Journal of Multivariate Analysis 143 (2016) 127–142

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data

Kai Dong [a], Herbert Pang [b,c], Tiejun Tong [a,*], Marc G. Genton [d]

[a] Department of Mathematics, Hong Kong Baptist University, Hong Kong
[b] School of Public Health, The University of Hong Kong, Hong Kong
[c] Department of Biostatistics and Bioinformatics, Duke University, USA
[d] CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia

ARTICLE INFO

Article history:
Received 25 June 2014
Available online 16 September 2015

AMS subject classifications:
62H15

Keywords:
Diagonal Hotelling's test
High-dimensional data
Microarray data
Null distribution
Optimal variance estimation

ABSTRACT

High-throughput expression profiling techniques bring novel tools and also statistical challenges to genetic research. In addition to detecting differentially expressed genes, testing the significance of gene sets or pathway analysis has been recognized as an equally important problem. Owing to the "large p small n" paradigm, the traditional Hotelling's T^2 test suffers from the singularity problem and therefore is not valid in this setting. In this paper, we propose a shrinkage-based diagonal Hotelling's test for both one-sample and two-sample cases. We also suggest 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, but 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.

1. Introduction

DNA microarrays allow us to acquire thousands or tens of thousands of gene expression values simultaneously, which introduces novel approaches to genetic research. One important goal of analyzing gene expression microarray data is to detect differentially expressed genes. Recently, biologists and medical scientists have also recognized that testing the significance of gene sets or pathway analysis is an equally important problem [10,20,5,17]. Specifically, if we want to know whether a certain gene set, Z, is significantly differentially expressed in two different treatments, A and B, the testing hypothesis is $H_0 : \boldsymbol{\mu}_{\mathbf{Z}A} = \boldsymbol{\mu}_{\mathbf{Z}B}$, where $\boldsymbol{\mu}_{\mathbf{Z}A}$ and $\boldsymbol{\mu}_{\mathbf{Z}B}$ are the mean vectors of $\mathbf{Z}$ in A and B, respectively. In statistics, this is essentially a two-sample multivariate testing problem. One classical method used to solve such testing problems is Hotelling's T^2 test [13], which is a generalization of Student's t test. This method works when the sample size, n, is larger than the data dimension, p. More generally, in a k-sample experiment, we are interested in whether or not there exist some differences among the k mean vectors of populations.

In this paper, we focus on one-sample and two-sample multivariate testing problems for high-dimensional small sample size data, or equivalently, for "large p small n" data. In such settings, Hotelling's T^2 test suffers from a singularity problem in the covariance matrix estimation and therefore is not valid in this setting. To overcome the singularity problem, some

* Corresponding author.
E-mail address: tongt@hkbu.edu.hk (T. Tong).

http://dx.doi.org/10.1016/j.jmva.2015.08.022
0047-259X/

remedies have been proposed in the literature; see, for example, the non-exact significance test and the randomization test in [6]. These approaches, however, are known to perform poorly in practice due to their complicated estimation of the degrees of freedom and some related issues [1]. In recent years, a number of approaches to improve Hotelling's T^2 test have emerged for testing high-dimensional data. In essence, these approaches can be classified into the following three categories, with the main difference among them how the covariance matrix is handled:

(1) In the first category, the covariance matrix is removed from Hotelling's T^2 statistic to avoid the covariance matrix estimation. This idea was first considered by Bai and Saranadasa [1]. Specifically, they proposed to use $(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)^T(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)$ to replace $(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)^T\mathbf{S}^{-1}(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)$ in Hotelling's T^2 statistic, where $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ are the sample mean vectors and $\mathbf{S}$ is the pooled sample covariance matrix. They demonstrated that the proposed test has better power than Hotelling's T^2 test under the requirement of p and n being of the same order. Recently, Zhang and Xu [37] and Chen and Qin [5] extended this method to "large p small n" data. We refer to the methods in this category as *the unscaled Hotelling's tests*.

(2) In the second category, a regularization method is applied to the covariance matrix estimation to resolve the singularity problem. In this direction, Chen et al. [4] have made a major contribution. They proposed a regularized Hotelling's T^2 test that estimates the covariance matrix by $\mathbf{S}+\lambda\mathbf{I}_p$, where $\mathbf{I}_p$ is the identity matrix and $\lambda > 0$ is a shrinkage parameter. This test works for both $p < n$ and $p \geq n$ cases. Note that a similar method was also proposed in [25], where the form of $\lambda\mathbf{S}+(1-\lambda)\mathbf{I}_p$ is used to estimate the covariance matrix with $0 \leq \lambda < 1$. In the special case of $\lambda = 0$, the test reduces to an unscaled Hotelling's test. We refer to the methods in this category as *the regularized Hotelling's tests*.

(3) In the third category, the covariance matrix is assumed to be diagonal. Under this assumption, the singularity problem is circumvented since a diagonal matrix is always invertible for non-zero entries, whether or not p is larger than n. This idea was first considered by Wu, Genton and Stefanski [35] and then revisited by several other researchers; see, for example, [28,27,22,29]. For more details, see Section 2.1. These methods are essentially all the same and we refer to them as *the diagonal Hotelling's tests*.

In our simulation studies, we note that the unscaled Hotelling's tests are often sensitive to the deviation of equal eigenvalues of the covariance matrix. If one eigenvalue is extremely large, then the performance of the test will be dominated by that individual component and thus a lower power will result. For more details, see the simulation studies in Section 4. In addition, even for the case of equal eigenvalues, Chen and Qin [5] suggested $n = [20\log(p)]$ to have a reasonably large power. For instance, n needs to be at least 46, 92 and 138 for $p = 10$, 100 and 1000, respectively. For high-dimensional data such as gene expression microarray data, however, it is not uncommon that n is very small, say for example less than 10 samples per group [23,8]. This has motivated researchers to consider more realistic testing methods for high-dimensional small sample size data, e.g., the regularized Hotelling's tests and the diagonal Hotelling's tests. Our additional simulation studies indicate that the existing regularized Hotelling's tests do not perform comparably to the diagonal Hotelling's tests when n is relatively small.

In view of the good performance of the diagonal Hotelling's tests, we also assume that the covariance matrix is diagonal in this paper. Before moving forward, we note that this diagonal covariance matrix assumption has been commonly used for high-dimensional small sample size data, e.g., [9,3,32]. In particular, Bickel and Levina [3] pointed out that if the estimated correlations are all very noisy, then we are probably better off without estimating them. This, in essence, is the assumption of a diagonal covariance matrix when n is relatively small. In discriminant analysis, Lee et al. [16] have also observed that discriminant rules with an inverse generalized matrix may not perform as well as diagonal discriminant rules for microarray data. Although very promising, the performance of the diagonal Hotelling's tests themselves can be suboptimal due to the unreliable estimates of the sample variances from the limited number of observations. This suggests that some modifications to the diagonal Hotelling's tests are necessary to further improve their performance. We note that one such attempt has already been made by Dinu et al. [7]. They proposed a modified diagonal Hotelling's test, called "SAM-GS", by adding a small constant to each gene-specific variance estimate to stabilize the variance estimation, an idea originated in the SAM test of Tusher, Tibshirani and Chu [33].

In this paper, we propose a shrinkage-based diagonal Hotelling's test for both one-sample and two-sample cases. The test is structured by replacing the sample variances in the diagonal Hotelling's tests by the optimal shrinkage estimation of variances in [32]. For the proposed shrinkage-based test, we then consider several different ways to derive the approximate null distribution under different scenarios of p and n. Simulation results show that the proposed method always performs comparably to existing competitors, especially when n is less than 10. In addition, to assess the performance of the proposed method using real data, we consider four gene expression data sets. A case study also demonstrates the advantage of the proposed shrinkage-based diagonal Hotelling's test. The remainder of the paper is organized as follows. The shrinkage-based diagonal Hotelling's tests are introduced in Section 2. In Section 3, we derive both a scaled chi-squared null distribution and a normal null distribution. Simulation studies and real data analysis are conducted in Sections 4 and 5, respectively.

2. Improving the diagonal Hotelling's tests

Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$, $i = 1, \ldots, n$, be independent and identically distributed (i.i.d.) random vectors from a multivariate normal distribution, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the population mean vector and $\boldsymbol{\Sigma}$ is the population covariance matrix. Let also $\bar{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i/n$ be the sample mean vector and $\mathbf{S} = \sum_{i=1}^n(\mathbf{X}_i-\bar{\mathbf{X}})(\mathbf{X}_i-\bar{\mathbf{X}})^T/(n-1)$ be the sample covariance