高次元データにおける平均ベクトルの検定

[1] 藤本 翔太，[1] 狩野 裕，[2]Muni S. Srivastava

[1] 大阪大学大学院基礎工学研究科，[2]Department of Statistics, University of Toronto

Introduction

- 設定

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

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

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

- 高次元データ（$n\leq p$）

$$T=n\bar{\boldsymbol{X}}_{n,p}^T\boldsymbol{S}_{n,p}^{-1}\bar{\boldsymbol{X}}_{n,p}\quad\Rightarrow\quad \text{定義されない}$$

- 高次元データにも対応可能な検定統計量

Dempster(1958) Bai and Saranadasa(1996)

$$T_D=n\bar{\boldsymbol{X}}_{n,p}^T\bar{\boldsymbol{X}}_{n,p}/\text{tr}\boldsymbol{S}_{n,p}\qquad T_{BS}=n\bar{\boldsymbol{X}}_{n,p}^T\bar{\boldsymbol{X}}_{n,p}-\text{tr}\boldsymbol{S}_{n,p}$$

- $H_0,(n,p)\to\infty$ における漸近正規性に次の条件を要求

$$\text{(A)}\quad \text{tr}\boldsymbol{\Sigma}_p^4=o\{(\text{tr}\boldsymbol{\Sigma}_p^2)^2\}$$

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

- Fujimoto et al.(2010,Submitted)

漸近分布の形状が $\boldsymbol{\Sigma}_p$ に依存して変化

$$\tilde{T}_{BS}=\frac{T_{BS}}{\sqrt{2\text{tr}\boldsymbol{\Sigma}_p^2}}\xrightarrow{d}\Phi_{\{\boldsymbol{\Sigma}_p\}},\quad (n,p)\to\infty$$

$\Phi_{\{\boldsymbol{\Sigma}_p\}}\Rightarrow$ Normal, Chi-Square, Convolution

定理 1 ($\text{tr}\boldsymbol{\Sigma}_p^2$ の推定量)：任意の $\varepsilon>0$ に対して，

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

$$\text{where}\quad \hat{a}=\frac{(n-1)^2}{(n+1)(n-2)}\left[\text{tr}\boldsymbol{S}_{n,p}^2-\frac{1}{n-1}(\text{tr}\boldsymbol{S}_{n,p})^2\right]$$

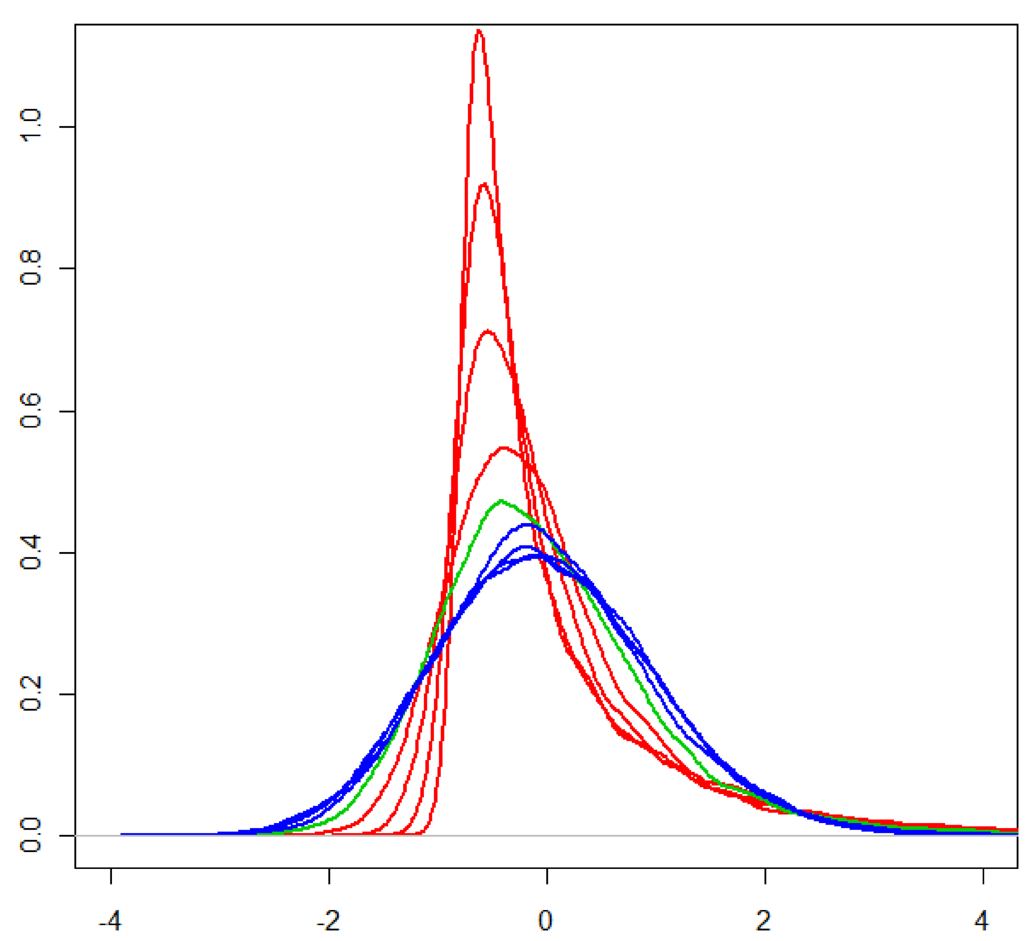
- Toy Model

$$\boldsymbol{\Sigma}_p=(1-\rho_p)\boldsymbol{I}_p+\rho_p\boldsymbol{1}_p\boldsymbol{1}_p^T,\quad \rho_p=p^{-\alpha}\quad(0<\alpha<1)$$

$$0<\alpha<1/2\ \Rightarrow\ \tilde{T}_{BS}\xrightarrow{d}(\chi_1^2-1)/\sqrt{2}$$

$$\alpha=1/2\ \Rightarrow\ \tilde{T}_{BS}\xrightarrow{d}\tfrac{1}{\sqrt{2}}N(0,1)+\tfrac{1}{\sqrt{2}}\tfrac{\chi_1^2-1}{\sqrt{2}}$$

$$1/2<\alpha<1\ \Rightarrow\ \tilde{T}_{BS}\xrightarrow{d}N(0,1)$$



$n=100,\ \ p=200,\ \ $ 繰り返し回数 $10,000$

赤：$\alpha=0.1-0.4$，緑：$\alpha=0.5$，青：$\alpha=0.6-0.9$

Proposed Test

- 任意の $\boldsymbol{\Sigma}_p$ に対して漸近正規する検定統計量の提案

$$T_{M_n}=\text{tr}\boldsymbol{X}_{n,p}^T\boldsymbol{M}_n\boldsymbol{X}_{n,p},\quad \boldsymbol{X}_{n,p}=(\boldsymbol{X}_p^{(1)},\ldots,\ \boldsymbol{X}_p^{(n)})^T$$

- $n\times n$ 行列 $\boldsymbol{M}_n$ は次を満たすようにとる

$$\text{(B)}\ \text{tr}\boldsymbol{M}_n=0;\ \boldsymbol{1}_n^T\boldsymbol{M}_n\boldsymbol{1}_n\neq0;\ \text{tr}\boldsymbol{M}_n^2\neq0;\ \text{tr}\boldsymbol{M}_n^4=o\{(\text{tr}\boldsymbol{M}_n^2)^2\}$$

- Moments

$$E(T_{M_n})=(\boldsymbol{1}_n^T\boldsymbol{M}_n\boldsymbol{1}_n)\boldsymbol{\mu}_p^T\boldsymbol{\mu}_p+\boxed{\text{tr}\boldsymbol{M}_n\text{tr}\boldsymbol{\Sigma}_p}$$

$$\text{Var}(T_{M_n})=2\text{tr}\boldsymbol{M}_n^2\text{tr}\boldsymbol{\Sigma}_p^2+4(\boldsymbol{1}_n^T\boldsymbol{M}_n^2\boldsymbol{1}_n)\boldsymbol{\mu}_p^T\boldsymbol{\Sigma}_p\boldsymbol{\mu}_p$$

定理 2 条件 (B) を仮定．H_0 の下，以下が成立つ：

$$\tilde{T}_{M_n}=\frac{T_{M_n}}{\sqrt{2\text{tr}\boldsymbol{M}_n^2\text{tr}\boldsymbol{\Sigma}_p^2}}\xrightarrow{d}N(0,1),\quad (n,p)\to\infty$$

- なぜ漸近正規性に $\boldsymbol{\Sigma}_p$ の条件を要求しないか？

1 以下

Lyapunov 条件 ($\delta=2$) $\Rightarrow$ $15\boxed{\dfrac{\text{tr}\boldsymbol{M}_n^4}{(\text{tr}\boldsymbol{M}_n^2)^2}}\boxed{\dfrac{\text{tr}\boldsymbol{\Sigma}_p^4}{(\text{tr}\boldsymbol{\Sigma}_p^2)^2}}$

0 に収束

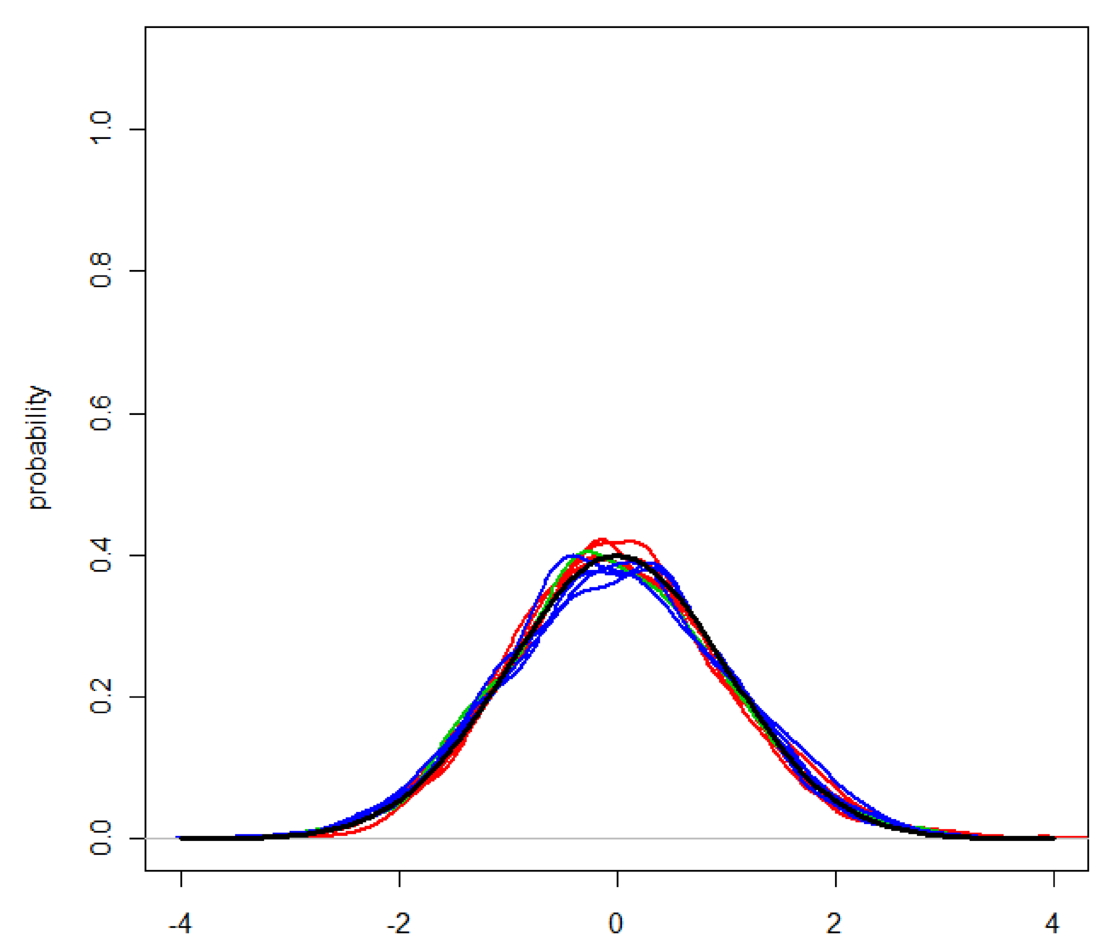
- 行列 $\boldsymbol{M}_n$ の候補

$$\boldsymbol{M}_n^{(1)}=(m_{ij})\ \text{with}\ m_{ii}=0\ \text{and}\ m_{ij}=\rho^{|i-j|}\quad(|\rho|<1)$$

$$\boldsymbol{M}_n^{(2)}=(m_{|i-j|})\ \text{with}\ m_0=0\ \text{and}\ \lim_{n\to\infty}\textstyle\sum_{k=1}^{n-1}|m_k|<\infty$$

- Simulation（左と同じデータを使用）

$$\boldsymbol{M}_n=\boldsymbol{M}_n^{(1)},\quad \text{tr}\boldsymbol{M}_n^4/(\text{tr}\boldsymbol{M}_n^2)^2=0.02$$



Further Works

- 漸近正規性を保ちながら検出力も大きくする

$$\max_{\boldsymbol{M}_n}\frac{\boldsymbol{1}_n^T\boldsymbol{M}_n\boldsymbol{1}_n}{\sqrt{\text{tr}\boldsymbol{M}_n^2}}\quad \textit{subject to}\quad \frac{\text{tr}\boldsymbol{M}_n^4}{(\text{tr}\boldsymbol{M}_n^2)^2}\leq c$$

- NOTE

制約条件なしの最適化 $\Rightarrow$ $\tilde{T}_{M_n}\approx\tilde{T}_{BS}$

- 問題点と今後の課題
  - $\boldsymbol{T}_{M_n}$ の取る値はデータの並び替えに依存

  $\Rightarrow$ Random Permutation Sample を使って客観化