Journal of Multivariate Analysis 114 (2013) 349–358

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

A two sample test in high dimensional data

Muni S. Srivastava [a], Shota Katayama [b,*], Yutaka Kano [b]

[a] Department of Statistics, University of Toronto, 100 St.George Street, Toronto, Ontario M5S 3G3, Canada

[b] Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

ARTICLE INFO

Article history:
Received 6 August 2011
Available online 27 August 2012

AMS subject classifications:
primary 62H15
secondary 62E20

Keywords:
High-dimensional data
Hypothesis testing
Behrens–Fisher problem
Asymptotic theory

ABSTRACT

In this paper we propose a test for testing the equality of the mean vectors of two groups with unequal covariance matrices based on N_1 and N_2 independently distributed p-dimensional observation vectors. It will be assumed that N_1 observation vectors from the first group are normally distributed with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$. Similarly, the N_2 observation vectors from the second group are normally distributed with mean vector $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$. We propose a test for testing the hypothesis that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. This test is invariant under the group of $p \times p$ nonsingular diagonal matrices. The asymptotic distribution is obtained as $(N_1, N_2, p) \to \infty$ and $N_1/(N_1 + N_2) \to k \in (0, 1)$ but N_1/p and N_2/p may go to zero or infinity. It is compared with a recently proposed non-invariant test. It is shown that the proposed test performs the best.

1. Introduction

Let $\boldsymbol{x}_{ij}$ be independently distributed as the multivariate normal distribution with the mean vector $\boldsymbol{\mu}_i$ and the positive definite covariance matrix $\boldsymbol{\Sigma}_i$ for $i = 1, 2$ and $j = 1, 2, \ldots, N_i$. For notational convenience, we shall denote it as $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),\ i = 1, 2$, where p denotes the dimension of the random vectors $\boldsymbol{x}_{ij}$. In this article, we consider the problem of testing the hypothesis

$$H : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \tag{1.1}$$

against the alternative

$$A : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2, \tag{1.2}$$

when the covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ of the two groups may be unequal. This problem has recently been considered by Chen and Qin [2] who proposed a test which we denote by T_{cq}. The test T_{cq} will be described in Section 2 from which it will be clear that it is a rather complicated test and requires considerable terms in programming and computing. Also, it is shown that the T_{cq} test is almost identical to a test that can be obtained by generalizing the Bai and Saranadasa [1] test when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. In addition, the test T_{cq}, although invariant under the group of orthogonal transformations, is not invariant under the units of measurements. That is, if we consider $\boldsymbol{D}\boldsymbol{x}_{ij}$ instead of $\boldsymbol{x}_{ij}$, where $\boldsymbol{D}$ is a nonsingular $p \times p$ diagonal matrix, the test T_{cq} changes, which is an undesirable feature. It may be noted that when N_i is less than p, no fully affine invariant test exists. Thus, in this article, we propose a test that is invariant under the transformation of the observation vector $\boldsymbol{x}_{ij}$ by nonsingular

* Corresponding author.
E-mail addresses: srivasta@utstat.utoronto.ca (M.S. Srivastava), sfujimoto@sigmath.es.osaka-u.ac.jp (S. Katayama), kano@sigmath.es.osaka-u.ac.jp (Y. Kano).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.08.014

$p \times p$ diagonal matrices. It will be shown that this new test, denoted by T, performs better than T_{cq}. To describe this new test T, we introduce some notations with $n_i = N_i - 1,\ i = 1, 2$:

$$\bar{\boldsymbol{x}}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{x}_{ij} \quad \text{and} \quad \boldsymbol{S}_i = \frac{1}{n_i}\sum_{j=1}^{N_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)'. \tag{1.3}$$

In high dimensional data, since N_i may be less than p, the sample covariance matrices $\boldsymbol{S}_i$ may be singular. However, the diagonal matrices consisting of only the diagonal elements of $\boldsymbol{S}_i = (s_{ijk}),\ i = 1, 2$, namely,

$$\hat{\boldsymbol{D}}_i = \operatorname{diag}(s_{i11}, \ldots, s_{ipp}), \quad i = 1, 2, \tag{1.4}$$

are non-singular matrices. Let

$$\hat{\boldsymbol{D}} = \frac{\hat{\boldsymbol{D}}_1}{N_1} + \frac{\hat{\boldsymbol{D}}_2}{N_2} = (\hat{d}_{ij}). \tag{1.5}$$

Then

$$\boldsymbol{R} = \hat{\boldsymbol{D}}^{-1/2}\left(\frac{\boldsymbol{S}_1}{N_1} + \frac{\boldsymbol{S}_2}{N_2}\right)\hat{\boldsymbol{D}}^{-1/2} = (r_{ij}) \tag{1.6}$$

is the sample correlation matrix, while $\boldsymbol{S}_i$ may not converge to $\boldsymbol{\Sigma}_i$ in probability since N_i may be less than p, $\hat{\boldsymbol{D}}_i$ converges in probability to $\boldsymbol{D}_i$, where

$$\boldsymbol{D}_i = \operatorname{diag}(\sigma_{i11}, \ldots, \sigma_{ipp}), \qquad \boldsymbol{\Sigma}_i = (\sigma_{ijk}), \quad i = 1, 2, \tag{1.7}$$

if $\max_{1\le k\le p}\sigma_{ikk} < \infty$ uniformly in p. Let

$$\boldsymbol{D} = \frac{\boldsymbol{D}_1}{N_1} + \frac{\boldsymbol{D}_2}{N_2} = (d_{ij}). \tag{1.8}$$

Then, $\hat{\boldsymbol{D}} \to \boldsymbol{D}$ in probability. Similar to the sample correlation matrix $\boldsymbol{R}$, we define the population correlation matrix $\mathcal{R}$ by

$$\mathcal{R} = \boldsymbol{D}^{-1/2}\left(\frac{\boldsymbol{\Sigma}_1}{N_1} + \frac{\boldsymbol{\Sigma}_2}{N_2}\right)\boldsymbol{D}^{-1/2} = (\rho_{ij}). \tag{1.9}$$

We note that under the null hypothesis H in (1.1),

$$\begin{aligned} E[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{D}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)] &= \operatorname{tr}\boldsymbol{D}^{-1}\left(\frac{\boldsymbol{\Sigma}_1}{N_1} + \frac{\boldsymbol{\Sigma}_2}{N_2}\right) \\ &= \operatorname{tr}\mathcal{R} = p. \end{aligned}$$

Also, under the null hypothesis H in (1.1),

$$\begin{aligned} \operatorname{Var}[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{D}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)] &= \operatorname{Var}(\bar{\boldsymbol{x}}_1'\boldsymbol{D}^{-1}\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2'\boldsymbol{D}^{-1}\bar{\boldsymbol{x}}_2 - 2\bar{\boldsymbol{x}}_1'\boldsymbol{D}^{-1}\bar{\boldsymbol{x}}_2) \\ &= \operatorname{Var}(\bar{\boldsymbol{x}}_1'\boldsymbol{D}^{-1}\bar{\boldsymbol{x}}_1) + \operatorname{Var}(\bar{\boldsymbol{x}}_2'\boldsymbol{D}^{-1}\bar{\boldsymbol{x}}_2) + 4\operatorname{Var}(\bar{\boldsymbol{x}}_1'\boldsymbol{D}^{-1}\bar{\boldsymbol{x}}_2) \\ &= \frac{2\operatorname{tr}(\boldsymbol{D}^{-1}\boldsymbol{\Sigma}_1)^2}{N_1^2} + \frac{2\operatorname{tr}(\boldsymbol{D}^{-1}\boldsymbol{\Sigma}_2)^2}{N_2^2} + \frac{4\operatorname{tr}\boldsymbol{D}^{-1}\boldsymbol{\Sigma}_1\boldsymbol{D}^{-1}\boldsymbol{\Sigma}_2}{N_1N_2} \\ &= 2\operatorname{tr}\left[\left(\frac{\boldsymbol{D}^{-1/2}\boldsymbol{\Sigma}_1\boldsymbol{D}^{-1/2}}{N_1}\right) + \left(\frac{\boldsymbol{D}^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{D}^{-1/2}}{N_2}\right)\right]^2 \\ &= 2\operatorname{tr}\mathcal{R}^2. \end{aligned}$$

Following Corollary 2.6 of [5], we have for $i = 1, 2$ and $j = 1, \ldots, p$ that $E(s_{ijj}^{-1}) = \sigma_{ijj}^{-1} + O(N_i^{-1})$. Hence, $s_{ijj}^{-1} = \sigma_{ijj}^{-1} + O_p(N_i^{-1})$. Thus,

$$\hat{\boldsymbol{D}}^{-1} = \boldsymbol{D}^{-1}[1 + O_p(N_m^{-1})], \qquad N_m = \min(N_1, N_2),$$

which implies

$$\begin{aligned} \hat{q}_n &= \frac{(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\hat{\boldsymbol{D}}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - p}{\sqrt{p}} \\ &= \frac{(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{D}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - p[1 + O_p(N_m^{-1})]}{\sqrt{p}}[1 + O_p(N_m^{-1})] \\ &= \tilde{q}_n + O_p\left(\frac{\sqrt{p}}{N_m}\right), \end{aligned} \tag{1.10}$$