Approximate and asymptotic distributions of Dempster trace criterion for high dimensional data

藤本翔太，狩野裕
大阪大学大学院基礎工学研究科

1. はじめに

正規分布に基づく 1 標本問題を考える．すなわち，n 個の標本 $\boldsymbol{X}_p^{(1)}, \cdots, \boldsymbol{X}_p^{(n)}$ が p 変量正規分布 $N_p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ から得られたときの，検定問題 $H_0 : \boldsymbol{\mu}_p = \mathbf{0}$ を考える．この検定問題に対する伝統的な方法は，Hotelling の T^2 統計量 $n\bar{\boldsymbol{X}}_{\boldsymbol{n,p}}^T \boldsymbol{S}_{n,p}^{-1} \bar{\boldsymbol{X}}_{\boldsymbol{n,p}}$ によるものである．ここに，$\bar{\boldsymbol{X}}_{n,p}$ は標本平均，$\boldsymbol{S}_{n,p}$ は標本共分散行列とする．しかしながら，高次元データ，すなわち $n < p$ の場合は $\boldsymbol{S}_{n,p}$ が特異になるため，この統計量は定義されない．そこで，Dempster(1958) は高次元データにも対応できる統計量として，Hotelling の T^2 統計量における $\boldsymbol{S}_{n,p}$ をそのトレースで置き換えた統計量 $\bar{\boldsymbol{X}}_{n,p}^T \bar{\boldsymbol{X}}_{n,p} / \mathrm{tr}\boldsymbol{S}$ を提案した．これを Dempster trace criterion (DTC) と呼ぶことにする．DTC の漸近分布による検定方法は Fujikoshi(2004) によって提案され，F 分布近似による方法は Srivastava(2007) によって提案されているが，どちらの方法も共分散行列 $\boldsymbol{\Sigma}_p$ に厳しい条件を課している．そこで本報告では，先行研究の結果をより現実的な条件の下に理論的に拡張する．

2. 共分散行列への条件

Fujikoshi(2004)，Srivastava(2007) では，共分散行列 $\boldsymbol{\Sigma}_p$ に次の条件を課している．

$$\text{(A)} \quad 0 < {}^{\exists} \lim_{p\to\infty} \frac{\boldsymbol{\Sigma}_p^i}{p} < \infty \quad (i = 1, 2, 3, 4)$$

条件 (A) は例えば次の (a),(b),(c) の場合は満たされるが，(d),(e) の場合は満たされない．

(a) $\boldsymbol{\Sigma}_p = aI_p$．ここに，$a > 0$ であり，I_p は $p \times p$ の単位行列を表す．
(b) Mild Spiked Model : $\lambda_i = C_1 p^{\alpha}$ $(i = 1, 2, \cdots, m)$ and $\lambda_j = C_2$ $(j = m+1, \cdots, p)$．ここに，λ_i $(i = 1, 2, \cdots, p)$ は $\boldsymbol{\Sigma}_p$ の固有値であり，$C_1, C_2 > 0$, $\alpha \leq 1/4$, $m < p$ とする．
(c) Autoregressive Model : $\boldsymbol{\Sigma}_p = (\rho^{|i-j|})$ $(0 < \rho < 1)$.
(d) Sharp Spiked Model : (b) において $\alpha > 1/4$ の場合．
(e) Compound Symmetry Structure : $\boldsymbol{\Sigma}_p = (1-\rho)I_p + \rho \mathbf{1}_p \mathbf{1}_p^T$ $(0 < \rho < 1)$．ここに，$\mathbf{1}_p$ は要素が全て 1 の p 次元列ベクトルを表す．

そこで，以上の全ての例を含むような次の条件を考える．

$$\text{(B)} \quad {}^{\exists}\delta_i \geq 0 \ \text{ s.t. } \ \delta_1 \geq \delta_2 \geq \delta_k \ (k = 3, 4, \cdots,) \text{ and } \ 0 < {}^{\exists} \lim_{p\to\infty} \mathrm{tr}\left(\frac{\boldsymbol{\Sigma}_p}{p^{\delta_i}}\right)^i < \infty \ \ (i = 1, 2, \cdots)$$

条件 (B) の下では，DTC の $(n, p) \to \infty$ における漸近分布が状況に応じて変化することが示された．さらに，DTC の F 分布近似は条件 (B) の下でも妥当であることが示された．当日，理論的な詳細とあわせて，2 つの方法を比較した簡単な数値実験の結果も報告する．

3. 参考文献

[1] Dempster, A. P. (1958), A high dimensional two sample significance test, *The Annals of Mathematical Statistics*, **29**, 995–1010.
[2] 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.
[3] Srivastava, M.S. (2007), Multivariate theory for analyzing high dimensional data, *Journal of the Japan Statistical Society*, **37**, 53–86.