Computational Statistics and Data Analysis 55 (2011) 2221–2233

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Shrinkage-based regularization tests for high-dimensional data with application to gene set analysis

Yanfeng Shen [a], Zhengyan Lin [a], Jun Zhu [a,b,*]

[a] *Department of Mathematics, Zhejiang University, Hangzhou 310027, China*
[b] *College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310029, China*

ARTICLE INFO

Article history:
Received 24 February 2010
Received in revised form 2 October 2010
Accepted 31 December 2010
Available online 6 January 2011

Keywords:
Feature selection
Gene set analysis
High dimensionality
MANOVA
Power
Regularization

ABSTRACT

Traditional multivariate tests such as Hotelling's test or Wilk's test are designed for classical problems, where the number of observations is much larger than the dimension of the variables. For high-dimensional data, however, this assumption cannot be met any longer. In this article, we consider testing problems in high-dimensional MANOVA where the number of variables exceeds the sample size. To overcome the challenges with high dimensionality, we propose a new approach called a shrinkage-based regularization test, which is suitable for a variety of data structures including the one-sample problem and one-way MANOVA. Our approach uses a ridge regularization to overcome the singularity of the sample covariance matrix and applies a soft-thresholding technique to reduce random noise and improve the testing power. An appealing property of this approach is its ability to select relevant variables that provide evidence against the hypothesis. We compare the performance of our approach with some competing approaches via real microarray data and simulation studies. The results illustrate that the proposed statistics maintains relatively high power in detecting a wide family of alternatives.

1. Introduction

With recent advances of high-throughput techniques, high-dimensional data, such as gene expression and proteomics data, has become increasingly popular in many modern statistical studies. These new classes of data have a remarkable feature in common—that is, the dimensionality of the variables or parameters p is much greater than the number of observations n. This "large p, small n" problem poses various computational and statistical challenges to classical statistical methodologies; see, for example, Donoho (2000) and Fan and Li (2006) for comprehensive surveys of statistical issues created by high dimensionality.

In this article, we focus on the problem of comparing mean vectors in high-dimensional MANOVA (HMANOVA), where the dimension of variables is much larger than the sample size. This statistical issue has become an important research topic in many practical settings, for instance, gene set analysis that aims to detect differentially expressed gene sets or pathways across several experimental conditions (e.g., Goeman et al., 2004; Nam and Kim, 2008; Tsai and Chen, 2009). In general, there are two main challenges with high dimensionality in HMANOVA. First, when $p > n$, the standard sample covariance matrix becomes singular with probability 1, and hence it cannot be inverted any more. Consequentially, many classical methods, such as Hotelling's T^2 test and Wilk's Λ test, break down completely in such high-dimensional settings. One intuitive and naive strategy is to use the generalized inverse matrix of the sample covariance matrix to generalize

[*] Corresponding author at: Department of Mathematics, Zhejiang University, Hangzhou 310027, China. Tel.: +86 571 8697 1731.
E-mail address: jzhu@zju.edu.cn (J. Zhu).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.12.013

Hotelling's T^2 test (Srivastava, 2007). However, the performance of such an approach is very unstable, since the high-dimensional covariance matrix cannot be efficiently estimated from a relatively small number of observations. Another possible and often efficient strategy is to regularize the sample covariance matrix. The simplest one may be that known as the independence rule (Tibshirani et al., 2002; Bickel and Levina, 2004; Fan and Fan, 2008), which regularizes the covariance matrix to a diagonal matrix. In fact, this rule makes an essential assumption that all variables are mutually independent, and hence dramatically reduces the number of parameters to be estimated. The same idea is also referred to as the "naive Bayes" assumption in the machine learning literature. A more general approach is to shrink the sample covariance matrix toward an identity matrix. Such a method has been successfully applied in the context of high-dimensional discriminant analysis (Friedman, 1989; Xu et al., 2009). In effect, this method can be viewed as a James–Stein estimator of the covariance matrix (Ledoit and Wolf, 2004). More recently, Tsai and Chen (2009) and Warton (2008) introduced this technique for testing statistical hypotheses in HMANOVA and high-dimensional multivariate regression, respectively. Another challenge is that a test for detecting high-dimensional alternatives could suffer from low power due to noise accumulation from a lot of irrelevant features that have no contribution to the discriminability power. Fan (1996) first pointed out the impact of high dimensionality in the testing problem of a Gaussian white noise model. This example indicates that feature selection and noise reduction are of great importance in high-dimensional data analysis. Feature selection techniques have recently received much attention in the context of high-dimensional classification for improving prediction accuracy and model interpretability (e.g., Hastie et al., 2001). However, little literature has considered incorporating feature selection techniques into a testing procedure. More recently, Wu et al. (2009) proposed an L_1 penalty version of the linear discriminant statistic to jointly test a two-sample problem and to select features in high-dimensional settings.

The aim of this paper is to establish a unified framework to simultaneously deal with the above-mentioned challenges in HMANOVA. Here we propose a hybrid approach named the shrinkage-based regularization test to generalize some classical multivariate methods. Our approach is novel in that, while a ridge regularization of the sample covariance matrix is used as one way out of the difficulty of singularity, a soft-thresholding technique (Donoho and Johnstone, 1994) is incorporated into the test statistics to reduce most noise and hence improve the testing power. Due to the nature of the soft-thresholding method, the approach allows researchers to select relevant variables or features that are important for detecting alternatives. Moreover, our approach is suited to a variety of commonly encountered problems, including the one-sample problem and one-way MANOVA. In order to make our testing method practical, we must determine two kinds of tuning parameter, and know the distribution of each statistic under the null hypothesis. From the consideration of power and computational expense, two simple but efficient strategies have been adopted to select the values of these tuning parameters, respectively. And we propose two kinds of randomization procedure to estimate the null distributions of the test statistics and calculate the p-values. It is worth noting that these randomization tests have exact level α for any sample size and any dimension of variables. Simulation and real human diabetes data studies demonstrate that our approach can have higher power than some competing methods in detecting a large family of alternatives.

2. One-sample problem for high-dimensional data

Let us begin with a one-sample problem to describe the core idea of our approach in more detail. Consider a random sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ from a p-dimensional normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$. In the one-sample problem, we wish to test a high-dimensional hypothesis

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

In practical applications, one may want to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\eta}_0$, where $\boldsymbol{\eta}_0$ is a given p-dimensional vector. In fact, such a problem can be easily transformed to problem (1) if we use a new sample $\mathbf{x}_1 - \boldsymbol{\eta}_0, \mathbf{x}_2 - \boldsymbol{\eta}_0, \ldots, \mathbf{x}_n - \boldsymbol{\eta}_0$ in place of the original one.

2.1. The impact of high dimensionality

We assume for the moment that the true covariance matrix $\boldsymbol{\Sigma}$ is positive definite and known. According to the Neyman–Pearson lemma, given a fixed alternative $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, the most powerful test for problem (1) is based on the statistic $n\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}$, where $\hat{\boldsymbol{\mu}} = n^{-1}\sum_{1\le i\le n}\mathbf{x}_i$ is the sample mean vector. It is easy to show that the exact power of this optimal test is

$$1 - \Phi(z_{1-\alpha} - \sqrt{n\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0}),$$

where Φ is the standard normal distribution function, α is a pre-specified significance level and $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. In practice, however, such a true alternative is not known. A simple and obvious strategy is to estimate $\boldsymbol{\mu}_0$ by the sample mean vector $\hat{\boldsymbol{\mu}}$, and then we get a Wald-type statistic $T = n\hat{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}$. Note that, under H_1, $(T - p - n\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0)/\sqrt{2p + 4n\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0}$ converges in distribution to a standard normal distribution as $p \to \infty$. Thus, the power of such a test at the alternative $\boldsymbol{\mu}_0$ can be approximated by

$$1 - \Phi\left(\frac{z_{1-\alpha} - n\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0/\sqrt{2p}}{\sqrt{1 + 2n\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0/p}}\right).$$