conditions (7) and (8), and the fact that observed dissimilarity d_{ij} is given from data, if dissimilarity f_{ij} is larger, then $\tilde{d}_{ij}$ is also larger. This means that if classification structures of objects i and j is not similar to each other, then dissimilarity of objects i and j is also not similar to each other. Using idea of fuzzy clustering, we define the fuzzy classification structure of an object i as follows: $\boldsymbol{u}_i = (u_{i1}, \cdots, u_{iK}), \quad i = 1, \cdots, n$, where u_{ik} satisfies conditions shown in equation (1). Then, we define f_{ij} as follows: $f_{ij} = \sum_{k=1}^{K}(u_{ik} - u_{jk})^2, \quad i, j = 1, \cdots, n$, and d_{ij} is assumed to be observed squared Euclidean distance which is $d_{ij} = \sum_{a=1}^{p}(x_{ia} - x_{ja})^2, \quad i, j = 1, \cdots, n$. Then equation (9) can be written as follows:

$$\tilde{d}_{ij} = \sum_{k=1}^{K}(u_{ik} - u_{jk})^2 \sum_{a=1}^{p}(x_{ia} - x_{ja})^2, \tag{10}$$

In equation (10), we can see that if objects i and j have similar degree of belongingness for the obtained clusters, then the dissimilarity between objects i and j becomes smaller. In this equation, if the observed data is dissimilarity, then we can rewrite equation (10) as follows: $\tilde{d}_{ij} = \sum_{k=1}^{K}(u_{ik} - u_{jk})^2 \hat{d}_{ij}$, where $\hat{d}_{ij}$ is an observed dissimilarity between objects i and j. Here, we assume that the observed dissimilarity $\hat{d}_{ij}$ can be explained by the estimate x_{ia} as the result of multidimensional scaling. (Gower (1966), Kruskal and Wish (1978)) Multidimensional Scaling (MDS) is a method for capturing efficient information from observed dissimilarity data by representing the data structure in lower dimensional spatial space. As a metric MDS (principal coordinate analysis), the model $d_{ij} \approx \{\sum_{a=1}^{p} d^{\kappa}(x_{ia}, x_{ja})\}^{\frac{1}{\kappa}}$ has been proposed. In this model, d_{ij} is an observed dissimilarity between objects i and j and x_{ia} is a point of an object i with respect to dimension a in p dimensional configuration space. $d^{\kappa}(x_{ia}, x_{ja})$ shows dissimilarity between objects i and j with respect to dimension a and usually $d^{\kappa}(x_{ia}, x_{ja}) = \mid x_{ia} - x_{ja} \mid^{\kappa}$. When $\kappa = 2$, this model can be rewritten as follows: $d_{ij}^2 = \sum_{a=1}^{p}(x_{ia} - x_{ja})^2 + \varepsilon_{ij}$. ε_{ij} is an error between the model and the data. Without loss of generality, we put $\hat{d}_{ij} \equiv d_{ij}^2$, then equation (10) can be rewritten as follows:

$$\tilde{d}_{ij} = \sum_{k=1}^{K}(u_{ik} - u_{jk})^2 (\sum_{a=1}^{p}(x_{ia} - x_{ja})^2 + \varepsilon_{ij}) = \sum_{k=1}^{K}(u_{ik} - u_{jk})^2 \sum_{a=1}^{p}(x_{ia} - x_{ja})^2 + \sum_{k=1}^{K}(u_{ik} - u_{jk})^2 \varepsilon_{ij}.$$

In this equation, the first term means that if the dissimilarity between classification structures of objects i and j is smaller, then the revised dissimilarity between objects i and j becomes smaller. In addition, the second term means if the absolute value of error (noise) of dissimilarity between objects i and j is larger, then subsequently the effect of the dissimilarity between classification structures of objects i and j to the noise is larger. That is, this revised dissimilarity can measure not only dissimilarity of objects but also dissimilarity of classification structures of the objects. In addition, this dissimilarity induces a learning process of the noise (error) of dissimilarity data. Therefore, we call this dissimilarity fuzzy self-organized dissimilarity. (Sato-Ilic (2004), Sato-Ilic and Kuwata (2005))

3. Fuzzy Clustering based Correlation of Variables

Suppose

$$s_{ij} \equiv \sum_{k=1}^{K} u_{ik} u_{jk}, \quad \hat{s}_{ij} \equiv \sum_{a=1}^{p} x_{ia} x_{ja}. \tag{11}$$

Then equation (10) is rewritten as:

$$\tilde{d}_{ij} = (s_{ii} + s_{jj})(\hat{s}_{ii} + \hat{s}_{jj}) + 4s_{ij}\hat{s}_{ij} - 2\{(s_{ii} + s_{jj})\hat{s}_{ij} + (\hat{s}_{ii} + \hat{s}_{jj})s_{ij}\}. \tag{12}$$

We normalize u_{ik} and x_{ia} for each object as:

$$u^*_{ik} \equiv \frac{u_{ik} - \frac{1}{K}}{\sigma_i^{(u)}},\ \ x^*_{ia} \equiv \frac{x_{ia} - \bar{x}_i}{\sigma_i^{(x)}},\ \ \sigma_i^{(u)} = \sqrt{\frac{\sum_{k=1}^{K}(u_{ik} - \frac{1}{K})^2}{K-1}},\ \ \sigma_i^{(x)} = \sqrt{\frac{\sum_{a=1}^{p}(x_{ia} - \bar{x}_i)^2}{p-1}},\ \ \bar{x}_i = \frac{\sum_{a=1}^{p} x_{ia}}{p}. \tag{13}$$

Then equation (10) should be changed as follows:

$$\tilde{d}^*_{ij} = \frac{1}{K-1}\sum_{k=1}^{K}(u^*_{ik} - u^*_{jk})^2 \frac{1}{p-1}\sum_{a=1}^{p}(x^*_{ia} - x^*_{ja})^2 = 4\{1 + s^*_{ij}\hat{s}^*_{ij} - (s^*_{ij} + \hat{s}^*_{ij})\}, \tag{14}$$

where

$$s^*_{ij} \equiv \frac{1}{K-1}\sum_{k=1}^{K} u^*_{ik}u^*_{jk},\ \ \hat{s}^*_{ij} \equiv \frac{1}{p-1}\sum_{a=1}^{p} x^*_{ia}x^*_{ja}. \tag{15}$$

From equation (13), s^*_{ij} and $\hat{s}^*_{ij}$ shown in equation (15) are correlation of degree of belongingness of objects i and j over the K clusters and correlation of objects i and j over p variables, respectively. From equation (14), we obtain the following fuzzy clustering based correlation as:

$$c_{ij} \equiv s^*_{ij} + \hat{s}^*_{ij} = s^*_{ij}\hat{s}^*_{ij} - \frac{\tilde{d}^*_{ij}}{4} + 1. \tag{16}$$

In equation (16), we can see that the proposed correlation consists of two kinds of correlations, $\hat{s}^*_{ij}$ and s^*_{ij}, which are correlation of objects and correlation of classification structures of the objects. In addition, from equation (16), we can see that the fuzzy clustering based correlation measures similarity of the two kinds of correlations which is represented by $s^*_{ij}\hat{s}^*_{ij}$ and the fuzzy clustering based dissimilarity between the objects which is denoted as $\tilde{d}^*_{ij}$. That is, the proposed correlation includes features of the fuzzy clustering based dissimilarity.

4. Principal Component Analysis used Fuzzy Clustering based Correlation

Suppose we obtain a data as a high-dimension low-sample size data. That is, in this data the number of variables (dimensions) is very much larger than the number of objects. We denote this situation as $p >> n$ in equation (1) and rewrite the data matrix of variables with respect to objects as follows:

$$\tilde{X} = X^t = (\tilde{x}_{ai}),\ \ a = 1, \cdots, p,\ i = 1, \cdots, n,\ \ p >> n. \tag{17}$$

For the data $X(p >> n)$, we cannot apply directly the principal component analysis (PCA), due to the inconsistency of eigenvalues of the sample covariance matrix with respect to variables, although one of the purposes of this analysis is the reduction of the number of dimensions (variables). Therefore, for $\tilde{X}$ shown in equation (17), we calculate the following fuzzy clustering based correlation with respect to variables:

$$\tilde{c}_{ab} \equiv \tilde{s}^*_{ab} + \tilde{\hat{s}}^*_{ab} = \tilde{s}^*_{ab}\tilde{\hat{s}}^*_{ab} - \frac{\tilde{\tilde{d}}^*_{ab}}{4} + 1, \tag{18}$$

where $\tilde{s}^*_{ab}$ shows correlation of degree of belongingness of variables a and b over the K clusters and $\tilde{\tilde{s}}^*_{ab}$ shows correlation of variables a and b over n objects. That is,

$$\tilde{s}^*_{ab} \equiv \frac{1}{K-1}\sum_{k=1}^{K}\tilde{u}^*_{ak}\tilde{u}^*_{bk}, \quad \tilde{\tilde{s}}^*_{ab} \equiv \frac{1}{n-1}\sum_{i=1}^{n}\tilde{x}^*_{ai}\tilde{x}^*_{bi},$$

$$\tilde{u}^*_{ak} \equiv \frac{\tilde{u}_{ak}-\frac{1}{K}}{\tilde{\sigma}^{(u)}_a}, \ \tilde{x}^*_{ai} \equiv \frac{\tilde{x}_{ai}-\bar{\tilde{x}}_a}{\tilde{\sigma}^{(x)}_a}, \ \tilde{\sigma}^{(u)}_a = \sqrt{\frac{\sum_{k=1}^{K}(\tilde{u}_{ak}-\frac{1}{K})^2}{K-1}}, \ \tilde{\sigma}^{(x)}_a = \sqrt{\frac{\sum_{i=1}^{n}(\tilde{x}_{ai}-\bar{\tilde{x}}_a)^2}{n-1}}, \ \bar{\tilde{x}}_a = \frac{\sum_{i=1}^{n}\tilde{x}_{ai}}{n}.$$

Equation (18) is basically the same as equation (16) and the only difference is whether it is the correlation of variables or the correlation of objects. Note that ordinary PCA uses the correlation matrix of variables $\tilde{\tilde{S}} = (\tilde{\tilde{s}}^*_{ab})$ to obtain the principal components when the data is normalized for each variable. However, due to the inconsistency of the eigenvalues of the sample covariance matrix in the case of the high-dimension low-sample size data, we cannot obtain any correct result of PCA by using only $\tilde{\tilde{S}}$. However, by adding the $\tilde{s}^*_{ab}$ which is the correlation of degree of belongingness of variables a and b over the K clusters to the ordinary correlation $\tilde{\tilde{s}}^*_{ab}$, we can use the correlation with respect to variables and obtain the result of the PCA. Then, the α-th principal component $\boldsymbol{z}_\alpha$ of $\tilde{X}$ ($p >> n$) is defined as follows:

$$\boldsymbol{z}_\alpha = \tilde{X}^t\boldsymbol{l}_\alpha = X\boldsymbol{l}_\alpha, \tag{19}$$

where $\boldsymbol{l}^t_\alpha = (l_{\alpha 1}, l_{\alpha 2}, \cdots, l_{\alpha p})$, and $\boldsymbol{l}_\alpha$ satisfies the conditions $\boldsymbol{l}^t_\alpha\boldsymbol{l}_\alpha = 1$, $\boldsymbol{l}^t_\alpha\boldsymbol{l}_\beta = 0$, $\alpha \neq \beta$. $\boldsymbol{l}_\alpha$ is obtained as the corresponding eigenvector for the α-th largest eigenvalue of $\tilde{C} = (\tilde{c}_{ab})$ shown in equation (18).

5. Numerical Example

We use gene expression data for prostate cancer. (Welsh et al. (2001)) The data consists of 32 objects (subjects) with respect to 12626 variables (genes) shown in equation (17). 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 without cancer. Figure 1 shows the result of the proposed PCA shown in equation (19). We use the result of the fuzzy clustering of variables when we assume the number of clusters is 2. This is a typical high-dimension low-sample size data and we cannot apply ordinary PCA to reduce the number of dimensions from the given dimension of 12626. Therefore, we calculate the proposed correlation shown in equation (18). In order to obtain the fuzzy clustering result, we use the FANNY method shown in equation (5). From the result shown in figure 1, we can see that the objects are successfully classified into the two given groups. That is, in this figure, 24 - 32 show objects without cancer and the other numbers show objects with cancer. This means that our proposed PCA could reduce the 12626 dimensions to three dimensions and also this method can obtain the adaptable result of objects for the externally given information of the classification structure of data. As a comparison, figure 2 shows the result of the ordinary PCA, from which we cannot see any clear classification. In addition, the cumulative proportion of the three components of the proposed PCA is 0.63 and the cumulative proportion of the three components of the ordinary PCA is 0.32. From these comparisons, we can see that the proposed PCA has a better performance when compared with the ordinary PCA.

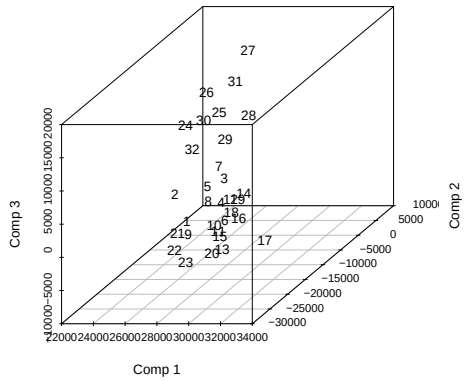

Fig. 1 Result of the Proposed Correlation based PCA

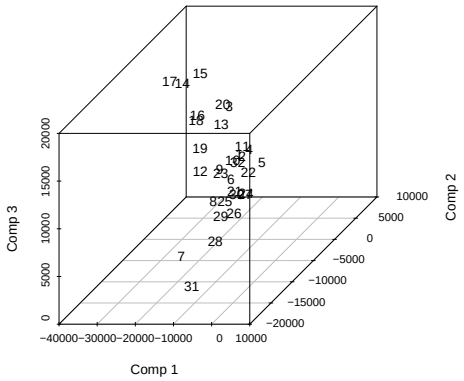

Fig. 2 Result of the Ordinary PCA

6. Conclusion

This paper presents a fuzzy clustering based correlation and its use for principal component analysis for high-dimension and low-sample size (HDLSS) data. Numerical example shows a better performance for the proposed PCA.

References

Ahn, J., Marron, J.S., Muller, K.M. and Chi, Y-Y. (2007) "The High-Dimension, Low-Sample-Size Geometric Representation Holds under Mild Conditions," *Biometrika*, 94, 3, 760-766.

Baik, J., Arous, G.B. and Peche, S. (2005) "Transition of the Largest Eigenvalue for Nonnull Complex Sample Covariance Matrices," *The Annals of Probability*, 33, 5, 1643-1697.

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

Gower, J.C. (1966) "Some Distance Properties of Latent Roots and Vector Methods used in Multivariate Analysis," *Biometrika*, 53, 325-338.

Hall, P., Marron, J.S. and Neeman, A. (2005) "Geometric Representation of High Dimension Low Sample Size Data," *Journal of Royal Statistical Society*, 67, Part 3, 427-444.

Johnstone, I.M. (2001) "On the Distribution of the Largest Eigenvalue in Principal Components Analysis," *The Annals of Statistics*, 29, 2, 295-327.

Jolliffe, I.T. (2002) *Principal Component Analysis*, 2nd ed., Springer.

Kaufman L., Rousseeuw, P.J. (1990) *Finding Groups in Data*, John Wiley & Sons.

Kruskal, J.B., Wish, M. (1978) *Multidimensional Scaling*, Sage Publications.

Sato-Ilic, M. (2004), "Self-Organized Fuzzy Clustering," *Intelligent Engineering Systems through Artificial Neural Networks*, 14, 579-584.

Sato-Ilic, M., Kuwata, T. (2005) "On Fuzzy Clustering based Self-Organized Methods," *FUZZ-IEEE2005*, 973-978.

Sato-Ilic, M. (2012) "Structural Classification based Correlation and its Application to Principal Component Analysis for High-Dimension Low-Sample Size Data," *2012 IEEE World Congress on Computational Intelligence*, 981-988.

Sato-Ilic, M. (2012) "On Fuzzy Clustering Based Correlation," *Procedia Computer Sciences, Elsevier*, 12, 230-235.

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