母共分散行列からの影響を受けない高次元平均ベクトルの検定法の提案

大阪大学大学院基礎工学研究科　藤本 翔太
大阪大学大学院基礎工学研究科　狩野 裕

1. はじめに

高次元平均ベクトルに対する検定問題を考える．すなわち，独立な n 個の p 次元確率変数ベクトル $\boldsymbol{X}_p^{(1)},\ldots,\boldsymbol{X}_p^{(n)}$ に対して $\boldsymbol{\mu}_p = E(\boldsymbol{X}_p^{(i)})$ とするとき，

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

を検定することを考える．高次元データ $(n \leq p)$ の場合，伝統的な検定方法である Hotelling の検定統計量による方法は使用できない．そのため，Bai and Saranadasa (1996), Srivastava (2009), Chen and Qin (2010) などによって，高次元データにも対応できる検定統計量が提案されており，(n,p) が共に大きくなる場合のそれらの H_0 における漸近分布も同時に導出されている．しかしながら，漸近分布の導出の際には，母共分散行列 $\boldsymbol{\Sigma}_p = \mathrm{Var}(\boldsymbol{X}_p^{(i)})$ に対して強い条件 $\mathrm{tr}\boldsymbol{\Sigma}_p^4 = o\{(\mathrm{tr}\boldsymbol{\Sigma}_p^2)^2\}$ が要求される．実際，$\boldsymbol{\Sigma}_p$ が Compound Symmetry の構造 $(1-\rho)\boldsymbol{I}_p + \rho\mathbf{1}_p\mathbf{1}_p^T, 0 < \rho < 1$ を持つ場合や，$\boldsymbol{\Sigma}_p$ の最大固有値が $O(p^{-\alpha}), \alpha > 1/2$ で与えられる場合にこの条件は満たされない．加えて，Fujimoto et al. (2010) では，先行研究における各統計量の H_0 の下での漸近分布の形状は $\boldsymbol{\Sigma}_p$ の固有値構造に依存して変化することが示されている．実際に検定を行う際に，これらの点は問題である．そこで，本報告では，漸近分布が $\boldsymbol{\Sigma}_p$ に依存して変化しないような検定統計量を提案する．

2. モデルと提案する検定統計量

Srivastava (2009) と同様に，以下のモデルを考える：

$$\boldsymbol{X}_p^{(i)} = \boldsymbol{\mu}_p + \boldsymbol{C}_p\boldsymbol{Z}_p^{(i)}, \quad i = 1,\ldots,n.$$

ここで，$\boldsymbol{C}_p$ は $p \times p$ 行列であり，$\boldsymbol{Z}_p^{(i)} = (z_{i1},\ldots,z_{ip})^T$ とする．また，z_{ij} は独立な確率変数で，$E(z_{ij}) = 0$, $E(z_{ij}^2) = 1$ および $E(z_{ij}^4) = \gamma < \infty$ を満たすものとする．高次元データにおける漸近理論を考えるために，次の条件を仮定する：**条件 A**. $n = n(p) \to \infty$ as $p \to \infty$.

今，上記の検定問題 (1) に対して次の検定統計量を考える：

$$T_{n,p} = \mathrm{tr}\boldsymbol{X}_{n,p}^T\boldsymbol{M}_n\boldsymbol{X}_{n,p}.$$

ここで，$\boldsymbol{X}_{n,p} = (\boldsymbol{X}_p^{(1)},\ldots,\ \boldsymbol{X}_p^{(n)})^T$ であり，$\boldsymbol{M}_n$ は次の条件 B を満たすような $n \times n$ 対称行列とする．**条件 B**. $\boldsymbol{M}_n = (m_{ij}), m_{ij} \geq 0, m_{ii} = 0, \mathbf{1}_n^T\boldsymbol{M}_n\mathbf{1}_n \neq 0, \mathrm{tr}\boldsymbol{M}_n^2 \neq 0, \mathrm{tr}\boldsymbol{M}_n^4 = o\{(\mathrm{tr}\boldsymbol{M}_n^2)^2\}$. このとき，条件 A および B の下で，$H_0$ における $T_{n,p}$ の漸近正規性が導かれる．すなわち，この漸近正規性は，条件 B を満たすような行列 $\boldsymbol{M}_n$ を適当に与え，それを用いて検定統計量を構成することによって達成される．詳細は当日報告する．

3. 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] Chen, S. X. and Qin, Y. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The annals of statistics* 38: 808–835.
[3] Fujimoto, S., Kano, Y. and Srivastava, M. S. (2010). Asymptotic distributions of some test criteria for the mean vector with fewer observations than the dimension. *Submitted*.
[4] Srivastava, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *Jounal of Multivariate Analysis*. 100: 518–532.