高次元データにおける Dempster's trace criterion の近似分布について

藤本 翔太，狩野 裕
大阪大学大学院基礎工学研究科

Introduction

- 設定

$$\boldsymbol{X}_p^{(1)}, \cdots, \boldsymbol{X}_p^{(n)} \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$$

$$\bar{\boldsymbol{X}}_{n,p} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_p^{(i)}, \quad \boldsymbol{S}_{n,p} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_p^{(i)} - \bar{\boldsymbol{X}}_{n,p})(\boldsymbol{X}_p^{(i)} - \bar{\boldsymbol{X}}_{n,p})^T$$

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

- 高次元の枠組み（$n \leq p$）

$$T_H^2 = n\bar{\boldsymbol{X}}_{n,p}^T \boldsymbol{S}_{n,p}^{-1} \bar{\boldsymbol{X}}_{n,p} \Rightarrow T_D^2 = \frac{n\bar{\boldsymbol{X}}_{n,p}^T \bar{\boldsymbol{X}}_{n,p}}{\mathrm{tr}\boldsymbol{S}_{n,p}}$$

- T_D^2 に関する先行研究

近似分布	漸近分布
Dempster (1958)	Bai and Saranadasa (1996)
Srivastava (2007)	Fujikoshi (2004)

- 次の条件 (A) を仮定

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

- かなり強い条件（$\boldsymbol{\Sigma}_p$ がほとんど単位行列）

- より現実的な条件で T_D^2 の近似分布，漸近分布を考える

Previous Works

- Srivastava (2007)

帰無仮説 H_0 の下

$$T_D^2 = \frac{(n-1)\boldsymbol{Y}_p^{(1)T}\boldsymbol{Y}_p^{(1)}}{\boldsymbol{Y}_p^{(2)T}\boldsymbol{Y}_p^{(2)} + \cdots + \boldsymbol{Y}_p^{(n)T}\boldsymbol{Y}_p^{(n)}}, \quad \boldsymbol{Y}_p^{(1)}, \cdots, \boldsymbol{Y}_p^{(n)} \overset{i.i.d.}{\sim} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_p)$$

$\boldsymbol{Y}_p^{(i)T}\boldsymbol{Y}_p^{(i)} \sim m\chi_r^2$（近似） $\Rightarrow$ $T_D^2 \sim F_{r,(n-1)r}$（近似）

2 次モーメントまで一致 $\Rightarrow$ $r = \dfrac{(\mathrm{tr}\boldsymbol{\Sigma}_p)^2}{\mathrm{tr}\boldsymbol{\Sigma}_p^2}$

- r を推定できれば近似的に検定可能

定理 1：条件 (A) を仮定．このとき任意の $\varepsilon > 0$ に対して，

$$\lim_{n\to\infty}\sup_p P\left(\left|\frac{\hat{a}_1}{\mathrm{tr}\boldsymbol{\Sigma}_p} - 1\right| > \varepsilon\right) = 0, \quad \lim_{n\to\infty}\sup_p P\left(\left|\frac{\hat{a}_2}{\mathrm{tr}\boldsymbol{\Sigma}_p^2} - 1\right| > \varepsilon\right) = 0,$$

$$\lim_{n\to\infty}\sup_p P\left(\left|\frac{\hat{r}}{r} - 1\right| > \varepsilon\right) = 0, \quad \text{where}$$

$$\hat{a}_1 = \mathrm{tr}\boldsymbol{S}_{n,p}, \quad \hat{a}_2 = \frac{(n-1)^2}{(n+1)(n-2)}\left[\mathrm{tr}\boldsymbol{S}_{n,p}^2 - \frac{1}{n-1}(\mathrm{tr}\boldsymbol{S}_{n,p})^2\right], \quad \hat{r} = \frac{\hat{a}_1^2}{a_2}.$$

- Bai and Saranadasa (1996), Fujikoshi (2004)

定理 2：条件 (A) を仮定し $\lim_{p\to\infty} n = \infty$ とする．このとき，

$$T_{BS}^2 = \frac{1}{\sqrt{2\mathrm{tr}\boldsymbol{\Sigma}_p^2}}\left(n\bar{\boldsymbol{X}}_{n,p}^T\bar{\boldsymbol{X}}_{n,p} - \mathrm{tr}\boldsymbol{S}_{n,p}\right) \xrightarrow{d} N(0,1) \quad \text{as } p \to \infty$$

$$T_F^2 = \frac{\mathrm{tr}\boldsymbol{\Sigma}_p}{\sqrt{2\mathrm{tr}\boldsymbol{\Sigma}_p^2}}\left(T_D^2 - 1\right) \xrightarrow{d} N(0,1) \quad \text{as } p \to \infty$$

- $\mathrm{tr}\boldsymbol{\Sigma}_p, \mathrm{tr}\boldsymbol{\Sigma}_p^2$ が未知の場合は $\hat{a}_1, \hat{a}_2$ で置き換えて検定

Main Results

- 条件 (A) を弱めた条件 (B)

$$\text{(B)} \quad {}^{\exists}\delta_i \geq 0\ (i = 1,2,\cdots) \text{ s.t. } \delta_1 \geq \delta_2 \geq \delta_k\ (k = 3,4,\cdots) \text{ and}$$

$$0 < {}^{\exists}\lim_{p\to\infty} \mathrm{tr}\left(\frac{\boldsymbol{\Sigma}_p}{p^{\delta_i}}\right)^i < \infty \quad (i = 1,2,\cdots)$$

- $\delta_i = 1/i$ ならば条件 (A)

- 近似分布について

定理 3：定理 1 は条件 (B) の下でも成立つ．

- 漸近分布について

定理 4：条件 (B) を仮定し，$\lim_{p\to\infty} n = \infty$ とする．

(B1) $\delta_2 > \delta_k\ (k = 3,4,\cdots)$

$\Rightarrow$ T_{BS}^2 and $T_F^2 \xrightarrow{d} N(0,1)$ as $p \to \infty$

(B2) $\delta_2 = \delta_k$ and $\lim_{p\to\infty} \mathrm{tr}\left(\dfrac{\boldsymbol{\Sigma}_p}{p^{\delta_i}}\right)^i = m\ (k = 3,4,\cdots)$

$\Rightarrow$ T_{BS}^2 and $T_F^2 \xrightarrow{d} \dfrac{\chi_m^2 - m}{\sqrt{2m}}$ as $p \to \infty$

- (B2) の例（Spiked Model）

$$\lambda_i = Cp^{\alpha}\ (i = 1, \cdots, m), \quad \lambda_j = 1\ (j = m+1, \cdots, p),$$

$$C > 0,\ \alpha > 0,\ m < p,\ \lambda_i : \boldsymbol{\Sigma}_p \text{の固有値}$$

- Monte Carlo Simulation

有意水準 0.05，繰り返し数 1000，次の Σ_p(CS)：

$$\boldsymbol{\Sigma}_p = (1-\rho)I_p + \rho\boldsymbol{1}_p\boldsymbol{1}_p^T \quad (0 \leq \rho < 1)$$

- $\rho = 0 \Rightarrow$ 条件 (B1)，それ以外 $\Rightarrow$ 条件 (B2)

ρ	n	p	T_D	T_F^N	T_{BS}^N	T_F^χ	T_{BS}^χ
0.8	40	40	0.054	0.098	0.073	0.067	0.075
	40	100	0.062	0.088	0.082	0.056	0.046
	80	100	0.048	0.077	0.070	0.060	0.064
	40	200	0.055	0.085	0.074	0.058	0.059
	80	200	0.069	0.091	0.074	0.053	0.063
	150	200	0.065	0.066	0.067	0.049	0.050

上添え字 N, χ は用いた漸近分布を表す
青数字：最も良い近似，赤数字：最も悪い近似

Conclusions

- 条件 (A) をより現実的な状況を許す条件 (B) に変更
  - 近似分布 $\Rightarrow$ そのままで OK
  - 漸近分布 $\Rightarrow$ 状況に応じて変化
- Monte Carlo Simulation の結果をみて
  - 近似分布による検定が一様に良いパフォーマンス
  - 漸近分布を間違えば当然悪い結果
- 今後の課題
  - 両者の検定方法の理論的比較