高次元データにおける幾つかの検定統計量の漸近分布について

藤本 翔太 [1], 狩野 裕 [1], Muni S. Srivastava[2]
[1] 大阪大学基礎工学研究科
[2]Department of Statistics, University of Toronto

正規分布に基づく多変量 1 標本問題を考える．すなわち，$\boldsymbol{X}_p^{(1)}, \cdots, \boldsymbol{X}_p^{(n)}$ が独立に同一の多変量正規分布 $N_p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ に従うとする．ここで，$\boldsymbol{\Sigma}_p$ は任意の p に対して正定値行列であることを仮定する．いま，次の検定問題を考える：$H_0 : \boldsymbol{\mu}_p = \boldsymbol{0}$ versus $H_1 : \boldsymbol{\mu}_p \neq \boldsymbol{0}$. この検定問題に対する伝統的な方法は，Hotelling の T^2 統計量を用いる方法であるが，標本サイズ n が変数の次元 p よりも小さな場合は，標本共分散行列が特異になり，定義されない．Dempster (1958, 1960) は T^2 統計量における標本共分散行列 $\boldsymbol{S}_{n,p}$ をそのトレース $\mathrm{tr}\boldsymbol{S}_{n,p}$ で置き換えた統計量を提案した：$T_D^2 = n\bar{\boldsymbol{X}}_{n,p}^T\bar{\boldsymbol{X}}_{n,p}/\mathrm{tr}\boldsymbol{S}_{n,p}$. ここに，$\bar{\boldsymbol{X}}_{n,p}$ は標本平均である．Bai and Saranadasa (1996), Fujikoshi (2004), Srivastava (2007) は T_D^2 を用いた検定を行うために，(n, p) が共に大きくなるという高次元漸近理論の枠組みでの漸近分布を導出している．Bai and Saranadasa (1996) は T_D^2 の漸近正規性を条件

$$\max_{1\leq i\leq p} \lambda_i = o\left(\sqrt{\mathrm{tr}\boldsymbol{\Sigma}_p^2}\right) \tag{1}$$

の下で示している．ここに，λ_i $(i = 1, 2, \cdots, p)$ は $\boldsymbol{\Sigma}_p$ の固有値である．また，Bai and Saranadasa (1996) では別の統計量 $T_{BS}^2 = n\bar{\boldsymbol{X}}_{n,p}^T\bar{\boldsymbol{X}}_{n,p} - \mathrm{tr}\boldsymbol{S}_{n,p}$. も提案しており，その漸近正規性も同じ条件の下で示している．近年， Fujikoshi (2004) や Srivastava (2007) によって T_D^2 と T_{BS}^2 の漸近正規性が

$$0 < \lim_{p\to\infty} \frac{\mathrm{tr}\boldsymbol{\Sigma}_p^i}{p} < \infty \quad (i = 1, 2, 3, 4). \tag{2}$$

の下でも導出されている．なお，T_D^2 と T_{BS}^2 はデータ行列の直交行列による変換 $\boldsymbol{X}_p^{(i)} \to c\Gamma\boldsymbol{X}_p^{(i)}$ に関して不変な統計量である．ここに，$c > 0$ であり，$\Gamma^T\Gamma = I_p$ である．また，I_p は $p \times p$ の単位行列を表す．しかしながら，これらの統計量はデータの単位変換 $\boldsymbol{X}_p^{(i)} \to D\boldsymbol{X}_p^{(i)}$ に関して不変ではない．ここに，$D = \mathrm{diag}(d_1, \cdots d_p), d_i > 0$ である．そこで，Srivastava and Du (2008) はデータの単位変換に関して不変な統計量 $T_S^2 = n\bar{\boldsymbol{X}}_{n,p}^T(\mathrm{diag}\boldsymbol{S}_{n,p})^{-1}\bar{\boldsymbol{X}}_{n,p}$ を提案している．ここに，$\mathrm{diag}\boldsymbol{A}$ は行列 $\boldsymbol{A}$ の対角要素だけを残し，非対角要素を全て 0 とした行列である．さらに，高次元漸近理論の枠組みにおける漸近正規性を次の条件の下で示している：

$$0 < \lim_{p\to\infty} \frac{\mathrm{tr}\mathcal{R}_p^i}{p} < \infty \quad (i = 1, 2, 3, 4), \tag{3}$$

ここに，$\mathcal{R}_p$ は $\boldsymbol{\Sigma}_p$ の相関行列である．

条件 (1) または (2) が満たされるならば，$p \to \infty$ のとき $r := (\mathrm{tr}\boldsymbol{\Sigma}_p)^2/\mathrm{tr}\boldsymbol{\Sigma}_p^2 \to \infty$ である．一方で，Cauchy-schwarz の不等式から $r \leq p$ であり，等号成立はある定数 $c > 0$ に対して $\boldsymbol{\Sigma}_p = cI_p$ であるときかつそのときに限られることがわかる．したがって，これらの条件は $\boldsymbol{\Sigma}_p$ が単位行列の定数倍に近い場合でしか成立ちそうにない．同じことが条件 (3) に対してもいえる．

具体的に例を見てみると，条件 (1), (2), (3) は $\boldsymbol{\Sigma}_p$ が例えば次のような場合は成立つ：

(a) Sphericity Model : $\boldsymbol{\Sigma}_p = aI_p$. ここに，$a > 0$.
(b) Autoregressive Model : $\boldsymbol{\Sigma}_p = a(\rho^{|i-j|})$. ここに，$a > 0$, $0 < \rho < 1$.

これらの条件が Autoregressive Model の場合に満たされることは，Szegö の定理 (e.g., Grenander and Szegö 1958) を使って示される．一方，条件 (1), (2), (3) は次の場合に満たされない．

(c) Compound Symmetry Structure : $\mathbf{\Sigma}_p = a(1-\rho)I_p + a\rho\mathbf{1}_p\mathbf{1}_p^T$. ここに，$a > 0,\ 0 < \rho < 1$ であり，$\mathbf{1}_p$ は要素すべてが 1 の p 次元列ベクトルを表す.

条件が満たされないことは，Compound Symmetry Structure の固有値が $\lambda_1 = a(1-\rho+p\rho),\ \lambda_2 = \cdots = \lambda_p = a(1-\rho)$ であることから容易に確認される. これは次のモデルの特別な場合である：

(d) Spiked Model (Johnstone, 2001) : $\lambda_i = c_i p^{\alpha_i}\ \ (i = 1, 2, \cdots, \ell),\ \ \lambda_j = c_j\ \ (j = \ell+1, \cdots, p)$. ここに，$\lambda_1 \geq \cdots \geq \lambda_p$ は $\mathbf{\Sigma}_p$ (または $\boldsymbol{\mathcal{R}}_p$) の固有値であり，$c_1 \geq \cdots \geq c_p > 0,\ \alpha_i \geq 0,\ m < p$ はすべて p に依存しないものする.

Spiked Model において，条件 (1) は $\alpha_1 < 1/2$ のとき満たされ，条件 (2) は $\alpha_1 \leq 1/4$ のとき満たされる. しかし，どちらの条件も $\alpha_1 \geq 1/2$ のときは満たされない. 以上の考察から，先行研究の条件は，$\mathbf{\Sigma}_p$(または $\boldsymbol{\mathcal{R}}_p$) の幾つかの固有値が大きく突出した場合に満たされないことがわかる.

次で提案する $\mathbf{\Sigma}_p$(または $\boldsymbol{\mathcal{R}}_p$) に関する新しい条件は，上記のすべての例を含んでいる：$\delta_1 \geq \delta_2 \geq \delta_k$ $(k = 3, 4, \cdots)$ であるようなある定数 $\delta_i > 0$ が存在して，

$$0 < \lim_{p\to\infty} \mathrm{tr}\left(\frac{\mathbf{\Sigma}_p}{p^{\delta_i}}\right)^i < \infty\ \ (i = 1, 2, \cdots). \tag{4}$$

本研究の目的は，検定統計量 $T_D^2,\ T_{BS}^2, T_S^2$ の高次元漸近理論の枠組みにおける漸近分布を条件 (4) の下に拡張し，より一般的な条件の下でも帰無仮説 H_0 を検定できるようにすることである. 本報告では，漸近分布は共分散行列に依存して変化することを示し，特に，ある共分散行列に対しては対応する漸近分布のパーセント点が容易にもとまらない場合があることを示す. そのような場合の対処法も簡単に紹介する. 漸近分布の導出に関する理論的な詳細，共分散行列と漸近分布の関係を示す例，そして簡単な数値実験の結果は当日報告する.

参考文献

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

[2] Dempster, A. P. (1958), A high dimensional two sample significance test, *The Annals of Mathematical Statistics*, **29**, 995–1010.

[3] Dempster, A. P. (1960), A significance test for the separation of two highly multivariate small samples, *Biometrics*, **16**, 41–56.

[4] Fujikoshi, Y., Himeno, T. and Wakaki, H. (2004), Asymptotic results of a high dimensional MANOVA test and power comparison when the dimension is large compared to the sample size, *Journal of the Japan Statistical Society*, **34**, 19–26.

[5] Grenander, U. and Szegö, G (1958), *Toeplitz forms and their applications*, 2nd edition, Chelsea Publishing Company, New York.

[6] Johnstone, I.M. (2007), On the distribution of the largest eigenvalue in principal components analysis, *The Annals of Statistics*, **29**, 295–327.

[7] Srivastava, M.S. (2007), Multivariate theory for analyzing high dimensional data, *Journal of the Japan Statistical Society*, **37**, 53–86.

[8] Srivastava, M.S. and Du, M. (2008), A test for the mean vector with fewer observations than the dimension, *Journal of Multivariate Analysis*, **99**, 386–402.