Journal of Statistical Planning and Inference 143 (2013) 929–943

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

A test for the mean vector in large dimension and small samples

Junyong Park*, Deepak Nag Ayyala

Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21250, USA

ARTICLE INFO

Article history:
Received 24 August 2011
Received in revised form
23 October 2012
Accepted 9 November 2012
Available online 19 November 2012

Keywords:
Asymptotic distribution
High dimension
Testing mean vector
Scalar transform invariant test

ABSTRACT

In this paper, we consider the problem of testing the mean vector in the multivariate setting where the dimension p is greater than the sample size n, namely a large p and small n problem. We propose a new scalar transform invariant test and show the asymptotic null distribution and power of the proposed test under weaker conditions than Srivastava (2009). We also present numerical studies including simulations and a real example of microarray data with comparison to existing tests developed for a large p and small n problem.

1. Introduction

Let X_i, $i=1,\ldots,n$ be independent and identically distributed p-dimensional random vectors with the mean vector μ and covariance matrix Σ. The one sample testing problem

$$H_0 : \mu = 0 \quad vs. \quad H_1 : \mu \neq 0 \tag{1}$$

with unknown μ and Σ, and the two sample case

$$H_0 : \mu_1 = \mu_2 \quad vs. \quad H_1 : \mu_1 \neq \mu_2 \tag{2}$$

with mean vectors μ_i and Σ, $i=1,2$ have been extensively studied by many researchers. For the one sample case (1), one typical test statistic is Hotelling's T^2, $n\overline{X}'S^{-1}\overline{X}$ where $\overline{X}$ and S are the sample mean vector and sample covariance matrix, respectively. However, Hotelling's T^2 has the limitation that it cannot be applied to the case of $p > n-1$ due to the singularity of S. Dempster (1958) (henceforth, D-test) and Bai and Saranadasa (1996) (henceforth BS-test) proposed test statistics which avoid the use of S^{-1}, however, these tests still have some limitations in the sense that they are based on the assumption that p increases at the same rate as the sample size n, $p/n \to c > 0$. In practice, lots of recent data sets have the situation, $p \gg n$, namely ultra-high dimension, for example, microarrays where thousands of genes are observed with only tens of samples to draw inferences from. For such a high dimensional data, Chen and Qin (2010) (henceforth CQ-test) modified the BS-test. This modification allowed the derivation of results similar to BS-test in Bai and Saranadasa (1996) but without any direct relationship between p and n. All these tests such as the D-test, the BS-test and the CQ-test are invariant under an orthogonal transformation, i.e., invariant under $X \to c\Gamma X$ where c is a nonzero constant and Γ is an orthogonal transformation, i.e., $\Gamma'\Gamma = I$ where I is an identity matrix. On the other hand, Srivastava and Du (2008) (henceforth SD-test) and Srivastava (2009) (henceforth S-test) proposed a test which has the property of scalar

* Corresponding author. Tel.: +1 410 455 2407; fax: +1 410 455 1066.
E-mail addresses: junpark@math.umbc.edu, junpark@umbc.edu (J. Park).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.11.001

transformation invariance, i.e., $X \to DX$ where $D = diag(d_1, d_2, \ldots, d_p)$ and d_i's are nonzero constants. One interesting result from Srivastava and Du (2008) is that the S-test obtains more power than the D-test and the BS-test when Σ is a diagonal matrix except $\sigma^2 I$. Srivastava and Du (2008) and Srivastava (2009) use the following conditions:

$$\lim_{p \to \infty} \frac{\mathrm{tr}(\mathcal{R}^i)}{p} < \infty, \quad i = 1,2,4 \tag{3}$$

for the derivation of asymptotic distribution of S-test where $\mathcal{R}$ is the population correlation matrix.

In this paper, we propose a new scalar transformation invariant test modifying the S-test with the following condition:

$$\mathrm{tr}(\mathcal{R}^4) = \mathrm{o}(\mathrm{tr}^2(\mathcal{R}^2)) \tag{4}$$

Condition (3) is stronger than (4) since $\mathrm{tr}(\mathcal{R}^4)/\mathrm{tr}^2(\mathcal{R}^2) = p^{-1}(\mathrm{tr}(\mathcal{R}^4)/p)/(\mathrm{tr}^2(\mathcal{R}^2)/p^2) = \mathrm{O}(p^{-1}) = \mathrm{o}(1)$ as $p \to \infty$ from the condition (3). Condition (4) allows $\mathrm{tr}(\mathcal{R}^4)/p \to \infty$ while (3) does not. We improve S-test in the sense that we present the asymptotic null distribution and power of our proposed test with weaker conditions than those in Srivastava (2009). We also provide numerical studies including simulations and a real example of microarray gene expression, which is a typical example of data with large dimension p and small sample n.

The rest of the paper is organized in the following way. In Section 2, we present an overview of existing tests in the one sample problem. In Section 3, we propose a new test in the one sample problem and show the asymptotic null distribution and power of the proposed test. Section 4 includes the extension of the proposed test to the two sample problem with the asymptotic null distribution and power. Section 5 presents remarks on comparison with some other tests, and Sections 6 and 7 include simulation studies and a real data example respectively. Concluding remarks are presented in Section 8.

2. Overview

In this section, we first begin with the one sample testing problem (1) and then extend to the two sample testing problem (2) later in Section 4. Consider p-dimensional observational vectors X_j, $1 \le j \le n$, generated from a factor model in multivariate analysis. The factor model has been used extensively in other literature, for example, Bai and Saranadasa (1996), Chen and Qin (2010) and Srivastava (2009). More formally, $X_j = (x_{j1}, x_{j2}, \ldots, x_{jp})'$ has the form of

$$X_j = \mu + CZ_j \tag{5}$$

where $Z_j = (z_{1j}, \ldots, z_{mj})'$ and z_{ij}'s are continuous random variables, $j = 1, \ldots, n$ and C is a $p \times m$ matrix for some $m \ge p$ such that $\Sigma = CC'$ is a positive definite matrix, say $\Sigma > 0$. The reason we consider $m \ge p$ is to preserve the basic characteristics of the covariance matrix so that the rank and eigenvalues are not affected by the transformation. In particular, Srivastava (2009) considered the case of $p = m$. See also Section 3 in Chen and Qin (2010). Note that the factor model covers multivariate normal distribution when Z_j's are multivariate normal vectors. S is the sample covariance matrix defined by $S = (1/(n-1)) \sum_{j=1}^n (X_j - \overline{X})(X_j - \overline{X})'$ where $\overline{X} = (1/n) \sum_{j=1}^n X_j$. Denote the diagonal matrix of the population variance and sample variance by $D_\sigma = diag(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp})$ and $D_S = diag(s_{11}, s_{22}, \ldots, s_{pp})$ where σ_{ii} and s_{ii} are the diagonal elements in Σ and S, respectively. The population correlation and the sample correlation matrix are $\mathcal{R} = D_\sigma^{-1/2} \Sigma D_\sigma^{-1/2} = (\rho_{ij})$ and $R = D_S^{-1/2} S D_S^{-1/2} = (r_{ij})$ where ρ_{ij} and r_{ij} are the elements in $\mathcal{R}$ and S, respectively.

In (5), we consider $Z_j = (z_{1j}, \ldots, z_{mj})'$ satisfying

$$E(z_{ij}) = 0, E(z_{ij}^2) = 1, \; E(z_{ij}^4) = 3 + \gamma < \infty \tag{6}$$

$$\text{the density of all } x_i \text{ for } 1 \le i \le p \text{ with respect to the Lebesgue measure is uniformly upper bounded} \tag{7}$$

where z_{ij}'s are independent for all $i = 1, \ldots, m$ and $1 \le j \le n$. If the marginal densities of x_i for all $1 \le i \le p$ is uniformly bounded, then the fourth moment of $1/s_{ii}$ is uniformly bounded. γ is the difference between the fourth moment of z_{ij} and that of a standard normal distribution.

For one sample test (1), Dempster (1958) and Bai and Saranadasa (1996) proposed orthonormal transformation invariant tests, namely T_D and T_{BS}, given by

$$T_D = \frac{n\overline{X}'\overline{X}}{\mathrm{tr}(S)} \tag{8}$$

$$T_{BS} = \frac{n\overline{X}'\overline{X} - \mathrm{tr}(S)}{\left[\frac{2(n-1)(n+1)}{(n-2)(n+1)} \left(\mathrm{tr}(S^2) - \frac{\mathrm{tr}^2(S)}{n-1} \right) \right]^{1/2}} \tag{9}$$

Bai and Saranadasa (1996)showed the asymptotic normality of T_D and T_{BS} and both tests achieve the same power in asymptotics. The main motivation of (8) and (9) is avoiding S^{-1} from Hotelling's T^2 test, however there is some limitation in theory such that these two statistics are efficient when $p/n \to c > 0$. Recently, Chen and Qin (2010) modified T_{BS} and