Effective PCA for High-Dimensional Data and Its Applications

Makoto Aoshima*
University of Tsukuba, Ibaraki, Japan aoshima@math.tsukuba.ac.jp

Kazuyoshi Yata
University of Tsukuba, Ibaraki, Japan yata@math.tsukuba.ac.jp

High-dimensional data situations occur in many areas of modern science. A common feature of high-dimensional data is that the data dimension is high, however, the sample size is relatively low. In this talk, we introduce a general spiked model called the power spiked model in high-dimensional settings. We derive relations among the dimension, the sample size and the high-dimensional noise structure. We first consider asymptotic properties of the conventional estimator of eigenvalues. We show that the estimator is affected by the high-dimensional noise structure directly, so that it becomes inconsistent. In order to overcome such difficulties in a high-dimensional situation, we focus on geometric representations of high-dimension, low-sample-size data. We develop new principal component analysis (PCA) methods called the noise-reduction methodology and the cross-data-matrix methodology. We show that the new PCA methods can enjoy consistency properties not only for eigenvalues but also for PC directions and PC scores in high-dimensional settings. Finally, we apply the PCA methods to cluster analysis. We show that the clustering method given by using the PC scores of the PCA methods can classify individuals into several groups effectively. We demonstrate how the methods perform well by using microarray data sets.

Key Words**:** Cluster analysis, cross-data-matrix methodology, microarray data, noise-reduction methodology