Fuzzy Clustering Based Correlation and Its Application to Principal Component Analysis

Mika Sato-Ilic
University of Tsukuba, Tsukuba, Japan mika@risk.tsukuba.ac.jp

Recently the analysis of high-dimension and low-sample size (HDLSS) data in which the number of variables (dimensions) is much larger than the number of objects 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. Principal component analysis (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. 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. In order to solve this problem, we propose a new correlation based on a fuzzy clustering result and a new PCA based on eigenvalues of this correlation. This correlation is derived from the self-organized dissimilarity of objects which can measure the dissimilarity considered classification structure of objects and learn the noise of the dissimilarity. Using this property, we can theoretically show that the proposed correlation measures similarity between two kinds of correlations which are the correlation of variables and the correlation of classification structures of the variables, and the self-organized dissimilarity. 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.

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