# 強スパイク固有値モデルにおける 高次元一標本検定とその応用について

東京理科大学・情報科学科　　石井　晶
Aki Ishii
Department of Information Sciences
Tokyo University of Science

筑波大学・数理物質系　　矢田　和善
Kazuyoshi Yata
Institute of Mathematics
University of Tsukuba

筑波大学・数理物質系　　青嶋　誠
Makoto Aoshima
Institute of Mathematics
University of Tsukuba

## 1　はじめに

平均が $\boldsymbol{\mu}$，共分散行列が $\boldsymbol{\Sigma}$ の $p$ 次元分布をもつ母集団から，$n$ 個の $p$ 次元データベクトル $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ を無作為に抽出し，$p \times n$ データ行列 $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ を定義する．ただし，$p > n$ とする．適当な直交行列 $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_p]$ で $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_p)$ と分解する．ただし，$\lambda_1 \geq \cdots \geq \lambda_p (> 0)$ とする．そのとき，$\boldsymbol{Z} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{H}^T(\boldsymbol{X} - [\boldsymbol{\mu}, ..., \boldsymbol{\mu}])$ を定義し，$\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_p]^T$, $\boldsymbol{z}_j = (z_{j1}, ..., z_{jn})^T$ と表記する．ただし，$\boldsymbol{Z}$ の成分は，4 次モーメントが一様有界になることを仮定する．さらに，各 $j$ で $z_{oj}^2 = \sum_{k=1}^n (z_{jk} - \bar{z}_j)^2/(n-1)$，$\bar{z}_j = n^{-1}\sum_{k=1}^n z_{jk}$ とおく．そのとき，$P(\liminf_{p\to\infty} z_{oj}^2 > 0) = 1$ となることを仮定する．

高次元統計解析において重要なことは，高次元共分散行列の固有空間の特徴に基づいた理論や方法論の構築である．特に，高次元共分散行列の固有値が次元数 $p$ の関数となることに注意しなければならない．実際，青嶋 [2]，青嶋・矢田 [3, 4] で解説されているように，ゲノムデータの最大固有値は次元数のべき乗関数となる．このような背景から，Aoshima and Yata [6] は，高次元データに対して固有値モデルを次のように 2 つに分類した．1 つ目は，強スパイク固有値モデル (strongly spiked eigenvalue

(SSE) モデル) と呼ばれ，以下のように定義される．

$$\liminf_{p\to\infty}\left\{\frac{\lambda_1^2}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\right\}>0 \tag{1.1}$$

2 つ目は，弱スパイク固有値モデル (non-SSE モデル) と呼ばれ，以下のように定義される．

$$\frac{\lambda_1^2}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\to 0 \quad (p\to\infty) \tag{1.2}$$

任意の高次元データは，上記 2 つの固有値モデルの何れかに分類される．弱スパイク固有値モデルに対しては，その固有空間の推測をはじめ，平均ベクトルの検定などの種々の統計量について，高次元漸近正規性という望ましい結果が得られる．弱スパイク固有値モデルにおける統計的推測の概説は，Aoshima and Yata [5] を参照のこと．その一方で，強スパイク固有値モデルに対しては，高次元漸近正規性が成り立たない．

標本共分散行列を $\boldsymbol{S}_n=(n-1)^{-1}(\boldsymbol{X}-\overline{\boldsymbol{X}})(\boldsymbol{X}-\overline{\boldsymbol{X}})^T=(n-1)^{-1}\sum_{i=1}^n(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T$ とする．ここで，$\overline{\boldsymbol{X}}=[\overline{\boldsymbol{x}},...,\overline{\boldsymbol{x}}]$，$\overline{\boldsymbol{x}}=\sum_{j=1}^n\boldsymbol{x}_j/n$ である．高次元平均ベクトルに対して，次の検定問題を考える．

$$H_0:\ \boldsymbol{\mu}=\boldsymbol{\mu}_0 \quad \text{vs.} \quad H_1:\ \boldsymbol{\mu}\neq\boldsymbol{\mu}_0 \tag{1.3}$$

ここで，$\boldsymbol{\mu}_0$ は $\boldsymbol{\mu}_0=\boldsymbol{0}$ など，既知のベクトルである．本稿では，一般性を失うことなく，$\boldsymbol{\mu}_0=\boldsymbol{0}$ と仮定する．

高次元小標本データに対する一標本検定においては，標本共分散行列の逆行列が存在しないため，ホテリングの $T^2$ 統計量を用いることはできない．そこで，Dempster [9, 10] や Srivastava [14] は，$\boldsymbol{X}$ が正規分布の場合に，高次元検定手法を提案した．一方，$\boldsymbol{X}$ が非正規分布の場合に，Bai and Saranadasa [7] は次の検定統計量を議論した．

$$T_{\mathrm{BS}}=\|\overline{\boldsymbol{x}}\|^2-\mathrm{tr}(\boldsymbol{S}_n)/n$$

$E(T_{\mathrm{BS}})=\|\boldsymbol{\mu}\|^2$ となることに注意する．Bai and Saranadasa [7] は，弱スパイク固有値モデルといくつかの正則条件のもと，$T_{\mathrm{BS}}$ に関する漸近正規性を示した．しかしながら，$T_{\mathrm{BS}}$ の漸近正規性は弱スパイク固有値モデルに非常に敏感であり，強スパイク固有値モデルのもとでは，精度が非常に悪くなる．そこで，本稿では，強スパイク固有値モデルに対する一標本検定について，Ishii, Yata and Aoshima [11, 12] で与えた検定手法を解説する．

## 2　単一強スパイク固有値モデルにおける高次元一標本検定

Ishii, Yata and Aoshima [11] では，単一強スパイク固有値モデルのもと，固有空間の推測や一標本検定，さらには共分散行列の同等性について結果を与えた．単一強スパイク固有値モデルは，次のように定義される．

**(A-i)**　$\dfrac{\sum_{s=2}^p\lambda_s^2}{\lambda_1^2}=o(1)\quad(p\to\infty)$

ここで，単一強スパイク固有値モデルは，強スパイク固有値モデル (1.1) の 1 つであることに注意する．いま，$z$ に関して，次のモーメント条件を考える．

**(A-ii)** $\dfrac{\sum_{r,s\geq 2}^{p}\lambda_r\lambda_s E\{(z_{rk}^2-1)(z_{sk}^2-1)\}}{n\lambda_1^2}=o(1)\quad(p\to\infty)$

さらに，第 1 主成分である $z_{1j},\ j=1,...,n$ には，必要に応じて，次の条件を仮定する．

**(A-iii)** $z_{1j},\ j=1,...,n$ は，互いに独立に標準正規分布 $N(0,1)$ に従う．

(A-iii) は第 1 主成分のみに正規性を仮定した緩い条件である．さらに，(A-iii) のもと $(n-1)z_{o1}^2$ は自由度 $n-1$ のカイ二乗分布 $\chi_{n-1}^2$ に従うことに注意する．Yata and Aoshima [16, 17] や Yata and Aoshima [15] は，「ノイズ掃き出し法」や「クロスデータ行列法」という新しい PCA を開発した．さらに，「$p\to\infty$ かつ $n\to\infty$」の枠組みで，固有値推定量の一致性について議論した．$\boldsymbol{S}_n$ の固有値を $\hat{\lambda}_1\geq\cdots\geq\hat{\lambda}_p\ (\geq 0)$ をとする．いま，$n\times n$ の行列

$$\boldsymbol{S}_D=(n-1)^{-1}(\boldsymbol{X}-\overline{\boldsymbol{X}})^T(\boldsymbol{X}-\overline{\boldsymbol{X}})$$

を $\boldsymbol{S}_n$ と正の固有値を共有する双対標本共分散行列という．ノイズ掃き出し法を用いると，固有値は次のように推定される．

$$\tilde{\lambda}_i=\hat{\lambda}_i-\frac{\mathrm{tr}(\boldsymbol{S}_D)-\sum_{j=1}^{i}\hat{\lambda}_j}{n-1-i}\quad(i=1,...,n-2)\tag{2.1}$$

その一方で，Ishii, Yata and Aoshima [11] は，「$p\to\infty$ だが，$n$ は固定」の枠組みで，$\tilde{\lambda}_1$ の漸近分布を導出した．

**定理 2.1** (Ishii, Yata and Aoshima [11])．(A-i) と (A-ii) を仮定する．「$p\to\infty$ だが，$n$ は固定」の枠組みで，次が成り立つ．

$$\frac{\tilde{\lambda}_1}{\lambda_1}=z_{o1}^2+o_P(1)$$

さらに，(A-iii) を仮定する．「$p\to\infty$ だが，$n$ は固定」の枠組みで，次が成り立つ．

$$(n-1)\frac{\tilde{\lambda}_1}{\lambda_1}\Rightarrow\chi_{n-1}^2$$

ここで，$\Rightarrow$ は分布収束を表す．

定理 2.1 の結果を用い，Ishii, Yata and Aoshima [11] では，単一強スパイク固有値モデルのもと，「$p\to\infty$ だが，$n$ は固定」の枠組みで有用な検定手法を提案した．$T_{\mathrm{BS}}$ に対し，次の結果を得た．

**補題 2.1** (Ishii, Yata and Aoshima [11])．(A-i) を仮定する．「$p\to\infty$ だが，$n$ は固定」の枠組みで，次が成り立つ．

$$\frac{\|\overline{\boldsymbol{x}}-\boldsymbol{\mu}\|^2-\mathrm{tr}(\boldsymbol{S}_D)/n}{\lambda_1}=\bar{z}_1^2-\frac{z_{o1}^2}{n}+o_P(1)$$

補題 2.1 をもとに，Ishii, Yata and Aoshima [11] では，次の検定統計量を提案した．

$$F_0=\frac{n\|\overline{\boldsymbol{x}}\|^2-\mathrm{tr}(\boldsymbol{S}_D)}{\tilde{\lambda}_1}+1$$

ここで，$E(\tilde{\lambda}_1(F_0-1)/n)=\|\boldsymbol{\mu}\|^2$ となることに注意する．したがって，$F_0$ の漸近帰無分布が以下のように得られる．

**定理 2.2** (Ishii, Yata and Aoshima [11]). (A-i) から (A-iii) を仮定する．$H_0$ のもと，「$p \to \infty$ だが，$n$ は固定」の枠組みで次が成り立つ．

$$F_0 \Rightarrow F_{1,n-1}$$

ここで，$F_{\nu_1,\nu_2}$ は自由度 $(\nu_1, \nu_2)$ の $F$ 分布にしたがう確率変数である．

よって，予め設定した $\alpha \in (0, 1/2)$ に対し，検定問題 (1.3) に対する検定ルールを次のように与えた．

$$H_0 \text{ を棄却する} \iff F_0 > F_{1,n-1}(\alpha)$$

ここで，$F_{\nu_1,\nu_2}(\alpha)$ は自由度 $(\nu_1, \nu_2)$ の $F$ 分布の上側 $\alpha$ 点である．いま，検定統計量 $F$ に対する第 1 種の過誤 (Size) を Size $(F)$ と表す．このとき，定理 2.2 の条件のもと，$F_0$ の第 1 種の過誤に対して，「$p \to \infty$ だが，$n$ は固定」の枠組みで，次の結果が得られる．

$$\text{Size } (F_0) = \alpha + o(1)$$

したがって，$n = 5$ 程度の高次元小標本データであっても，上記の検定手法により，第 1 種の過誤を $\alpha$ に抑えた一標本検定を行うことができる．

## 3　データ変換を用いた高次元一標本検定

Ishii, Yata and Aoshima [12] では，強スパイク固有値モデルから弱スパイク固有値モデルへのデータ変換を用いた新たな検定手法を提案した．また，強スパイク固有値モデルのもと，「$p \to \infty$ かつ $n \to \infty$」の枠組みで理論の構築を行った．必要に応じて，次の仮定をおく．

**(A-iv)** $E(z_{rl}^2 z_{sl}^2) = E(z_{rl}^2)E(z_{sl}^2) = 1$, $E(z_{rl}z_{sl}z_{tl}) = 0$, $E(z_{rl}z_{sl}z_{tl}z_{ul}) = 0$ $(r \neq s, t, u)$

ここで，(A-iv) は母集団の正規性を緩めた仮定である．いま，$\Psi_r$ を次のようにおく．

$$\Psi_r = \sum_{s=r}^{p} \lambda_s^2 \quad (r \geq 1)$$

Aoshima and Yata [6] と同様に，次の条件を満たす強スパイク固有値モデルを考える．

**(A-v)** 次元数 $p$ に依存しないある自然数 $k$ に対し，

**(i)** $1 \leq r < s \leq k$ のとき，$\liminf_{p\to\infty}(\lambda_r/\lambda_s - 1) > 0$

**(ii)** $\liminf_{p\to\infty} \dfrac{\lambda_k^2}{\Psi_k} > 0$ かつ $\dfrac{\lambda_{k+1}^2}{\Psi_{k+1}} \to 0$ $(p \to \infty)$

つまり，$k$ は強スパイクする固有値の個数であり，強スパイクする $k$ 個の固有空間を除くと，残りの固有空間は弱スパイク固有値モデルとなる．Aoshima and Yata [6] で与えられたデータ変換を用いて，強スパイク固有値モデルから弱スパイク固有値モデルにデータを変換する．次の正射影行列を考える．

$$\boldsymbol{A} = \boldsymbol{I}_p - \sum_{j=1}^{k} \boldsymbol{h}_j \boldsymbol{h}_j^T = \sum_{j=k+1}^{p} \boldsymbol{h}_j \boldsymbol{h}_j^T$$

$\boldsymbol{A}$ は，最初の $k$ 個の固有空間の直交補空間への正射影行列である．すると，$\boldsymbol{A}\boldsymbol{x}_j$ の期待値と分散は次のようになる．$E(\boldsymbol{A}\boldsymbol{x}_j) = \boldsymbol{A}\boldsymbol{\mu}$ $(= \boldsymbol{\mu}_*$とおく$)$，

$$\mathrm{Var}(\boldsymbol{A}\boldsymbol{x}_j) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A} = \sum_{j=k+1}^{p} \lambda_j \boldsymbol{h}_j \boldsymbol{h}_j^T \ (= \boldsymbol{\Sigma}_*\text{とおく})$$

$\lambda_{\max}(\boldsymbol{\Sigma}_*)$ を $\boldsymbol{\Sigma}_*$ の最大固有値とすると，$\mathrm{tr}(\boldsymbol{\Sigma}_*^2) = \Psi_{k+1}$ かつ $\lambda_{\max}(\boldsymbol{\Sigma}_*) = \lambda_{k+1}$ となることに注意する．したがって，(A-v) のもとで，

$$\lambda_{\max}^2(\boldsymbol{\Sigma}_*)/\mathrm{tr}(\boldsymbol{\Sigma}_*^2) \to 0 \quad (p \to \infty)$$

となるので，$\boldsymbol{A}\boldsymbol{x}_j$ は弱スパイク固有値モデルとなる．変換後のデータを用いて，次の統計量を考える．

$$T_{\mathrm{DT}} = \|\boldsymbol{A}\overline{\boldsymbol{x}}\|^2 - \frac{\mathrm{tr}(\boldsymbol{A}\boldsymbol{S})}{n} = 2\frac{\sum_{l<l'}^{n} \boldsymbol{x}_l^T \boldsymbol{A} \boldsymbol{x}_{l'}}{n(n-1)} = 2\frac{\sum_{l<l'}^{n} \left(\boldsymbol{x}_l^T \boldsymbol{x}_{l'} - \sum_{j=1}^{k} x_{jl} x_{jl'}\right)}{n(n-1)}$$

ここで，任意の $j, l$ に対し，

$$x_{jl} = \boldsymbol{h}_j^T \boldsymbol{x}_l$$

である．$\Delta_* = \|\boldsymbol{\mu}_*\|^2$ とする．$E(T_{\mathrm{DT}}) = \Delta_*$ であり，$\mathrm{Var}(T_{\mathrm{DT}}) = K_*$ とおくと，

$$K_* = K_{1*} + K_{2*};\ K_{1*} = 2\frac{\mathrm{tr}(\boldsymbol{\Sigma}_*^2)}{n(n-1)},\ K_{2*} = 4\frac{\boldsymbol{\mu}_*^T \boldsymbol{\Sigma}_* \boldsymbol{\mu}_*}{n}$$

である．$m = \min\{p, n\}$ とおく．(A-v) のもと，必要に応じて，次を仮定する．

**(A-vi)** $\limsup_{m\to\infty} \dfrac{\Delta_*^2}{K_{1*}} < \infty$

**定理 3.1** (Ishii, Yata and Aoshima [12])**.** (A-iv) から (A-vi) のもと，$m \to \infty$ で次が成り立つ．

$$\frac{T_{\mathrm{DT}} - \Delta_*}{K_*^{1/2}} = \frac{T_{\mathrm{DT}} - \Delta_*}{K_{1*}^{1/2}} + o_P(1) \Rightarrow N(0,1)$$

したがって，Ishii, Yata and Aoshima [12] では，データ変換を用いて，次の検定統計量を構築した．

$$\widehat{T}_{\mathrm{DT}} = 2\frac{\sum_{l<l'}^{n} \left(\boldsymbol{x}_l^T \boldsymbol{x}_{l'} - \sum_{j=1}^{k} \tilde{x}_{jl} \tilde{x}_{jl'}\right)}{n(n-1)}$$

ここで，$\tilde{x}_{jl}$ はノイズ掃き出し法を用いた $x_{jl}$ の推定量である．また，強スパイクしている固有値の数である $k$ も未知であるため，Aoshima and Yata [6] で与えられた推定方法を用いて $k$ を推定する．(A-v) のもと，次を仮定する．

**(A-vii)** $\dfrac{\lambda_1^2}{n\mathrm{tr}(\boldsymbol{\Sigma}_*^2)} \to 0 \ \ (m \to \infty)$，　**(A-viii)** $\liminf_{p\to\infty} \dfrac{\Delta_*}{\Delta} > 0 \ \ (\Delta \neq 0)$

通常は強スパイクしている固有値の数 $k$ より，次元数 $p$ の方が遥かに大きいため，(A-viii) は緩い仮定である．また，(A-viii) は，データ変換を行った上で，検定問題 (1.3) を扱うことが保証されるという意味である．このとき，以下の結果を得た．

**定理 3.2** (Ishii, Yata and Aoshima [12]). (A-iv) から (A-viii) のもと，$m \to \infty$ で次が成り立つ．

$$\frac{\widehat{T}_{\mathrm{DT}} - \Delta_*}{\widehat{K}_{1*}^{1/2}} \Rightarrow N(0,1)$$

ここで，$\widehat{K}_{1*}$ は $K_{1*}$ のクロスデータ行列法による一致推定量である．

Ishii, Yata and Aoshima [12] では，定理 3.2 より，予め設定した $\alpha \in (0, 1/2)$ に対し，検定問題 (1.3) に対する検定ルールを次のように与えた．

$$H_0 \text{ を棄却する} \iff \frac{\widehat{T}_{\mathrm{DT}}}{\widehat{K}_{1*}^{1/2}} > z_\alpha \tag{3.1}$$

ここで，$z_\alpha$ は $N(0,1)$ の上側 $\alpha$ 点である．いま，検定統計量 $F$ に対する検出力 (Power) を Power $(F)$ と表す．検定ルール (3.1) を用いると，第 1 種の過誤と検出力は以下のようになる．

**定理 3.3** (Ishii, Yata and Aoshima [12]). (A-iv), (A-v), (A-vii), (A-viii) を仮定する．$m \to \infty$ で次が成り立つ．

$$\text{Size } (\widehat{T}_{\mathrm{DT}}) = \alpha + o(1), \quad \text{Power } (\widehat{T}_{\mathrm{DT}}) = \Phi\Big( \frac{\Delta_*}{K_*^{1/2}} - z_\alpha \Big( \frac{K_{1*}}{K_*} \Big)^{1/2} \Big) + o(1)$$

ここで，$\Phi(\cdot)$ は $N(0,1)$ の累積分布関数である．

## 4　多標本問題への応用とシミュレーション

$g$ 個の $p$ 次元母集団 $\pi_i$, $i = 1, ..., g$ を仮定する．各母集団の平均ベクトルを $\boldsymbol{\mu}_i$，分散共分散行列を $\boldsymbol{\Sigma}_i$ $(i = 1, ..., g)$ とする．母集団 $\pi_i$ から，互いに独立な標本 $\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}$ をとる．一般性を失うことなく，$n_1 = \min\{n_1, ..., n_g\}$ と仮定する．次の検定問題を考える．

$$H_0 : \sum_{i=1}^{g} b_i \boldsymbol{\mu}_i = \mathbf{0} \quad \text{vs.} \quad H_1 : \sum_{i=1}^{g} b_i \boldsymbol{\mu}_i \neq \mathbf{0} \tag{4.1}$$

ここで，$b_i$ $(i = 1, ..., g)$ は次元数 $p$ に依存せず，0 でないスカラーである．いま，$j = 1, ..., n_1$ について，Bennett [8] や Anderson [1] で与えられている以下のデータ変換を行う．

$$\boldsymbol{x}_j = b_1 \boldsymbol{x}_{1j} + \sum_{i=2}^{g} b_i \sqrt{\frac{n_1}{n_i}} \Big( \boldsymbol{x}_{ij} - \frac{1}{n_1} \sum_{j=1}^{n_1} \boldsymbol{x}_{ij} + \frac{1}{\sqrt{n_1 n_i}} \sum_{j'=1}^{n_i} \boldsymbol{x}_{ij'} \Big) \tag{4.2}$$

このとき，以下が成り立つ．

$$E(\boldsymbol{x}_j) = \sum_{i=1}^{g} b_i \boldsymbol{\mu}_i, \, \mathrm{Var}(\boldsymbol{x}_j) = \sum_{i=1}^{g} b_i^2 (n_1/n_i) \boldsymbol{\Sigma}_i$$

いま，各母集団 $\pi_i$ に対し，母集団分布を $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ と仮定する．そのとき，$\boldsymbol{x}_1, ..., \boldsymbol{x}_{n_1}$ は，互いに独立に $N_p(\sum_{i=1}^{g} b_i \boldsymbol{\mu}_i, \ \sum_{i=1}^{g} b_i^2 (n_1/n_i) \boldsymbol{\Sigma}_i)$ に従う．Nishiyama et al. [13] では，変換 (4.2) を用い，弱スパイク固有値モデルのもと，(4.1) に対する検定手法を提案した．この節では，強スパイク固有値モデル (1.1) のもと，新たな検定手法を与え

る．$n = n_1$，$\boldsymbol{\mu} = \sum_{i=1}^{g} b_i \boldsymbol{\mu}_i$，$\boldsymbol{\Sigma} = \sum_{i=1}^{g} b_i^2 (n_1/n_i)\boldsymbol{\Sigma}_i$ とする．通常，$\pi_i$ $(i = 1, ..., g)$ が強スパイク固有値モデルであれば，$\boldsymbol{x}_j$ も強スパイク固有値モデルとなることに注意する．例えば，$i \geq 2$ に対し $\mathrm{tr}(\boldsymbol{\Sigma}_1^2) \geq \mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ かつ $\liminf_{p\to\infty} \lambda_{\max}^2(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_1^2) > 0$ であれば，

$$\liminf_{p\to\infty} \frac{\lambda_{\max}^2(\boldsymbol{\Sigma})}{\mathrm{tr}(\boldsymbol{\Sigma}^2)} \geq \liminf_{p\to\infty} \frac{b_1^2 \lambda_{\max}^2(\boldsymbol{\Sigma}_1)}{(\sum_{i=1}^{g} b_i^2)^2 \mathrm{tr}(\boldsymbol{\Sigma}_1^2)} > 0$$

が成り立つ．したがって，$\boldsymbol{x}_j$ を用い，(3.1) で与えた一標本検定を行うことで，(4.1) の検定を行うことができる．

しかしながら，上記の多標本検定では，理論的には母集団の正規性という高次元データに対しては厳しい仮定が必要である．そこで，母集団の正規性に関して，シミュレーションにより本検定手法の頑健性を検証する．$g = 4$ とする．$i = 1, ..., 4$ に対し，$b_i = 1$，$\boldsymbol{\mu}_i = \mathbf{0}$，$\boldsymbol{\Sigma}_i = c_i \mathrm{diag}(p^{2/3}, p^{1/3}, 1, ..., 1)$ とし，$c_1 = 1$ かつ $c_2 = c_3 = c_4 = 2$ と設定した．以下の 3 つの場合について考察をした．

(a) $\boldsymbol{x}_{ij}, j = 1, ..., n_i$ が互いに独立に $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ に従う場合．ここで，$n_1 = \lceil p^{1/2} \rceil$，$p = 2^s (s = 6, ..., 12)$ かつ $n_i = 2n_1$ $(i = 2, 3, 4)$ とした．

(b) $\boldsymbol{x}_{ij} = \boldsymbol{\Sigma}_i^{1/2} (z_{1j(i)}, ..., z_{pj(i)})^T + \boldsymbol{\mu}_i$ $(j = 1, ..., n_i)$ と表記する．$z_{rj(i)} = (v_{rj(i)} - \nu)/\sqrt{2\nu}$ について，$v_{rj(i)}, j = 1, ..., n_i; r = 1, ..., p$ が互いに独立に自由度 $\nu$ のカイ二乗分布に従う場合．ここで，$\nu = 5s$ $(s = 1, ..., 7)$，$n_1 = 50$ かつ $n_i = 100$ $(i = 2, 3, 4)$ とした．

(c) $\boldsymbol{x}_{ij} = \boldsymbol{\Sigma}_i^{1/2} (z_{1j(i)}, ..., z_{pj(i)})^T + \boldsymbol{\mu}_i$ $(j = 1, ..., n_i)$ と表記する．$(z_{1j(i)}, ..., z_{pj(i)})^T, j = 1, ..., n_i$ が互いに独立に平均ベクトル $\mathbf{0}$，分散共分散行列 $\boldsymbol{I}_p$ で，自由度 $\nu$ の $p$ 変量 $T$ 分布 $t_p(\boldsymbol{I}_p, \nu)$ に従う場合．ここで，$\nu = 5s$ $(s = 1, ..., 7)$，$n_1 = 50$ かつ $n_i = 100$ $(i = 2, 3, 4)$ とした．

各 (a) から (c) に対し，2000 回の繰り返し計算により，本検定手法の性能を評価した．$r = 1, ..., 2000$ に対し，(4.1) の $H_0$ が棄却 (または，採択) されたとき，$P_r = 1$ (または，0) と定義する．第 1 種の過誤を $\overline{\alpha} = \sum_{r=1}^{2000} P_r/2000$ で推定した．図 1 は，各 (a) から (c) について，$\overline{\alpha}$ をプロットしたものである．理論的に示した通り，(a) では，高次元で良い性能を示していることが見てとれる．また，(b) では，本検定手法の頑健性が見てとれる．一方で，(c) では，$T$ 分布の裾の重さに影響を受けているように思われる．

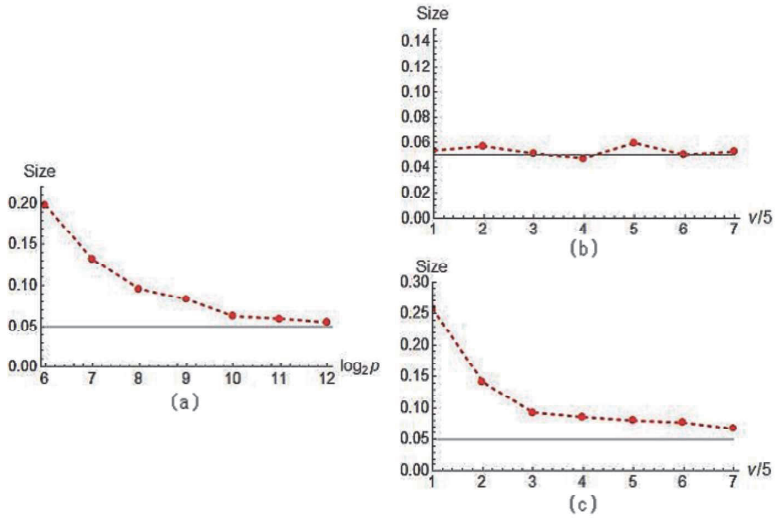


図 1 : (a) から (c) における第 1 種の過誤 $\alpha$ をプロットした．左部は (a) $\boldsymbol{x}_{ij}, j = 1, ..., n_i$ が互いに独立に $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ に従う場合．右上部は (b) $\boldsymbol{x}_{ij} = \boldsymbol{\Sigma}_i^{1/2}(z_{1j(i)}, ..., z_{pj(i)})^T + \boldsymbol{\mu}_i$ $(j = 1, ..., n_i)$ と表記するとき，$z_{rj(i)} = (v_{rj(i)} - \nu)/\sqrt{2\nu}$ について，$v_{rj(i)}, j = 1, ..., n_i; r = 1, ..., p$ が互いに独立に自由度 $\nu$ のカイ二乗分布に従う場合．右下部は (c) $\boldsymbol{x}_{ij} = \boldsymbol{\Sigma}_i^{1/2}(z_{1j(i)}, ..., z_{pj(i)})^T + \boldsymbol{\mu}_i$ $(j = 1, ..., n_i)$ について，$(z_{1j(i)}, ..., z_{pj(i)})^T, j = 1, ..., n_i$ が互いに独立に平均ベクトル $\mathbf{0}$，分散共分散行列 $\boldsymbol{I}_p$ で，自由度 $\nu$ の $p$ 変量 $T$ 分布 $t_p(\boldsymbol{I}_p, \nu)$ に従う場合．

## 謝辞

本研究は，科学研究費補助金 基盤研究 (A) 15H01678 研究代表者: 青嶋 誠「大規模複雑データの理論と方法論の総合的研究」，学術研究助成基金助成金 挑戦的研究 (萌芽) 17K19956 研究代表者: 青嶋 誠「非スパースモデリングによるビッグデータの新展開」，科学研究費補助金 若手研究 18K18015 研究代表者: 石井 晶「高次元データの統計的推測：スパイク性とスパース性」，及び，科学研究費補助金 基盤研究 (C) 18K03409 研究代表者: 矢田 和善「高次元データにおける高次漸近理論の開拓とその応用」から研究助成を受けています．

## 参考文献

[1] Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd edn. New York: Wiley.

[2] 青嶋 誠 (2018).「日本統計学会賞受賞者特別寄稿論文: 高次元統計解析: 理論と方法論の新しい展開」『日本統計学会誌』 **48**, 89-111.

[3] 青嶋誠，矢田和善 (2013).「論説: 高次元小標本における統計的推測」『数学』 **65**, 225-247.

[4] 青嶋誠，矢田和善 (2013).「日本統計学会研究業績賞受賞者特別寄稿論文: 高次元データの統計的方法論」『日本統計学会誌』 **43**, 123-150.

[5] Aoshima, M. and Yata, K. (2017). Statistical inference for high-dimension, low-sample-size data, *American Mathematical Society, Sugaku Expositions*, **30**, 137-158.

[6] Aoshima, M. and Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models, *Statistica Sinica*, **28**, 43-62.

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

[8] Bennett, B.M. (1951). Note on a solution of the generalized Behrens-Fisher problem, *Annals of the Institute of Statistical Mathematics*, **2**, 87-90.

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

[10] Dempster, A.P. (1960). A significance test for the separation of two highly multivariate small samples, *Biometrics*, **16**, 41-50.

[11] Ishii, A., Yata, K. and Aoshima, M. (2016). Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context, *Journal of Statistical Planning and Inference*, **170**, 186-199.

[12] Ishii, A., Yata, K. and Aoshima, M. (2019). Inference on high-dimensional mean vectors under the strongly spiked eigenvalue model, *Japanese Journal of Statistics and Data Science*, in press.

[13] Nishiyama, T., Hyodo, M., Seo, T., Pavlenko, T. (2013). Testing linear hypotheses of mean vectors for high-dimension data with unequal covariance matrices, *Journal of Statistical Planning and Inference*, **143**, 1898-1911.

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

[15] Yata, K. and Aoshima, M. (2010). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix, *Journal of Multivariate Analysis*, **101**, 2060-2077.

[16] Yata, K. and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations, *Journal of Multivariate Analysis*, **105**, 193-215.

[17] Yata, K., Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings, *Journal of Multivariate Analysis*, **122**, 334-354.

Department of Information Sciences
Tokyo University of Science
Chiba 278-8510
Japan
E-mail address: a.ishii@rs.tus.ac.jp

東京理科大学・情報科学科　石井　晶