DOCTORAL THESIS

New developments in multiple testing and multivariate testing for high-dimensional data

Hu, Zongliang

Date of Award:
2018

Abstract

This thesis aims to develop some new and novel methods in advancing multivariate testing and multiple testing for high-dimensional small sample size data.

In Chapter 2, we propose a likelihood ratio test framework for testing normal mean vectors in high-dimensional data under two common scenarios: the one-sample test and the two-sample test with equal covariance matrices. We derive the test statistics under the assumption that the covariance matrices follow a diagonal matrix structure. In comparison with the diagonal Hotelling's tests, our proposed test statistics display some interesting characteristics. In particular, they are a summation of the log-transformed squared t-statistics rather than a direct summation of those components. More importantly, to derive the asymptotic normality of our test statistics under the null and local alternative hypotheses, we do not need the requirement that the covariance matrices follow a diagonal matrix structure. As a consequence, our proposed test methods are very flexible and readily applicable in practice. Monte Carlo simulations and a real data analysis are also carried out to demonstrate the advantages of the proposed methods.

In Chapter 3, we propose a pairwise Hotelling's method for testing high-dimensional mean vectors. The new test statistics make a compromise on whether using all the correlations or completely abandoning them. To achieve the goal, we perform a screening procedure, pick up the paired covariates with strong correlations, and construct a classical Hotelling's statistic for each pair. While for the individual covariates without strong correlations with others, we apply squared t statistics to account for their respective contributions to the multivariate testing problem. As a consequence, our proposed test statistics involve a combination of the collected pairwise Hotelling's test statistics and squared t statistics. The asymptotic normality of our test statistics under the null and local alternative hypotheses are also derived under some regularity conditions. Numerical studies and two real data examples demonstrate the efficacy of our pairwise Hotelling's test.

In Chapter 4, we propose a regularized t distribution and also explore its applications in multiple testing. The motivation of this topic dates back to microarray

studies, where the expression levels of thousands of genes are measured simultaneously by the microarray technology. To identify genes that are differentially expressed between two or more groups, one needs to conduct hypothesis test for each gene. However, as microarray experiments are often with a small number of replicates, Student's t-tests using the sample means and standard deviations may suffer a low power for detecting differentially expressed genes. To overcome this problem, we first propose a regularized t distribution and derive its statistical properties including the probability density function and the moments. The noncentral regularized t distribution is also introduced for the power analysis. To demonstrate the usefulness of the proposed test, we apply the regularized t distribution to the gene expression detection problem. Simulation studies and two real data examples show that the regularized t-test outperforms the existing tests including Student's t-test and the Bayesian t-test in a wide range of settings, in particular when the sample size is small.

Keywords: High-dimensional data, Hotelling's test, Likelihood ratio test, Log-transformed squared t-statistic, Noncentral regularized t distribution, Paired correlations, Power, Regularized t distribution, Screening, Type I error

Table of Contents