Abstract:

In recent years, substantial work has been done on high-dimension, low-sample-size (HDLSS) asymptotic theory by letting $p \to \infty$ while $n$ is fixed, where $p$ is the data dimension and $n$ is the sample size. For instance, see Hall et al. (2005), Ahn et al. (2007) and Jung and Marron (2009) under the assumption that the population distribution has normality or $\rho$-mixing dependency. On the other hand, Yata and Aoshima (2009) developed the HDLSS asymptotic theory without assuming either the normality or $\rho$-mixing dependency for the spiked covariance model introduced by Johnstone (2001).

In this talk, we first show that HDLSS data sets fall into two types of geometric representations depending on whether a $\rho$-mixing-type dependency appears in variables or not. When a $\rho$-mixing-type dependency appears in variables, a HDLSS data set converges to a $p$-dimensional surface of the unit sphere with increasing $p$. Other than that, a HDLSS data set converges to a point on the axes. If the data structure appears Gaussian, we recommend that the experimenter should apply the noise-reduction methodology, given by Yata and Aoshima (2011b), to estimation of eigenvalues for a HDLSS data set. We show that the noise-reduction methodology gives a consistent estimator of eigenvalues along with its limiting distribution by evaluating the amount of noise with the help of a $\rho$-mixing-type geometric representation. If the data structure appears non-Gaussian, we recommend the cross-data-matrix methodology given by Yata and Aoshima (2010b). The cross-data matrix was introduced by Yata and Aoshima (2010a) for estimation of the intrinsic dimension of HDLSS data. We show that the cross-data-matrix methodology gives a consistent estimator of eigenvalues by using the singular values of the cross-data matrix along with its limiting distribution. We apply the method to estimation for PC directions and PC scores, and give an application for clustering of a HDLSS data set. We also consider the sample size determination for inference on eigenvalues. Finally, we examine how those methods perform well by using microarray data.
Acknowledgement:

This research was partially supported by Grant-in-Aid for Young Scientists (B), Japan Society for the Promotion of Science (JSPS), under Contract Number 23740066.

REFERENCES


Speaker: Makoto Aoshima  
_Institute of Mathematics,  
University of Tsukuba, Japan_

Title: Effective classification for high-dimension, non-Gaussian data and sample size determination

Abstract:

We propose an effective discriminant procedure for high-dimensional and non-Gaussian data. In addition, we give a sample size determination for the proposed discriminant rule to control misclassification rates being no more than a prespecified value. We do not assume that the covariance matrices are common since it is a rather strong assumption and, most importantly, such an assumption is difficult to verify especially for non-Gaussian and high-dimensional data.

High-dimensional data situations occur in many areas of modern science. A common feature of high-dimensional data is that, while the data dimension $p$ is high, the sample size $n$ is relatively small. The inverse of a sample covariance matrix does not exist in such a situation, and hence typical discriminant analysis such as Fisher discriminant analysis (FDA) does not work straightforwardly for high-dimensional data sets. In recent years, several authors proposed FDA-type discriminant methods for high-dimensional data. Dudoit et al. (2002), Bickel and Levina (2004) considered an inverse matrix defined by only diagonal elements of a sample covariance matrix. Yata and Aoshima (2011) considered using a ridge-type inverse covariance matrix derived by the noise reduction methodology. These methods assumed that the population distributions are Gaussian or those covariance matrices are common. On the other hand, Fan and Fan (2008) proposed a features annealed independence rule that selects a subset of important features for classification. Hall et al. (2008) considered distance-based classifiers and asymptotic normality about a discriminant function.

In this talk, we attempt drawing much information from the heteroscedasticity on high-dimensional classification. We pay special attention to a new quadratic-type discriminant rule given by Aoshima and Yata (2011) for non-Gaussian and high-dimensional data. This rule draws information about heteroscedasticity via the difference of the trace of the covariance matrices.
We show that the quadratic-type discriminant rule assures higher accuracy as the difference of the trace grows. We also show that the discriminant function holds the asymptotic normality when the dimension grows. With the help of the asymptotic normality, we give a sample size determination for the discriminant rule to control misclassification rates being no more than a prespecified value.

Acknowledgement:

This research was partially supported by Grant-in-Aid for Scientific Research (B) (No. 22300094) and Grant-in-Aid for Challenging Exploratory Research (No. 23650142), Japan Society for the Promotion of Science (JSPS).

REFERENCES


The Third International Workshop in Sequential Methodologies
June 14–16, 2011

Area C.6.1: Wednesday, June 15, 5:00pm
Math Corner Room 380X

Speaker: Rasul A. Khan
Department of Mathematics,
Cleveland State University, Ohio

Title: Two-Stage and Sequential Estimation of the Scale Parameter of a Gamma Distribution with Fixed-Width Intervals

Abstract:
Two-stage and sequential procedures are developed for fixed-width interval estimation of the parameter $\beta$ in a gamma distribution $G(\alpha,\beta)$ when $\alpha$ is known. Exact properties are obtained for the two-stage procedure while some asymptotics and approximations are given for the operating characteristics of the sequential procedure, and some numerical computations are included.

Acknowledgement:
This is joint work with Professor Shelly Zacks of Binghamton University, New York.