DOCTORAL THESIS

High-dimensional covariance matrix estimation with application to Hotelling's tests

Dong, Kai

Date of Award:
2015

ABSTRACT

In recent years, high-dimensional data sets are widely available in many scientific areas, such as gene expression study, finance and others. Estimating the covariance matrix is a significant issue in such high-dimensional data analysis. This thesis focuses on high-dimensional covariance matrix estimation and its application.

First, this thesis focuses on the covariance matrix estimation. In Chapter 2, a new optimal shrinkage estimation of the covariance matrices is proposed. This method is motivated by the quadratic discriminant analysis where many covariance matrices need to be estimated simultaneously. We shrink the sample covariance matrix towards the pooled sample covariance matrix through a shrinkage parameter. Some properties of the optimal shrinkage parameter are investigated and we also provide how to estimate the optimal shrinkage parameter. Simulation studies and real data analysis are also conducted. In Chapter 4, we estimate the determinant of the covariance matrix using some recent proposals for estimating high-dimensional covariance matrix. Specifically, a total of nine covariance matrix estimation methods will be considered for comparison. Through extensive simulation studies, we explore and summarize some interesting comparison results among all compared methods. A few practical guidelines are also made on the sample size, the dimension, and the correlation of the data set for estimating the determinant of high-dimensional covariance matrix. Finally, from a perspective of the loss function, the comparison study in this chapter also serves as a proxy to assess the performance of the covariance matrix estimation.

Second, this thesis focuses on the application of high-dimensional covariance matrix estimation. In Chapter 3, we consider to estimate the high-dimensional covariance matrix based on the diagonal matrix of the sample covariance matrix and apply it to the Hotelling's tests. In this chapter, we propose a shrinkage-based diagonal Hotelling's test for both one-sample and two-sample cases. We also propose several different ways to derive the approximate null distribution under different scenarios of p and n for our proposed shrinkage-based test. Simulation studies show that the

proposed method performs comparably to existing competitors when n is moderate or large, and it is better when n is small. In addition, we analyze four gene expression data sets and they demonstrate the advantage of our proposed shrinkage-based diagonal Hotelling's test.

Apart from the covariance matrix estimation, we also develop a new classification method for a specific type of high-dimensional data, RNA-sequencing data. In Chapter 5, we propose a negative binomial linear discriminant analysis for RNA-Seq data. By Bayes' rule, we construct the classifier by fitting a negative binomial model, and propose some plug-in rules to estimate the unknown parameters in the classifier. The relationship between the negative binomial classifier and the Poisson classifier is explored, with a numerical investigation of the impact of dispersion on the discriminant score. Simulation results show the superiority of our proposed method. We also analyze four real RNA-Seq data sets to demonstrate the advantage of our method in real-world applications.

Keywords: Covariance matrix, Discriminant analysis, High-dimensional data, Hotelling's test, Log determinant, RNA-sequencing data

Contents