Journal of Multivariate Analysis 149 (2016) 192–198

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Note(s)

A simpler spatial-sign-based two-sample test for high-dimensional data

Yang Li, Zhaojun Wang, Changliang Zou *

LPMC and Institute of Statistics, Nankai University, Tianjin, China

ARTICLE INFO

Article history:
Received 28 May 2015
Available online 28 April 2016

AMS subject classifications:
62H15
62F03

Keywords:
Asymptotic normality
Bias correction
Large p small n
Scalar-invariance
Spatial median

ABSTRACT

This article concerns the tests for the equality of two location parameters when the data dimension is larger than the sample size. Existing spatial-sign-based procedures are not robust with respect to high dimensionality, producing tests with the type-I error rates that are much larger than the nominal levels. We develop a correction that makes the sign-based tests applicable for high-dimensional data, allowing the dimensionality to increase as the square of the sample size. We show that the proposed test statistic is asymptotically normal under elliptical distributions and demonstrate that it has good size and power in a wide range of settings by simulation.

1. Introduction

Assume that $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i} \in \mathbb{R}^p,\ i = 1, 2$ are two independent random samples from p-variate distributions $F_1(\mathbf{x} - \boldsymbol{\mu}_1)$ and $F_2(\mathbf{x} - \boldsymbol{\mu}_2)$ located at p-variate centers $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, with covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. We are concerned about testing the hypothesis

$$\mathcal{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad \mathcal{H}_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 \tag{1}$$

for high-dimensional data which have dimension p that increases to infinity as the number of observations n tends to infinity. The need of high-dimensional two-sample tests arises from various applications; see Chen and Qin [3] for some related discussion.

Conventionally, we deal with this problem by using the famous Hotelling's T^2 test statistic $T_n^2 = n_1 n_2 (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\top \mathbf{S}_n^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)/n$, where $n = n_1 + n_2$, $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ are the two sample means and $\mathbf{S}_n$ is the pooled sample covariance matrix. However, the T_n^2 cannot work for "large p, small n" cases, i.e., $p > n$. Some authors have suggested replacing $\mathbf{S}_n$ with either the identity matrix or the diagonal matrix of $\mathbf{S}_n$. See Bai and Saranadasa [1], Srivastava and Du [10], Chen and Qin [3], Srivastava et al. [11], Feng et al. [5], Gregory et al. [6] and Feng et al. [4].

Statistical performances of the moment-based tests mentioned above would be degraded when the non-normality is severe, especially for heavy-tailed distributions. Some authors considered using multivariate sign-and/or-rank-based approaches to construct robust tests. Under traditional conditions with fixed p, the "inner centering and inner standardization" sign-based procedure is often used (see Randles [9]; Oja [8]), i.e., using the test statistic Q_n^2 =

* Corresponding author.
E-mail address: nk.chlzou@gmail.com (C. Zou).

http://dx.doi.org/10.1016/j.jmva.2016.04.004
0047-259X/

$p\sum_{i=1}^{2} n_i \hat{\mathbf{U}}_i^\top \hat{\mathbf{U}}_i$, where $U(\mathbf{x}) = \|\mathbf{x}\|^{-1}\mathbf{x}I(\mathbf{x} \neq 0)$, $\hat{\mathbf{U}}_i = n_i^{-1}\sum_{j=1}^{n_i} \hat{\mathbf{U}}_{ij}$ and $\hat{\mathbf{U}}_{ij} = U(\boldsymbol{\Omega}^{-1/2}(\mathbf{X}_{ij} - \hat{\boldsymbol{\mu}}))$. $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\Omega}$ are the Hettmansperger & Randles's [7] estimates (HRE) of location and scatter matrix for the pooled sample, satisfying

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i} \hat{\mathbf{U}}_{ij} = \mathbf{0} \quad \text{and} \quad pn^{-1}\sum_{i=1}^{2}\sum_{j=1}^{n_i} \hat{\mathbf{U}}_{ij}\hat{\mathbf{U}}_{ij}^\top = \mathbf{I}_p.$$

Of course, this test statistic is not applicable for the case of $p > n$ either.

Recently, Wang et al. [13] and Feng et al. [4] proposed high-dimensional spatial-sign-based tests for the one-sample and two-sample problems, respectively. The test statistic proposed in the former one is essentially in a similar fashion to Chen and Qin's [3] test statistic, it is but not directly applicable for the two-sample problem due to the bias from estimating the location parameter. Feng et al. [4] further proposed a scalar-invariant test, which is particularly useful when different components have different scales in high-dimensional data. To circumvent the difficulty of estimating additional biases yielded by using the estimation of location parameter to replace the true one, they suggested a "leave-one-out" test statistic which is computationally extensive, especially when n is not too small.

This work is a sequel to Feng et al. [4]; we propose a simpler test statistic and develop a bias correction to the proposed statistic that makes it robust with respect to high dimensionality. The new test is much more computationally efficient compared to Feng et al.'s [4] method. Simulation comparisons show that our procedure has good size and power for a wide range of dimensions, sample sizes and distributions. Finite-sample studies also demonstrate that the proposed method works reasonably well when the underlying distribution is not elliptical. We describe in detail the proposed method in Section 2, and investigate its numerical performance in Section 3. Technical details are included in the Appendix which is presented in the Supplementary Material (see Appendix B).

2. Spatial-sign-based high-dimensional tests

2.1. Model, assumptions and existing works

Let $\{\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i}\}$, $i = 1, 2$ be two independently and identically distributed (i.i.d.) random vectors $\det(\boldsymbol{\Sigma}_i)^{-1/2} f_i(\|\boldsymbol{\Sigma}_i^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_i)\|)$, $i = 1, 2$. Denote $\boldsymbol{\varepsilon}_{ij} = \boldsymbol{\Sigma}_i^{-1/2}(\mathbf{X}_{ij} - \boldsymbol{\mu}_i)$ and $\mathbf{u}_{ij} = U(\boldsymbol{\varepsilon}_{ij})$. The module $\|\boldsymbol{\varepsilon}_{ij}\|$ and the direction $\mathbf{u}_{ij}$ are independent. Furthermore, the direction vector $\mathbf{u}_{ij}$ is uniformly distributed on the p-dimensional unit sphere. Clearly, $\mathrm{E}(\mathbf{u}_{ij}) = 0$ and $\mathrm{cov}(\mathbf{u}_{ij}) = p^{-1}\mathbf{I}_p$.

Motivated by HRE (Hettmansperger & Randles [7]), Feng et al. [4] suggested a simplified version of HRE without considering the off-diagonal elements of the covariance matrix. Denote $\mathbf{D}_i = \mathrm{diag}(d_{i1}, \ldots, d_{ip})$, $i = 1, 2$, where d_{ij} is the jth diagonal element of $\boldsymbol{\Sigma}_i$. We find the estimates of $\boldsymbol{\mu}_i$ and $\mathbf{D}_i$, $(\hat{\boldsymbol{\mu}}_i, \hat{\mathbf{D}}_i)$, which satisfy

$$\frac{1}{n_i}\sum_{j=1}^{n_i} U(\boldsymbol{\epsilon}_{ij}) = 0 \quad \text{and} \quad \frac{p}{n_i}\mathrm{diag}\Big\{\sum_{j=1}^{n_i} U(\boldsymbol{\epsilon}_{ij})U(\boldsymbol{\epsilon}_{ij})^\top\Big\} = \mathbf{I}_p, \tag{2}$$

where $\boldsymbol{\epsilon}_{ij} = \hat{\mathbf{D}}_i^{-1/2}(\mathbf{X}_{ij} - \hat{\boldsymbol{\mu}}_i)$. A recursive algorithm was developed to solve these equation. Feng et al. [4] proposed the following test statistic

$$R_n = -\frac{1}{n_i n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} U^\top(\hat{\mathbf{D}}_{1,i}^{-1/2}(\mathbf{X}_{1i} - \hat{\boldsymbol{\mu}}_{2,j}))U(\hat{\mathbf{D}}_{2,j}^{-1/2}(\mathbf{X}_{2j} - \hat{\boldsymbol{\mu}}_{1,i})),$$

where $\hat{\boldsymbol{\mu}}_{i,j}$ and $\hat{\mathbf{D}}_{i,j}$ are the corresponding location vectors and scatter matrices using "leave-one-out" samples $\{\mathbf{X}_{ik}\}_{k\neq j}$. Under $\mathcal{H}_0$ the expectation of R_n is asymptotically negligible compared to its standard deviation. This facilitates the construction of the test, because there is no need to estimate its expectation. However, the calculation of R_n is of order $O(pn^3)$, and thus this procedure is computationally complex when n_i is large.

2.2. The proposed test

To overcome the drawback of R_n and make the spatial-sign-based test more feasible for large n situations, we suggest the following test statistic

$$T_n = -\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} U^\top(\hat{\mathbf{D}}_1^{-1/2}(\mathbf{X}_{1i} - \hat{\boldsymbol{\mu}}_2))U(\hat{\mathbf{D}}_2^{-1/2}(\mathbf{X}_{2j} - \hat{\boldsymbol{\mu}}_1)),$$

where $\hat{\boldsymbol{\mu}}_i$ and $\hat{\mathbf{D}}_i$, $i = 1, 2$, are the estimators of the location parameters and diagonal matrices using (2) with the samples $\{\mathbf{X}_{ij}\}_{j=1}^{n_i}$. That is, we use the two entire samples to estimate the location parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ rather than the "leave-one-out" samples as in R_n. In contrast to R_n, this scalar-transformation-invariant test statistic requires only $O(n^2 p)$ computation. Its asymptotic null distribution is summarized in Theorem 1 given below.