Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

ELSEVIER

An adaptive test for the mean vector in large-p-small-n problems

Q1 Yanfeng Shen [a,*], Zhengyan Lin [b]

[a] Department of Mathematics, Zhejiang Normal University, Jinhua 321004, China

[b] Department of Mathematics, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 13 December 2013
Received in revised form 25 November 2014
Accepted 3 March 2015
Available online xxxx

Keywords:
High-dimensional data
Hypothesis testing
Power
Testing mean vector
Variable selection

ABSTRACT

The problem of testing the mean vector in a high-dimensional setting is considered. Up to date, most high-dimensional tests for the mean vector only make use of the marginal information from the variables, and do not incorporate the correlation information into the test statistics. A new testing procedure is proposed, which makes use of the covariance information between the variables. The new approach is novel in that it can select important variables that contain evidence against the null hypothesis and reduce the impact of noise accumulation. Simulations and real data analysis demonstrate that the new test has higher power than some competing methods proposed in the literature.

1. Introduction

The problem of hypotheses testing about the mean vector for high-dimensional data is frequently encountered in many Q2 contemporary statistical studies. For instance, in the analysis of biomedical data, it is of great interest to identify differentially expressed pathways or gene sets across different experimental treatments (e.g. Goeman et al., 2004; Nam and Kim, 2008; Tsai and Chen, 2009; Chen et al., 2011). One remarkable aspect of such high-dimensional data is that the number of variables is much larger than the number of observations. This is often refereed to as "large-p-small-n" paradigm, which causes most classical methodologies to break down such as famous Hotelling's T^2 test. As such, new statistical approaches are needed to Q3 develop for high-dimensional data analysis.

In the present study, we focus mainly on both one-sample problem and two-sample problem in a high-dimensional setting. These topics have received great attention in last decades, and many testing approaches have been proposed by different authors. To generalize Hotelling's T^2 test for high-dimensional inference, a simple and popular strategy is to replace the sample covariance with a given nonsingular matrix. For example, Dempster (1958) and Bai and Saranadasa (1996) (henceforth BS-test) modified the T^2 test statistics by replacing the sample covariance matrix by the identity matrix, and Bai and Saranadasa (1996) established their asymptotic distributions when both the sample size and the number of variables tend to infinity. Furthermore, Chen and Qin (2010) (henceforth CQ-test) provided a modification version of the test proposed by Bai and Saranadasa (1996), and also obtained its asymptotic distributions under much weaker conditions than those in Bai and Saranadasa (1996). Note that these three tests are invariant under an orthogonal transformation. On the other hand, Srivastava and Du (2008) (henceforth SD-test) introduced another Hotelling-like test, which replaces the

[*] Corresponding author. Tel.: +86 579 8229 8188.
E-mail address: syf@zjnu.cn (Y. Shen).

http://dx.doi.org/10.1016/j.csda.2015.03.004
0167-9473/© 2015 Elsevier B.V. All rights reserved.

sample covariance matrix by the diagonal of the sample covariance matrix. It is worth to note that this test has the property of scalar transformation invariance.

All of the tests mentioned earlier have a drawback in common—that is, they typically do not make use of the correlation structure between the variables. However, in microarray gene expression studies, it is often the case that different genes in the same genetic regulatory network could be highly correlated (e.g. Ackermann and Strimmer, 2009, Fan et al., 2012). Intuitively, the above methods do discard the critical information of dependence and hence they may not be optimal. It is reasonable to believe that we could gain more power by incorporating the correlation information into the analysis.

A challenging issue is that how to make use of the covariance structure. Obviously, we cannot directly use the sample covariance matrix, since it becomes singular under high-dimensional settings. Recently, Cai et al. (in press) (henceforth CLX-test) applied a regularization technique to obtain a sparse estimator of the precision matrix, and then suggested a test statistic based on the estimated precision matrix. In this work, we introduce another strategy to make use of the correlation information in high-dimensional testing context. Motivating our approach is the observation that, although the whole covariance matrix cannot be estimated well in high dimension, its low-order submatrices can be estimated well. Thus, we could choose a subset of variables by an optimal criteria, and then utilize the covariance structure from this subset to construct the test statistic.

The purpose of this paper is to provide a new testing approach that can utilize the dependence between the variables to further increase testing power for high-dimensional data. In particular, we pick up a subset of variables such that the corresponding optimal criteria obtains the maximum, and then make use of the information from selected variables to make statistical inference. Our approach is novel that, it can automatically select relevant variables or features that contain evidence against the null hypothesis. As a result, it can efficiently reduce the impact of noise accumulation, and hence it leads to a substantial increase in power in case of testing against sparse alternatives. We also propose a simple greedy algorithm for searching the optimal subsets to avoid the need for exhaustive searches. In addition, we adopt a permutation procedure to estimate the null distribution of our test statistic and compute the p-values. At last, we investigate the performance of this new method with some competing methods via both simulation studies and real data analysis.

2. An adaptive testing procedure for high-dimensional one-sample problem

In this section, we begin with a one-sample problem to introduce the key idea of our new procedure in more detail. Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ be a random sample generated from a p-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$. The one-sample problem is to test the following hypothesis:

$$H_0 : \boldsymbol{\mu} = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \mathbf{0}, \tag{1}$$

with unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. We also assume that the covariance matrix $\boldsymbol{\Sigma}$ is positive definite.

2.1. Motivation and a new testing procedure

When the dimension p is smaller than the sample size n, typical Hotelling's T^2 test for problem (1) is based on the test statistic

$$T^2 = n\bar{\mathbf{x}}^T\mathbf{S}^{-1}\bar{\mathbf{x}},$$

where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ is the sample mean vector, and $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ is the sample covariance matrix. Theoretical properties of Hotelling's T^2 test have been well studied when p is fixed; see, for example Anderson (2003). On the other hand, Bai and Saranadasa (1996) investigated the asymptotic properties of T^2 under $p/n \to y \in (0, 1)$, and obtained its asymptotic power that has the form of

$$\Phi\left(-z_{1-\alpha} + \sqrt{\frac{n(1-y)}{2y}}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right), \tag{2}$$

where α is the nominal significance level, $z_{1-\alpha}$ denotes the $100(1-\alpha)$th quantile of the standard normal distribution, and Φ is the distribution function of a standard normal random variable.

In high-dimensional settings where $p \geq n$, Hotelling's T^2 test can no longer be applied since the sample covariance matrix $\mathbf{S}$ becomes singular. To deal with the challenges associated with high dimensionality, one intuitive and feasible strategy is dimension reduction. An interesting approach proposed by Lopes et al. (2011) is to use a random projection technique to select a lower-dimensional feature space, and then make inferences by using Hotelling's T^2 test in the selected subspaces. However, the performance of such a random projection may not be optimal because the selected subspaces could include many noise variables that have no contribution to the detection power. To address this issue, we here introduce an alternative approach based on a supervised-learning strategy. In particular, we pick up an optimal subset of features such that it maximizes the asymptotic power of Hotelling's T^2 test in (2).

Please cite this article in press as: Shen, Y., Lin, Z., An adaptive test for the mean vector in large-p-small-n problems. Computational Statistics and Data Analysis (2015), http://dx.doi.org/10.1016/j.csda.2015.03.004