ARTICLE IN PRESS

Journal of Statistical Planning and Inference ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A feasible high dimensional randomization test for the mean vector

Rui Wang [a], Xingzhong Xu [a,b,*]

[a] School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081, China
[b] Beijing Key Laboratory on MCAACI, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Article history:
Received 2 August 2017
Received in revised form 27 February 2018
Accepted 16 June 2018
Available online xxxx

Keywords:
Asymptotic power function
High dimension
Randomization test
Symmetry assumption

ABSTRACT

The strength of randomization tests is that they are exact tests under certain symmetry assumption for distributions. In this paper, we propose a randomization test for the mean vector in high dimensional setting. We give an implementation of the proposed randomization test procedure, which is computationally feasible. So far, the asymptotic behaviors of randomization tests have only been studied in fixed dimension case. We investigate the asymptotic behavior of the proposed randomization test in high dimensional setting. It turns out that even if the symmetry assumption is violated, the proposed randomization test still has correct level asymptotically. The asymptotic power function is also given. Our theoretical and simulation results show that the proposed test has a wide application scope while still has good power behavior.

1. Introduction

Suppose $X_1, \ldots, X_n$ are p-variate independent and identically distributed (i.i.d.) random vectors with mean vector μ and covariance matrix Σ. In this paper, we consider the problem of testing the hypotheses

$$H_0 : \mu = 0_p \quad \text{versus} \quad H_1 : \mu \neq 0_p. \tag{1}$$

A classical test statistic for hypotheses (1) is Hotelling's T^2, defined as $n\bar{X}^T S^{-1}\bar{X}$, where $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ is the sample mean vector and $S = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$ is the sample covariance matrix. Under normal distribution, Hotelling's T^2 is the likelihood ratio test and enjoys desirable properties in fixed p setting. See, for example, Anderson (2003). However, Hotelling's test cannot be defined when $p > n - 1$ due to the singularity of S. In a seminal paper, Bai and Saranadasa (1996) considered two sample testing problem and proposed a statistic by removing S^{-1} from Hotelling's T^2 statistic. They studied the asymptotic properties of their test statistic when p/n tends to a positive constant. Many subsequent papers generalized the idea of Bai and Saranadasa (1996) to more general models (Srivastava and Du, 2008; Chen and Qin, 2010; Wang et al., 2015). The critical values of existing high dimensional tests mostly rely on the asymptotic normality of the test statistics. We call it asymptotic method. However, if Σ has spiked eigenvalues, the asymptotic normality of the test statistics is not be valid (Katayama et al., 2013). In this case, the asymptotic methods cannot satisfactorily control the level.

The randomization test method is a tool to determine the critical value for a given test statistic. The idea of randomization tests dates back to Fisher (1935). See Romano (1990) for a general construction of randomization test. Its strength is in that the resulting test procedure has exact level under mild condition. There are many papers concerning the theoretical properties of randomization tests for fixed p case. See, for example, Romano (1990), Zhu (2000) and Chung and Romano

* Corresponding author at: School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081, China.
E-mail address: xuxz@bit.edu.cn (X. Xu).

https://doi.org/10.1016/j.jspi.2018.06.003
0378-3758/

(2016). In high dimensional setting, randomization tests are widely used in applied statistics (Subramanian et al., 2005; Efron and Tibshirani, 2007; Ko et al., 2016). However, little is known about the theoretical properties of the randomization test in high dimensional setting.

In this paper, we consider the following randomization method. Suppose $T(X_1, \ldots, X_n)$ is certain test statistic for hypotheses (1). Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. Rademacher variables ($\Pr(\epsilon_i = 1) = \Pr(\epsilon_i = -1) = 1/2$) which are independent of data. The randomization test rejects the null hypothesis when $T(X_1, \ldots, X_n)$ is greater than the $1-\alpha$ quantile of the conditional distribution of $T(\epsilon_1 X_1, \ldots, \epsilon_i X_i, \ldots, \epsilon_n X_n)$ given $X_1, \ldots, X_n$, and accepts the null hypothesis otherwise, where α is the significant level and the $1-\alpha$ quantile of a distribution function $F(\cdot)$ is defined as $\inf\{y : F(y) \geq 1-\alpha\}$. In fixed p setting, it is well known that randomization tests consume much more computing time than the asymptotic method, which historically hampered the use of randomization tests. The goal of this paper is to show that in high dimensional setting, randomization tests can be computationally feasible and have desirable statistic properties. Inspired by the work of Bai and Saranadasa (1996) and Chen and Qin (2010), we propose a randomization test for hypotheses (1). We give an implementation of the proposed randomization test, which is computationally feasible. We also investigate the asymptotic behavior of the test procedure. Our results show that even if the null distribution of X_1 is not symmetric, the randomization test is still asymptotically exact under fairly general assumptions. Hence the test procedure is robust. In particular, the proposed test can be applied to situations where the asymptotic method is not valid. We also derive the asymptotic power function of the proposed test. To the best of our knowledge, this is the first work which gives the asymptotic behavior of randomization tests in high dimensional setting. A simulation study is carried out to examine the numerical performance of the proposed test and compare with the asymptotic method and the bootstrap method. Compared with its competitors, the proposed test reduces the size distortion while still possesses reasonable test power.

The rest of the paper is organized in the following way. In Section 2, we propose a randomization test and give a fast implementation. In Section 3, we investigate the asymptotic behavior of the proposed test. The simulation results are reported in Section 4. The technical proofs are presented in Appendices.

2. Test procedure

Consider testing the hypotheses (1) in high dimensional setting. It is known that Hotelling's T^2 cannot be defined when $p > n-1$. Bai and Saranadasa (1996) removed S^{-1} from Hotelling's T^2 statistic and proposed a statistic which has good power behavior in high dimensional setting. Their idea can also be used for testing hypotheses (1) and the statistic becomes $\bar{X}^T\bar{X}$. The asymptotic properties of the statistic requires p/n tends to a positive constant. Chen and Qin (2010) found that the restriction on p and n can be considerably relaxed by removing the diagonal elements in the statistic of Bai and Saranadasa (1996). For hypotheses (1), their statistic is $\sum_{i\neq j} X_i^T X_j$. Inspired by the statistics of Bai and Saranadasa (1996) and Chen and Qin (2010), we consider the statistic

$$T(X_1, \ldots, X_n) = \sum_{j<i} X_i^T X_j. \tag{2}$$

Bai and Saranadasa (1996) and Chen and Qin (2010) used asymptotic method to determine the critical value of $T(X_1, \ldots, X_n)$. However, as will be shown in Section 3, the critical value determined by the asymptotic method is not valid in some important cases. Hence it is desirable to find a better critical value.

The bootstrap and the randomization method are two popular methods to determine the critical value. However, the bootstrap method may not be a good choice for our problem. To see this, consider the following basic bootstrap procedure. Let $X_1^*, \ldots, X_n^*$ be a bootstrap sample which is randomly drawn from $\{X_1 - \bar{X}, \ldots, X_n - \bar{X}\}$ with replacement. Denote by

$$\mathcal{L}(T(X_1^*, \ldots, X_n^*)|X_1, \ldots, X_n), \tag{3}$$

the distribution of $T(X_1^*, \ldots, X_n^*)$ conditioning on $X_1, \ldots, X_n$. Then the bootstrap critical value for the statistic $T(X_1, \ldots, X_n)$ is defined to be the $1-\alpha$ quantile of the bootstrap distribution (3). If this bootstrap method worked well, the bootstrap distribution (3) should mimic the null distribution of $T(X_1, \ldots, X_n)$. Then one may expect that the first two moments of (3) are close to the first two moments of $T(X_1, \ldots, X_n)$. Under the null hypothesis,

$$\mathrm{E}\, T(X_1, \ldots, X_n) = 0, \quad \mathrm{Var}(T(X_1, \ldots, X_n)) = \frac{n(n-1)}{2}\,\mathrm{tr}(\Sigma^2).$$

Also, it is straightforward to show that

$$\mathrm{E}(T(X_1^*, \ldots, X_n^*)|X_1, \ldots, X_n) = 0, \quad \mathrm{Var}(T(X_1^*, \ldots, X_n^*)|X_1, \ldots, X_n) = \frac{n(n-1)}{2}\Big(\frac{n-1}{n}\Big)^2\,\mathrm{tr}(S^2).$$

However, as pointed out by Bai and Saranadasa (1996), $\big((n-1)/n\big)^2\,\mathrm{tr}(S^2)$ is not even a ratio consistent estimator of $\mathrm{tr}(\Sigma^2)$ in high dimensional setting.

Now we consider the following randomization method. Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. Rademacher variables which are independent of data. Denote by

$$\mathcal{L}\big(T(\epsilon_1 X_1, \ldots, \epsilon_i X_i, \ldots, \epsilon_n X_n)|X_1, \ldots, X_n\big) \tag{4}$$

Please cite this article in press as: Wang R., Xu X., A feasible high dimensional randomization test for the mean vector. J. Statist. Plann. Inference (2018), https://doi.org/10.1016/j.jspi.2018.06.003.