非正規母集団からの高次元データの平均の検定

山田　隆行　**リスク解析センター　特任助教**

Introduction

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be a random sample drown from a population. We shall assume the following model:

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}, \quad \boldsymbol{z} \sim F. \tag{1}$$

Assume that $E[\boldsymbol{z}] = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{z}) = \boldsymbol{I}_p$. The interest for the model (**??**) is to test

$$H_0 : \boldsymbol{\mu} = \boldsymbol{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{0}.$$

Hotelling's T^2 test is valid for the case in which $n > p$. When $p > N$, S becomes singular, so T^2 cannot be defined. In this case, Bai and Saranadasa [1] have proposed other non-exact tests for two sample problem. Srivastava and Du [5] proposed other test based on the criterion $\bar{\boldsymbol{x}}'\boldsymbol{D}_S^{-1}\bar{\boldsymbol{x}}$ with $\boldsymbol{D}_S = \mathrm{diag}(s_{11}, \ldots, s_{pp})$ for $S = (s_{ij})$. These results were firstly built under the assumption that F is p-dimensional normal distribution. Generalization for non-normality have been studied. Bai and Saranadasa [1] have showed that their test is robust under the condition $\mathrm{C_{BS}}$ that $E[z_i^4] = 3 + \gamma$ for $\boldsymbol{z} = (z_1, \ldots, z_p)'$ and $E[\prod_{i=1}^p z_i^{\nu_i}] = 0$ (and 1) when there is at least one $\nu_i = 1$ (there are two ν_i's equal to 2, correspondingly), whenever $\nu_1 + \cdots + \nu_p = 4$. Srivastava [6] have shown that Srivastava and Du [5]'s test is robust under the condition $\mathrm{C_S}$ that $z_1, \ldots, z_p$ are iid, and $E[z_i^4] = 3 + \gamma$. For two sample problem of mean vector, Chen and Qin [2] proposed a test base on Bai and Saranadasa [1]'s criterion. They showed asymptotic normality of Bai and Saranadasa [1]'s criterion under the condition $\mathrm{C_{CQ}}$ that $E[z_i^4] = 3 + \gamma$ and $E[\prod_{i=1}^p z_{\ell_i}^{\nu_i}] = \prod_{i=1}^q E[z_{\ell_i}^{\nu_i}]$ for a positive integer q such that $\sum_{i=1}^q \nu_i \leq 8$.

In this paper, we treat Bai and Saranadasa [1]'s testing statistic reduced to the one sample problem, which the testing statistic is defined as

$$T_{\mathrm{BS}} = N\bar{\boldsymbol{x}}'\bar{\boldsymbol{x}} - \mathrm{tr}\,\boldsymbol{S}.$$

We will derive asymptotic null distribution of $T_{\mathrm{BS}}^* = \{n/(Np)\}^{1/2}T_{\mathrm{BS}}$ under the asymptotic framework A1 and A2:

$$\mathrm{A1} : p = O(N) \text{ as } N \to \infty, \quad \mathrm{A2} : N = O(p) \text{ as } p \to \infty.$$

In order to derive asymptotic null distribution under A1, we assume the following assumptions:

$$E[(\boldsymbol{z}'\boldsymbol{\Sigma}\boldsymbol{y})^4] = o(p^4); \tag{2}$$
$$E[(\boldsymbol{z}'\boldsymbol{\Sigma}^2\boldsymbol{z})^2] = O(p^2); \tag{3}$$
$$E[(\boldsymbol{z}'\boldsymbol{\Sigma}\boldsymbol{y})^2\boldsymbol{z}'\boldsymbol{\Sigma}^2\boldsymbol{y}] = O(p^5); \tag{4}$$
$$a_i = (1/p)\,\mathrm{tr}\,\boldsymbol{\Sigma}^i = O(1), \quad i = 1, \ldots, 4, \tag{5}$$

where $\boldsymbol{y}$ and $\boldsymbol{z}$ are i.i.d. as F. These assumptions imply $\mathrm{C_{BS}}$, $\mathrm{C_S}$ and $\mathrm{C_{CQ}}$, so our assumptions are milder than them. Besides, for A2, F is assumed as spherical distribution such that

$$E[\boldsymbol{z}_2|\boldsymbol{z}_1] = \boldsymbol{0} \tag{6}$$

for any partition $\boldsymbol{z}' = (\boldsymbol{z}_1' : \boldsymbol{z}_2')'$. In addition, we assume

$$\gamma_4 = \sup_{1 \leq i \leq p} E[|z_i|^4] < \infty, \tag{7}$$

for $\boldsymbol{z} = (z_1, \ldots, z_p)'$.

Asymptotic distributions

Proposition 1. *Under the asymptotic framework* A1 *and assumptions* (**??**)*,* (**??**) *and* (**??**)*,* T_{BS}^* *converges in distribution to the normal distribution with the mean 0 and the variance* $2\lim_{\mathrm{A1}}(1/p)\,\mathrm{tr}\,\boldsymbol{\Sigma}^2$.

Proposition 2. *Assume that* F *is spherical distribution, and assume conditions* (**??**)*,* (**??**) *and* (**??**)*. Under the asymptotic framework* A2 T_{BS}^* *converges in distribution to the normal distribution with the mean 0 and the variance* $2\lim_{\mathrm{A2}}(1/p)\,\mathrm{tr}\,\boldsymbol{\Sigma}^2$.

Himeno and Yamada [3] proposed unbiased estimator $\widetilde{a}_2$ of a_2 under non-normality, which is given by $\widetilde{a}_2 = \frac{N-1}{N(N-2)(N-3)p}\Big\{(N-1)(N-2)\,\mathrm{tr}\,\boldsymbol{S}^2 + (\mathrm{tr}\,\boldsymbol{S})^2 - \frac{N}{N-1}\sum_{i=1}^N((\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}))^2\Big\}$. Consistency is proved under asymptotic framework A1 and the assumptions (**??**), (**??**), (**??**) and (**??**), and so $T = T_{\mathrm{BS}}^*/(2\widetilde{a}_2)^{1/2} \xrightarrow{d} N(0,1)$. Besides, under the assumption that the population distribution F is normal, $T_N = T_{\mathrm{BS}}^*/(2\hat{a}_2)^{1/2} \xrightarrow{d} N(0,1)$ under A1, where $\hat{a}_2$ is the unbiased and the consistent estimator of a_2, which is given in Srivastava [4], defined as $\hat{a}_2 = \frac{n^2}{(n-1)(n+2)}\frac{1}{p}\Big\{\mathrm{tr}\,\boldsymbol{S}^2 - \frac{1}{n}(\mathrm{tr}\,\boldsymbol{S})^2\Big\}$.

In order to check the performance of the asymptotic approximations we did small scale simulation. Generate the data based on the model (**??**). We consider 3 cases for the population distribution F; Case1: F is multivariate normal distribution with the mean $\boldsymbol{0}$ and the covariance matrix $\boldsymbol{I}_p$; Case2: F is scaled multivariate T distribution with 5 degrees of freedom, the mean $\boldsymbol{0}$ and the covariance matrix $\boldsymbol{I}_p$; Case3: For $c_1, \ldots, c_p$ are i.i.d. χ_1^2, chi-squared distribution with 1 degrees of freedom, $z_i = (c_i - 1)/2^{1/2}$, $i = 1, \ldots, p$. For the structure of the dispersion matrix $\boldsymbol{\Sigma}$, we selected $\boldsymbol{\Sigma} = (0.2^{|i-j|})$. We reject the null hypothesis H_0 if T (T_N) is larger than upper α percentile point of the standard normal distribution.

References

[1] Z. Bai, H. Saranadasa, Effect of high dimension: an example of a two sample problem, Statist. Sinica 6 (1996) 311–329.

[2] X.C. Chen, Y.L. Qin, A two sample test for high dimensional data with applications to gene-set testing, Ann. Statist. 28 (2010) 808–835.

[3] T. Himeno, T. Yamada, Estimation for some functions of covariance matrix in high dimension under non-normality, under preparation.

[4] M.S. Srivastava, Some tests concerning the covariance matrix in high dimensional data, J. Japan Statist. Soc. 35 (2005) 251–272.

[5] M.S. Srivastava, D. Meng, A test for the mean vector with fewer observations than the dimension, J. Multivariate Anal. 99 (2008) 386–402.

[6] M.S. Srivastava, A test for the mean vector with fewer observations than the dimension under non-normality, J. Multivariate Anal. 100 (2009) 518–532.

Table 1: Actual error probabilities of the first kind when the nominal is 0.05 based on 10,000 repetition

n	p	Case1		Case2		Case3		n	p	Case1		Case2		Case3	
		T	T_N	T	T_N	T	T_N			T	T_N	T	T_N	T	T_N
20	20	0.067	0.067	0.070	0.037	0.059	0.040	100	20	0.063	0.064	0.066	0.050	0.057	0.051
20	60	0.065	0.065	0.063	0.012	0.056	0.034	100	60	0.060	0.060	0.061	0.029	0.059	0.052
20	100	0.058	0.060	0.064	0.005	0.060	0.034	100	100	0.062	0.062	0.058	0.019	0.054	0.046
60	20	0.068	0.067	0.064	0.045	0.064	0.052								
60	60	0.061	0.062	0.059	0.024	0.054	0.044								
60	100	0.059	0.059	0.058	0.014	0.056	0.045								