Journal of Multivariate Analysis 140 (2015) 162–170

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

A high dimensional two-sample test under a low dimensional factor structure

Yingying Ma [a], Wei Lan [b,*], Hansheng Wang [c]

[a] School of Economics and Management, Beihang University, Beijing, 100191, PR China

[b] Statistics School and Center of Statistical Research, Southwestern University of Finance and Economics, Chengdu, Sichuan, 610074, PR China

[c] Guanghua School of Management, Peking University, Beijing, 100871, PR China

ARTICLE INFO

Article history:
Received 20 August 2014
Available online 14 May 2015

AMS subject classification:
62H

Keywords:
China stock market
High-dimensional data
Hypothesis testing
Latent factor structure
Two-sample test

ABSTRACT

Existing high dimensional two-sample tests usually assume that different elements of a high dimensional predictor are weakly dependent. Such a condition can be violated when data follow a low dimensional latent factor structure. As a result, the recently developed two-sample testing methods are not directly applicable. To fulfill such a theoretical gap, we propose here a Factor Adjusted two-Sample Testing (FAST) procedure to accommodate the low dimensional latent factor structure. Under the null hypothesis, together with fairly weak technical conditions, we show that the proposed test statistic is asymptotically distributed as a weighted chi-square distribution with a finite number of degrees of freedom. This leads to a totally different test statistic and inference procedure, as compared with those of Bai and Saranadasa (1996) and Chen and Qin (2010). Simulation studies are carried out to examine its finite sample performance. A real example on China stock market is analyzed for illustration purpose.

1. Introduction

Let $X_{ki} = (X_{ki1}, \ldots, X_{kip})^\top \in \mathbb{R}^p$ be the information collected from the ith subject in the kth ($k = 1, 2$) group, with $E(X_{ki}) = \mu_k \in \mathbb{R}^p$ and $\mathrm{cov}(X_{ki}) = \Sigma$. Then, the classical two-sample test concerns about testing the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 \neq \mu_2$. This is a problem of fundamental importance, and has been commonly encountered in many scientific applications, including economics, finance, genetics, and many others. The corresponding testing procedures (e.g., Hotelling's T^2 test) are well studied when p is fixed [2]. However, when p is much larger than n, the traditional methods, such as Hotelling's T^2 test, cannot be computed, as the sample covariance matrix is not invertible. This makes two-sample test for ultra high dimensional data a problem of importance; see, for example, [3,6,15], and [4].

Recently, a number of useful methods have been developed for high dimensional two-sample test. Nevertheless, their applicability relies on one critical assumption. That is different elements of X_{ki} for $k = 1, 2$ should be weakly dependent. Mathematically, this requires that $tr(\Sigma^4) = o\{tr^2(\Sigma^2)\}$; see, for example, condition (2.8) in [17] and condition (3.7) in [6]. It is worthy mentioning that such an assumption can be violated if the eigenvalues of Σ are dominated by a few top ones. This can happen if X_{ki} follows a low dimensional latent factor structure [9,16]. For the purpose of illustration, we assume that $X_{ki} = BZ_{ki} + \varepsilon_{ki}$, where each element of the factor loading $B \in \mathbb{R}^{p\times d}$, common factor $Z_{ki} \in \mathbb{R}^d$, and random error ε_{ki} are

* Corresponding author.
E-mail address: lanwei@swufe.edu.cn (W. Lan).

http://dx.doi.org/10.1016/j.jmva.2015.05.005
0047-259X/

all independently generated from a standard normal distribution, with $d > 0$ is the finite number of common factors. In this setting, one can verify that $tr(\Sigma^4) = dp^4\{1 + o(1)\}$ and $tr(\Sigma^2) = dp^2\{1 + o(1)\}$. As a result, $tr(\Sigma^4)/tr^2(\Sigma^2) \to 1/d \neq 0$, which makes condition (2.8) in [17] and condition (3.7) in [6] invalid. As a result, how to construct a testing procedure for data of this type becomes a problem of interest.

The aforementioned problem is also empirically motivated. Consider, for example, if our interest is to test Monday effect in China stock market, for assessing stock market efficiency [10]. To this end, we treat trading days (not stocks) as samples, and classify different trading days into two groups, according to whether they are Monday ($k = 1$) or not ($k = 2$). For a given trading day i in the kth group, we use X_{kij} to stand for the jth stock return on this particular trading day in percentage (%). Thus, $X_{ki} = (X_{ki1}, \ldots, X_{kip})^\top \in \mathbb{R}^p$ records all the stock returns on the (k, i)th day. Then, one way to evaluate Monday effect is to test whether $\mu_1 = \mu_2$. It is remarkable that for this problem the existing methods of [3] and [6] cannot be directly applied. The reason is that different stock returns are all heavily correlated with at least one common factor, that is the market index [14,7]. This makes the leading eigenvalue of the covariance matrix Σ extremely large, which violates condition (3.8) in [3] and condition (3.7) in [6] seriously. Thus, this application calls for a new method, which can accommodate such a highly singular eigenvalue structure.

Motivated by the theoretical and practical demand, we aim to develop a two-sample testing procedure for data admitting a low dimensional latent factor structure. Specifically, we investigate the asymptotic distribution of $\|\bar{X}_1 - \bar{X}_2\|$ under a low dimensional latent factor model setup, where $\bar{X}_k$ stands for the sample mean of the kth group. We demonstrate theoretically that such a simple discrepancy measure is asymptotically distributed as a weighted chi-square distribution with a finite degrees of freedom. That leads to a totally different test statistic and inference procedure, as compared with those of Bai and Saranadasa [3] and Chen and Qin [6]. Extensive simulation studies are conducted to demonstrate its finite sample performance. A real data example is also presented for illustration purpose.

The rest of the paper is organized as follows. Section 2 introduces the methodology with both model assumptions and asymptotic theories. Section 3 presents numerical studies based on both simulation and real dataset. The article is concluded with a short discussion in Section 4. All technical details are relegated to Appendix.

2. Methodology

2.1. Model and notation

Let $X_{ki} = (X_{ki1}, \ldots, X_{kip})^\top \in \mathbb{R}^p$ be a p-dimensional vector collected from the ith subject ($1 \leq i \leq n_k$) in the kth ($1 \leq k \leq 2$) group. Write $E(X_{ki}) = \mu_k = (\mu_{k1}, \ldots, \mu_{kp})^\top \in \mathbb{R}^p$ and assume $\mathrm{cov}(X_{ki}) = \Sigma = (\sigma_{j_1 j_2}) \in \mathbb{R}^{p\times p}$. Then, the hypotheses of interest are given by

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2. \tag{2.1}$$

To test the null hypotheses in (2.1), it is natural to consider the following Hotelling's test statistic. That is,

$$T_{\text{Hotelling}} = \left(\frac{n_1 n_2}{n_1 + n_2}\right)\left(\bar{X}_1 - \bar{X}_2\right)^\top \hat{\Sigma}^{-1}\left(\bar{X}_1 - \bar{X}_2\right),$$

where $\hat{\Sigma} = \sum_{k=1}^{2}\sum_{i=1}^{n_k}(X_{ki} - \bar{X}_k)(X_{ki} - \bar{X}_k)^\top/(n_1 + n_2 - 2)$ is the sample covariance matrix. Assuming p is fixed and X_{ki} is normally distributed. Under the null hypothesis of (2.1), $T_{\text{Hotelling}}$ follows a Hotelling's T^2 distribution with $(p, n_1 + n_2 - 2)$ degrees of freedom. Nevertheless, if the data dimension p is considerably larger than the sample size n, the story changes. In that situation, $\hat{\Sigma}$ is not invertible. As a result, the test statistic $T_{\text{Hotelling}}$ is no longer computable.

2.2. Existing methods

To fix the aforementioned problem with $p \gg n$, Bai and Saranadasa [3] proposed the following test statistic

$$T_{BS} = \|\bar{X}_1 - \bar{X}_2\|^2 - \left(\frac{n_1 + n_2}{n_1 n_2}\right) tr(\widehat{\Sigma}).$$

Assuming appropriate regularity conditions and also $p/n \to c$ for some constant $c > 0$, Bai and Saranadasa [3] demonstrates that $T_{BS}/\mathrm{var}^{1/2}(T_{BS})$ follows a standard normal distribution asymptotically. Such a test was further improved by Chen and Qin [6] to the following test statistic

$$T_{CQ} = \frac{\sum_{i\neq j}^{n_1} X_{1i}^\top X_{1j}}{n_1(n_1 - 1)} + \frac{\sum_{i\neq j}^{n_2} X_{2i}^\top X_{2j}}{n_2(n_2 - 1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_{1i}^\top X_{2j}}{n_1 n_2}.$$

Under the null hypothesis of (2.1), Chen and Qin [6] proved that $T_{CQ}/\mathrm{var}^{1/2}(T_{CQ})$ follows a standard normal distribution, even if $p/n \to \infty$.