On error bounds for high-dimensional asymptotic distribution of L_2-type test statistic

Takahiro Nishiyama[a], Masashi Hyodo[b] and Tatjana Pavlenko[c]

[a] Department of Business Administration, Senshu University
[b] Department of Mathematical Sciences, Osaka Prefecture University
[c] Department of Mathematics, KTH Royal Institute of Technology

1. Introduction

This paper is concerned with the canonical testing problem in modern statistical inference, namely the two-sample test for equality of means of independent multivariate populations with very large dimensions. Precisely, let $\boldsymbol{x}_{gk} = (x_{g1k}, \ldots, x_{gpk})^\top$, $k \in \{1, \ldots, n_g\}$, be n_g iid random vectors with $\boldsymbol{x}_{gk} \sim \mathcal{N}_p(\boldsymbol{\mu}_g, \Sigma_g)$, $g \in \{1, 2\}$, where $\boldsymbol{\mu}_g \in \mathbb{R}^p$, $\Sigma_g \in \mathbb{R}^{p \times p}_{>0}$, represent the usual parameters.

We are interested in testing the hypothesis $\mathcal{H} : \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = 0$, vs. $\mathcal{A} : \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| > 0$ where $\| \cdot \|$ denotes L_2-norm. A common feature of the modern high-dimensional data is that the dimension of $\boldsymbol{x}_{gk}$ exceeds the number of sampled observations, the so-called "*large-p-small-n*" problem which breaks down applicability of the classical Hotelling's T^2-test. The best-known test procedure which accommodates high-dimensional data and allows for $\Sigma_1 \neq \Sigma_2$ is the L_2-type test statistics introduced by Chen and Qin (2010), (hereafter called for Ch-Q test) who tested $\mathcal{H}$ based on the unbiased estimator of the squared L_2-norm of the difference $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ written as

$$T_n = \|\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2\|^2 - \sum_{g=1}^{2} \mathrm{tr}(S_g)/n_g,$$

where $n = n_1 + n_2$, $\overline{\boldsymbol{x}}_g$ and S_g are the sample mean and sample covariance matrix of gth population and $\mathrm{E}(T_n) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$. To the unbiasedness property of T_n, Chen and Qin (2010) quantified its variance

$$\sigma_n^2 = \mathrm{var}(T_n) = \sum_{g=1}^{2} \frac{2\mathrm{tr}(\Sigma_g^2)}{n_g(n_g - 1)} + \frac{4\mathrm{tr}(\Sigma_1 \Sigma_2)}{n_1 n_2},$$

under $\mathcal{H}$, and show that the distribution of T_n admits a normal limit after appropriate rescaling.

The normal approximation, while representing the conventional class of limiting laws is often too loose for large-p-small-n settings or fails to capture the tail behavior of the resulting distribution, as it occurs when dealing with the hypothesis testing using the L_2-type test statistics, see e.g. results of empirical study of Nishiyama et al. (2013) and Hyodo et al. (2014). In what follows, we first present a theoretical analysis of this problem which requires understanding of the rate of convergence of T_n to its normal limit, and then introduce two new approximations which are more accurate in the asymptotic regime where both n and p tend to infinity.

2. Main results

2.1. Our approximations

To provide the intuition behind the approximations which we propose, let $\psi_{\widetilde{T}_n}(t)$ denote the characteristic functions of $\widetilde{T}_n = T_n/\sigma_n$ and consider expansion of the cumulant generating function

$$\ln(\psi_{\widetilde{T}_n}(t)) = \sum_{j=1}^{\infty} \kappa_j(\widetilde{T}_n)\frac{(it)^j}{j!}$$

where the constants $\kappa_j(\widetilde{T}_n)$ are known as *cumulants* of $\widetilde{T}_n$ (see e.g. Muirhead (1982)). Also, without loss of generality, let $\lambda_r(\Lambda)$ be the r-th largest eigenvalue of the matrix $\Lambda = (n/n_1)\Sigma_1 + (n/n_2)\Sigma_2$ and let $\Delta = \lambda_1(\Lambda)^2/\mathrm{tr}(\Lambda^2)$.

Our primary tool in constructing the approximations is the so-called *cumulant matching technique* which requires existence of all cumulants of $\widetilde{T}_n$ up the third order. After some cumbersome but straightforward manipulations, we obtain the first three cumulants of $\widetilde{T}_n$ as $\kappa_1 = 0,\ \kappa_2 = 1,\ \kappa_3 = 6a$, where $a = 4b/(3\sigma_n^3)$ and

$$b = \sum_{g=1}^{2} \frac{(n_g-2)\mathrm{tr}(\Sigma_g^3)}{n_g^2(n_g-1)^2} + \frac{3\mathrm{tr}(\Sigma_1^2\Sigma_2)}{n_1^2 n_2} + \frac{3\mathrm{tr}(\Sigma_1\Sigma_2^2)}{n_1 n_2^2}.$$

Further, by approximating $\ln\psi_{\widetilde{T}_n}(t)$ up to the third order cumulant, we obtain $\ln\psi_{\widetilde{T}_n}(t) \approx -t^2/2 + a(it)^3$, which in turn provides the approximation of the characteristic function of $\widetilde{T}_n$ as

$$\psi_{\widetilde{T}_n}(t) \approx e^{-t^2/2} + a(it)^3 e^{-t^2/2} \tag{1}$$

By this result, the normal approximation of $\widetilde{T}_n$ is immediately achieved by inverting the first term on the right side of (1), that is $F_{\widetilde{T}_n}(t) = \Phi(t) + o(1)$, as $\Delta = 0$.

On the other hand, inverting the right side of (1) term by term provides the approximating distribution of $\widetilde{T}_n$ of the form $F_2(x) = \Phi(x) + a(1-x^2)\phi(x)$, where $\phi(x)$ denotes the density of $\Phi(x)$. This latter representation of $F_2(x)$ is also known as the higher order Edgeworth expansion of the distribution of $\widetilde{T}_n$.

Another avenue of research delivers a χ^2-approximation of $\widetilde{T}_n$; this look on the problem is motivated by the work of Buckley and Eagleson (1988) and Zhang (2005), who studied approximation of noncentral χ^2-type mixtures by using cumulant matching to single noncentral χ^2 random variable. More precisely, let V_d denote the χ^2-distributed random variable with d degrees of freedom and let $G_d(\cdot)$ denote the corresponding cumulative distribution function. The formulation of the approximation rests on the fact that, by setting $d = 2/(9a^2) = \sigma_n^6/(8b^2)$, the first three cumulants of the two random variables, $\widetilde{T}_n$ and $C_d = (V_d - d)/\sqrt{2d}$, are exactly the same. Using this fact, we can approximate $F_{\widetilde{T}_n}(x)$ by $F_3(x) = G_d(\sqrt{2d}x + d)$.

2.2. Error bounds for approximations

In the following we assess the theoretical properties of our approximations. To establish the rate of convergence, we first obtain the explicit bounds for the Kolmogorov distance between the distribution of $\widetilde{T}_n$ and its approximations $\Phi(x)$, $F_2(x)$, and $F_3(x)$. Now, the following theorem, which is the main theoretical result of this paper, states the weak convergence for the distribution of $\widetilde{T}_n$ to its approximate limits and provides the corresponding error bounds in terms of Δ.

Theorem 1. *The distribution of $\widetilde{T}_n$ satisfies the following properties under $\mathcal{H}$:*

(i) For any $n_1, n_2, p, \Sigma_1, \Sigma_2$ such that $\Delta < 1/8$,

$$\sup_{x\in\mathbb{R}} |F_{\widetilde{T}_n}(x) - F_2(x)| \leq \frac{3\Delta}{2\pi\omega}\left\{2 + \frac{8!^{1/4}}{8(1-8\Delta)^2}\right\} + \frac{8\Delta(2+\omega)}{9\pi\omega^2}.$$

(ii) For any $n_1, n_2, p, \Sigma_1, \Sigma_2$ such that $\Delta < 1/8$,

$$\sup_{x\in\mathbb{R}} |F_{\widetilde{T}_n}(x) - F_3(x)| \leq \frac{3\Delta}{2\pi\omega}\left\{2 + \frac{8!^{1/4}}{8(1-8\Delta)^2}\right\} + \frac{\{10 + 3(1-8\Delta)^{-2}\}\Delta}{2\pi}.$$

(iii) For any $n_1, n_2, p, \Sigma_1, \Sigma_2$ such that $\Delta < 1/6$,

$$\sup_{x\in\mathbb{R}} |F_{\widetilde{T}_n}(x) - \Phi(x)| \leq \frac{2\Delta^{1/2}}{2\pi\omega}\left\{3\sqrt{\pi} + \frac{6!^{1/4}\sqrt{2}}{3(1-6\Delta)^{3/2}}\right\},$$

where ω is omega constant which is a mathematical constant defined by $\omega\exp(\omega) = 1$, and $\omega \approx 0.56714$.

The following corollary is prepared for evaluation of the convergence rate of the proposed approximations in a general high-dimensional asymptotic regime.

Corollary 1. *Assume that for $g \in \{1, 2\}$ the eigenvalues of Σ_g admit the representation*

$$\lambda_r(\Sigma_g) = a_r(g)p^{\beta_{r(g)}},\ r \in \{1, \ldots, t_g\} \text{ and } \lambda_r(\Sigma_g) = c_{r(g)},\ r \in \{t_g + 1, \ldots, p\},$$

where $a_{r(g)}$, $c_{r(g)}$ and $\beta_{r(g)}$ are positive and fixed constants and t_g is fixed positive integer. Let further $\beta_{(1)} = \max\{\beta_{1(1)}, \beta_{2(1)}\} < 1/2$. Then, for any n such that either both n_1 and n_2 are fixed or $n_1/n_2 \to \gamma \in (0, \infty)$ as $n_1, n_2 \to \infty$, it holds that

(i) $\sup_{x\in\mathbb{R}} |F_{\widetilde{T}_n}(x) - F_2(x)| = O(p^{2\beta_{(1)}-1})$,

(ii) $\sup_{x\in\mathbb{R}} |F_{\widetilde{T}_n}(x) - F_3(x)| = O(p^{2\beta_{(1)}-1})$,

(iii) $\sup_{x\in\mathbb{R}} |F_{\widetilde{T}_n}(x) - \Phi(x)| = O(p^{\beta_{(1)}-1/2})$.

We now return to the Ch-Q statistic presented in Section 1 and exploit our proposed approximations to furnish a test of significance of $\mathcal{H} : \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = 0$, vs. $\mathcal{A} : \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| > 0$ at the level $0 < \alpha < 1$. By using the result (2.2) in Hall (1983) along with Corollary 1, we state the following.

Corollary 2. *Assume the same conditions as for Corollary 1. Then, for any point x in the compact subset of* $\mathbb{R}$, $F_{\widetilde{T}_n}\{x + a(x^2 - 1)\} = \Phi(x) + o(p^{\beta_{(1)} - 1/2})$.

Direct applications of Corollaries 1 and 2 provide the critical values for the corresponding approximate tests which can be calibrated based on χ^2 and normal quantiles, respectively.

3. Application to two sample test

The Ch-Q test with the proposed approximations rests on the adjusted α-quantiles, $q_2(\alpha) = z_\alpha + 4b(z_\alpha^2 - 1)/(3\sigma_n^3)$ and $q_3(\alpha) = (\chi_d^2(\alpha) - d)/(2d)^{1/2}$, both depending on unknown quantities, σ_n, b, and d. Now, for practical applications of the proposed approximations, we use $\widehat{q}_2(\alpha)$ and $\widehat{q}_3(\alpha)$ which are constructed by plugging the estimators of $\mathrm{tr}(\Sigma_g^2)$, $\mathrm{tr}(\Sigma_g^3)$, $\mathrm{tr}(\Sigma_g\Sigma_h)$ and $\mathrm{tr}(\Sigma_g^2\Sigma_h)$ into σ_n, b, and d. The obtained estimated quantiles serve as critical values for the corresponding approximate level α tests are obtained as follows:

$$\text{reject } \mathcal{H} \Longleftrightarrow T_n > \widehat{\sigma}_n \widehat{q}_2(\alpha),$$
$$\text{reject } \mathcal{H} \Longleftrightarrow T_n > \widehat{\sigma}_n \widehat{q}_3(\alpha).$$

The following theorem provides the ratio consistency of the estimated quantiles of T_n in a general high-dimensional asymptotic regime.

Theorem 2. *For fixed $\alpha \in (0, 1)$, under the same conditions as for Corollary 1,*

$$\frac{\widehat{\sigma}_n \widehat{q}_2(\alpha)}{\sigma_n q_2(\alpha)} = 1 + O_p(n_1^{-1/2}) + O_p(n_2^{-1/2}), \quad \frac{\widehat{\sigma}_n \widehat{q}_3(\alpha)}{\sigma_n q_3(\alpha)} = 1 + O_p(n_1^{-1/2}) + O_p(n_2^{-1/2}).$$

4. Numerical study

We evaluate empirical quantiles of $\widetilde{T}_n$ to assess the accuracy of the proposed approximations $q_2(\alpha)$ and $q_3(\alpha)$ for some selected parameters. Also, we evaluate the sizes of the proposed tests, and for comparison exploit the existing Ch-Q procedure.

References

[1] Buckley, M.J., Eagleson, G.K., 1988. An approximation to the distribution of quadratic forms in normal random variables. *Austral. J. Statist.*, **30**, 150–159.

[2] Chen, S.X., Qin, Y.L., 2010. A two-sample test for high dimensional data with applications to gene-set testing. *Ann. Statist.*, **38**, 808–835.

[3] Hall, P., 1983. Inverting an edgeworth expansion. *Ann. Statist.*, **11**, 569–576.

[4] Hyodo, M., Takahashi, S., Nishiyama, T., 2014. Multiple comparisons among mean vectors when the dimension is larger than the total sample size. *Commun. Statist. Simul. Comput.*, **43**, 2283–2306.

[5] Muirhead, R.J., 1982. *Aspects of multivariate statistical theory.* Wiley, New York.

[6] Nishiyama, T., Hyodo, M., Seo, T., Pavlenko, T., 2013. Testing linear hypotheses of mean vectors for high-dimension data with unequal covariance matrices. *J. Stat. Plan. Inference.*, **143**, 1898–1911.

[7] Zhang, J.T., 2005. Approximate and asymptotic distributions of chi-squared-type mixtures with applications. *J. Amer. Statist. Assoc.*, **100**, 273–285.