Journal of Multivariate Analysis 155 (2017) 200–216

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Linear hypothesis testing in high-dimensional one-way MANOVA

Jin-Ting Zhang*, Jia Guo, Bu Zhou

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore

ARTICLE INFO

Article history:
Received 10 July 2016
Available online 9 January 2017

AMS 2010 subject classifications:
primary 62H15
secondary 62F05

Keywords:
High-dimensional data
L^2-norm based test
χ^2-type mixtures
One-way MANOVA
Welch–Satterthwaite χ^2 approximation

ABSTRACT

In recent years, with the rapid development of data collecting technologies, high-dimensional data have become increasingly prevalent. Much work has been done for testing hypotheses on mean vectors, especially for high-dimensional two-sample problems. Rather than considering a specific problem, we are interested in a general linear hypothesis testing (GLHT) problem on mean vectors of several populations, which includes many existing hypotheses about mean vectors as special cases. A few existing methodologies on this important GLHT problem impose strong assumptions on the underlying covariance matrix so that the null distributions of the associated test statistics are asymptotically normal. In this paper, we propose a simple and adaptive test based on the L^2-norm for the GLHT problem. For normal data, we show that the null distribution of our test statistic is the same as that of a chi-squared type mixture which is generally skewed. Therefore, it may yield misleading results if we blindly approximate the underlying null distribution of our test statistic using a normal distribution. In fact, we show that the null distribution of our test statistic is asymptotically normal only when a necessary and sufficient condition on the underlying covariance matrix is satisfied. This condition, however, is not always satisfied and it is not an easy task to check if it is satisfied in practice. To overcome this difficulty, we propose to approximate the null distribution of our test statistic using the well-known Welch–Satterthwaite chi-squared approximation so that our new test is applicable without any assumption on the underlying covariance matrix. Simple ratio-consistent estimators of the unknown parameters are obtained. The asymptotic and approximate powers of our new test are also investigated. The methodologies are then extended for non-normal data. Four simulation studies and a real data application are presented to demonstrate the good performance of our new test compared with some existing testing procedures available in the literature.

1. Introduction

With data collecting technology development, high-dimensional data, whose dimensionality may be in hundreds or thousands and may be much larger than the sample size, are becoming increasingly prevalent. Examples include DNA microarray and high-frequency financial data among others. The work of this paper was motivated by the peripheral blood mononuclear cells (PBMC) data, available at http://www.ncbi.nlm.nih.gov/gds with accession ID GDS1615. The PBMC data set, which has been studied by [4], consists of 42 normal, 26 ulcerative colitis (UC) and 59 Crohn's disease (CD) tissues,

* Corresponding author.
E-mail address: stazjt@nus.edu.sg (J.-T. Zhang).

http://dx.doi.org/10.1016/j.jmva.2017.01.002
0047-259X/

each having 22,283 gene expression levels. Of interest is to test whether the three groups of tissues have the same mean expression levels. This is a one-way MANOVA (multivariate analysis of variance) testing problem for high-dimensional data. A general one-way MANOVA problem for high-dimensional data can be described as follows.

Suppose we have the following k independent samples:

$$\mathbf{y}_{i1}, \ldots, \mathbf{y}_{in_i} \text{ are i.i.d. with } \mathrm{E}(\mathbf{y}_{i1}) = \boldsymbol{\mu}_i, \qquad \mathrm{cov}(\mathbf{y}_{i1}) = \boldsymbol{\Sigma}, \quad i = 1, \ldots, k, \tag{1.1}$$

where the data dimension p is large and may be much larger than the total sample size $n = n_1 + \cdots + n_k$, one of the goals is to test whether the k mean vectors are equal:

$$\mathcal{H}_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_k \quad \text{vs.} \quad \mathcal{H}_1 : \mathcal{H}_0 \text{ is not true.} \tag{1.2}$$

High-dimensional data analysis is challenging because the sample covariance matrices involved in the classical testing procedures are degenerate so that the classical testing procedures, such as the well-known Hotelling's T^2 test and the Lawley–Hotelling trace test, among others, are no longer powerful or well defined. To overcome this difficulty, a number of authors have proposed many alternatives in the past decades. When $k = 2$, the problem (1.2) reduces to a two-sample problem for high-dimensional data. When the covariance matrices of the two samples are the same, the first two articles, published over fifty years ago [8,9], proposed some non-exact tests using an F distribution approximation. In 1996, [2] then proposed a non-exact test via a modification of the Hotelling's T^2 test. More recently, a number of different approaches have been proposed, including scale-invariant tests [5,10,27], empirical likelihood ratio tests [18,35], simulation or permutation-based tests [19,36], random-projection and subspace-based tests [21–23,31,41,44], and nonparametric tests [14,20,34], among others. In most of this literature, strong assumptions on the underlying covariance matrices are required, an exception being [17], where a weighted two-sample test with a proper choice of the weight matrix was proposed; another is [40], where a simple and adaptive L^2-norm based test with the Welch–Satterthwaite χ^2 approximation was considered. When the covariance matrices of the two samples are different, applications of the above two-sample tests may lead to misleading results. To overcome this difficulty, several alternatives have been proposed, including U-statistic based tests [1,7,12] and scale-invariant tests [11,15,29], among others.

For a general $k > 2$, when the covariance matrices of the k samples are the same, for the one-way MANOVA problem (1.2), [13] derived the asymptotic normality of several classical MANOVA tests in a high-dimensional setting, [26] constructed a MANOVA test, extending the work of [2], [30] proposed a scale-invariant MANOVA test for non-normal data, [24] investigated the asymptotics of the Dempster trace criterion, and [6] proposed a linear transformation-based test that can take the dependence structure of the variables into account. When the k samples have different covariance matrices, [42] proposed an approximate solution to the k-sample Behrens–Fisher problem by transforming a k-sample problem into a one-sample problem, [28] proposed a general linear hypothesis test under multivariate linear regression models for normal data, and [38] proposed a MANOVA test based on variation matrices due to hypothesis and the unbiased estimator of the covariance matrices.

In this paper, we are interested in testing the following general linear hypothesis testing (GLHT) problem:

$$\mathcal{H}_0 : \tilde{\mathbf{G}}\mathbf{M} = \mathbf{0} \quad \text{vs.} \quad \mathcal{H}_1 : \tilde{\mathbf{G}}\mathbf{M} \neq \mathbf{0}, \tag{1.3}$$

where $\mathbf{M} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)^\top$ is a $k \times p$ matrix collecting all the k mean vectors and $\tilde{\mathbf{G}} : q \times k$ is a known full-rank coefficient matrix with $\mathrm{rank}(\tilde{\mathbf{G}}) = q < k$. [32] considered a hypothesis testing problem similar to (1.3) for a one-sample problem of transposable data where an observation is a realization of a random matrix. When the group sample sizes of the k samples (1.1) are the same, our GLHT problem may be transformed into the problem they considered. Note that the GLHT problem (1.3) includes the one-way MANOVA test (1.2) as a special case. In fact, the GLHT problem (1.3) reduces to the one-way MANOVA test (1.2) when we set $\tilde{\mathbf{G}}$ to be either

$$\tilde{\mathbf{G}}_1 = \left(\mathbf{I}_{k-1}, -\mathbf{1}_{k-1}\right) \quad \text{or} \quad \tilde{\mathbf{G}}_2 = \left(-\mathbf{1}_{k-1}, \mathbf{I}_{k-1}\right), \tag{1.4}$$

where $\mathbf{I}_r$ and $\mathbf{1}_r$ denote the identity matrix of size r and the r-dimensional vector of 1's, respectively. Actually, the GLHT problem (1.3) is very general. It includes not only the one-way MANOVA test (1.2) but also various post hoc and contrast tests as special cases since any post hoc and contrast tests can be written in the form (1.3). For example, when the one-way MANOVA test is rejected, it is of interest to further test if $\boldsymbol{\mu}_1 = 3\boldsymbol{\mu}_2$ or if a contrast is zero, e.g., $\boldsymbol{\mu}_1 - 3\boldsymbol{\mu}_2 + 2\boldsymbol{\mu}_3 = 0$. In fact, these two testing problems can be written in the form (1.3) with $\tilde{\mathbf{G}} = \left(\mathbf{e}_{1,k} - 3\mathbf{e}_{2,k}\right)^\top$ and $\tilde{\mathbf{G}} = \left(\mathbf{e}_{1,k} - 3\mathbf{e}_{2,k} + 2\mathbf{e}_{3,k}\right)^\top$, respectively, where $\mathbf{e}_{r,\ell}$ denotes throughout a unit vector of length ℓ whose rth entry is 1 while all the others equal 0.

Although the GLHT problem (1.3) is very important, we are aware of only a few articles devoted to test it. [28] considered testing (1.3) in the context of multivariate linear regression models for normal data and [30] for non-normal data. Recently, [43] considered this GLHT problem in the context of one-way MANOVA under heteroscedasticity using a test statistic based on U-statistics of the k samples (1.1). In these three articles, strong assumptions are imposed on the underlying covariance matrices of the k samples (1.1) so that the null distributions of the associated test statistics will tend to normal as the sample sizes tend to infinity. This shows that the above three tests are useful only for some underlying covariance matrices. To overcome this difficulty, in this paper, we propose and study an L^2-norm based test for the GLHT problem (1.3), extending the work of [40] so that the resulting test is useful for any underlying covariance matrices. In Section 5, we demonstrate via simulations that in terms of size controlling, our new test indeed works well for the one-way