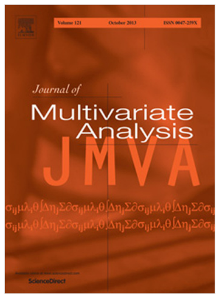
Accepted Manuscript



On two-sample mean tests under spiked covariances

Rui Wang, Xingzhong Xu

PII: S0047-259X(17)30416-5
DOI: https://doi.org/10.1016/j.jmva.2018.05.004
Reference: YJMVA 4363

To appear in: *Journal of Multivariate Analysis*

Received date : 11 July 2017

Please cite this article as: R. Wang, X. Xu, On two-sample mean tests under spiked covariances, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.05.004

On two-sample mean tests under spiked covariances

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

Abstract

This paper considers testing the means of two p-variate normal samples in high dimensional settings. We show that under the null hypothesis, a necessary and sufficient condition for the asymptotic normality of the test statistic of Chen and Qin (2010) is that the eigenvalues of the covariance matrix are concentrated around their average. However, this condition is not satisfied when the variables are strongly correlated. To characterize the correlations between variables, we adopt a spiked covariance model. Under the spiked covariance model, we derive the asymptotic distribution of the test statistic of Chen and Qin (2010) and correct its critical value. The recently proposed random projection test procedures suggest that the power of tests may be boosted using the projected data. By maximizing an average signal to noise ratio, we find that the optimal projection subspace is the orthogonal complement of the principal subspace. We propose a new test procedure through the projection onto the estimated orthogonal complement of the principal subspace. The asymptotic normality of the new test statistic is proved and the asymptotic power function of the test is given. Theoretical and simulation results show that the new test outperforms the competing tests substantially under the spiked covariance model.

Keywords: high dimension, mean test, principal subspace, spiked covariance model

1. Introduction

Suppose that for $k \in \{1, 2\}$, $X_{k,1}, \ldots, X_{k,n_k}$ are independent identically distributed (iid) p-dimensional normal random vectors with unknown mean vector μ_k and covariance matrix $\mathbf{\Sigma}$. We consider the hypothesis testing problem

$$\mathcal{H}_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad \mathcal{H}_1 : \mu_1 \neq \mu_2. \tag{1}$$

In this paper, the high-dimensional setting is adopted, i.e., the dimension p varies as n increases, where $n = n_1 + n_2 - 2$. Testing hypotheses (1) is important in many fields, including biology, finance and economics.

A classical test statistic for hypotheses (1) is Hotelling's T^2 test statistic $(\bar{X}_1 - \bar{X}_2)^\top \mathbf{S}^{-1}(\bar{X}_1 - \bar{X}_2)$, where $\bar{X}_1$ and $\bar{X}_2$ are the two sample means and

$$\mathbf{S} = n^{-1} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (X_{k,i} - \bar{X}_k)(X_{k,i} - \bar{X}_k)^\top$$

is the pooled sample covariance matrix. However, Hotelling's test statistic is not defined when $p > n$. Moreover, Bai and Saranadasa [4] showed that even if $p \leq n$, Hotelling's test suffers from low power when p is comparable to n. Perhaps, the main reason for the low power of Hotelling's test is that $\mathbf{S}$ is a poor estimator of $\mathbf{\Sigma}$ when p is large compared with n; see [9] and the references therein. For testing hypotheses (1) in high-dimensional settings, many test statistics are based on the estimation of $(\mu_1 - \mu_2)^\top \mathbf{A}(\mu_1 - \mu_2)$ for a positive definite matrix $\mathbf{A}$. Bai and Saranadasa [4] proposed a test based on

$$T_{BS} = \|\bar{X}_1 - \bar{X}_2\|^2 - (1/n_1 + 1/n_2)\,\mathrm{tr}\,\mathbf{S},$$

*Corresponding author
Email address: xuxz@bit.edu.cn (Xingzhong Xu)