

| | |
|-------------|--|
| Title | On the Distribution of the Largest Eigenvalue in High Dimension, Low Sample Size Context (Asymptotic Expansions for Various Models and Their Related Topics) |
| Author(s) | 矢田, 和善; 青嶋, 誠 |
| Citation | 数理解析研究所講究録 (2013), 1860: 120-128 |
| Issue Date | 2013-11 |
| URL | http://hdl.handle.net/2433/195290 |
| Right | |
| Type | Departmental Bulletin Paper |
| Textversion | publisher |

On the Distribution of the Largest Eigenvalue in High Dimension, Low Sample Size Context

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

情報化の進展に伴い、高次元データの統計解析が益々重要になってきている。2000 年以降、確率論と理論物理の方面から、ランダム行列の理論に基づく幾つかの重要な結果がもたらされた。Baik and Silverstein [5], Johnstone [7], Paul [10] 等による、標本固有値の漸近的性質の導出である。しかしながら、これらの研究は、データの次元数 p と標本数 n が $n/p \rightarrow c > 0$ を満たす場合を考え、高次元において標本数は次元数と同程度を仮定し、母集団分布には正規分布もしくは類する条件を仮定していた。次元数は優に 10,000 を超えるが標本数は高々 100 程度といった高次元小標本においては、標本数を次元数と同程度には仮定できない。それゆえ、 n が p に依存しないような設定で、もしくは、 $n = n(p)$ であっても $n/p \rightarrow 0$ となる設定で、高次元漸近理論を展開する必要がある。

Yata and Aoshima [11, 12] は、高次元小標本のもとで‘クロスデータ行列法’とよばれるノンパラメトリックな方法論を考案した。クロスデータ行列法は、データセットを 2 分割して掛け合わせ、クロスデータ行列という非正則な行列を定義し、これに基づいて高速かつ高精度な汎用性の高い推測を可能にする。Yata and Aoshima [11] は、クロスデータ行列の特異値分解に基づいて固有値の推定と漸近分布を求め、さらに固有ベクトルや主成分スコアの推定も与えて、それらが高次元小標本のもとで一致性をもつことを証明した。一方、高次元小標本データ空間を幾何学的に捉えるための研究もある。Ahn et al. [1], Hall et al. [6], Jung et al. [9], Yata and Aoshima [13] は、標本数 n を高々 100 程度に固定して次元数 p を $p \rightarrow \infty$ としたときの高次元データ空間の幾何学的表現を見つけている。Hall et al. [6] は高次元データの挙動を幾何学的に捉え、Ahn et al. [1] は標本共分散行列が有する幾何学的表現を導出し、Jung et al. [9] は固有ベクトルについて幾何学的表現を論じた。しかしながら、これらは、母集団分布が正規分布もしくは類するものであることを仮定した。それに対し、Yata and Aoshima [13] は分布に関する

限定を取り去って非正規分布の場合も扱い、先行研究では見つけられなかった高次元小標本データの2つの幾何学的表現を発見した。ある非正規性の尺度に関する閾値を境にした、固有空間の球面集中現象と座標軸集中現象である。Yata and Aoshima [13] は、固有空間の球面集中現象に基づいて‘ノイズ掃き出し法’とよばれる固有空間のセミパラメトリックな推定法を考案した。Aoshima and Yata [2, 3] は、高次元小標本において重要となる8つの統計的推測問題を取り上げ、それぞれに対して幾何学的表現に基づいた統計量をクロスデータ行列法で構築し、それらの漸近分布を導出して、推測の精度を保証する標本数の設計に至るまでの一連の基礎理論と方法論を築き上げた。なお、青嶋・矢田 [4] は、高次元小標本の統計的推測について、最新の動向を詳しく解説している。

本論文は‘ $p \rightarrow \infty$ だが n は固定’という枠組みで高次元小標本漸近理論を扱い、ノイズ掃き出し法による最大固有値の推定について漸近分布を導出する。平均に p 次のベクトル $\boldsymbol{\mu}_{(p)}$ 、共分散行列に p 次の非負定値対称行列 $\boldsymbol{\Sigma}_{(p)} (\geq \mathbf{O})$ をもつ母集団を考える。 n 個の p 次データベクトル $\mathbf{x}_{1(p)}, \dots, \mathbf{x}_{n(p)}$ を無作為に抽出して、データ行列 $\mathbf{X}_{(p)} : p \times n = [\mathbf{x}_{1(p)}, \dots, \mathbf{x}_{n(p)}]$ を定義する。ただし、 $p > n$ である。 $\boldsymbol{\Sigma}_{(p)}$ の固有値を $\lambda_{1(p)} \geq \dots \geq \lambda_{p(p)} (\geq 0)$ とし、適当な直交行列 $\mathbf{H}_{(p)} = [\mathbf{h}_{1(p)}, \dots, \mathbf{h}_{p(p)}]$ で $\boldsymbol{\Sigma}_{(p)} = \mathbf{H}_{(p)} \boldsymbol{\Lambda}_{(p)} \mathbf{H}_{(p)}^T$ 、 $\boldsymbol{\Lambda}_{(p)} = \text{diag}(\lambda_{1(p)}, \dots, \lambda_{p(p)})$ と分解する。そのとき $\mathbf{X}_{(p)} - [\boldsymbol{\mu}_{(p)}, \dots, \boldsymbol{\mu}_{(p)}] = \mathbf{H}_{(p)} \boldsymbol{\Lambda}_{(p)}^{1/2} \mathbf{Z}_{(p)}$ とおき、 $\mathbf{Z}_{(p)} = [\mathbf{z}_{1(p)}, \dots, \mathbf{z}_{p(p)}]^T$ 、 $\mathbf{z}_{i(p)} = (z_{i1(p)}, \dots, z_{in(p)})^T$ と表記する。ただし、 $\mathbf{Z}_{(p)}$ の成分は、4次モーメントが一様有界と仮定する。今後、次元数を意識して付した添え字 (p) は省いて表記する。

2 高次元小標本における双対空間の幾何学的表現

2.1 $\boldsymbol{\mu} = \mathbf{0}$ のとき

高次元小標本データを解析する上で鍵となるのは、データがもつ特有の幾何学的表現である。まず、簡単のために $\boldsymbol{\mu} = \mathbf{0}$ とし、標本共分散行列を $\mathbf{S}_o = n^{-1} \mathbf{X} \mathbf{X}^T$ とする。そのとき、双対な標本共分散行列 $\mathbf{S}_{oD} = n^{-1} \mathbf{X}^T \mathbf{X}$ は \mathbf{S}_o と正の固有値を共有する。 \mathbf{S}_{oD} の固有値を $\hat{\lambda}_{o1} \geq \dots \geq \hat{\lambda}_{on} (\geq 0)$ とし、 $\hat{\lambda}_{oj}$ に対する固有ベクトルを $\hat{\mathbf{u}}_{oj}$ とし、スペクトル分解を $\mathbf{S}_{oD} = \sum_{j=1}^n \hat{\lambda}_{oj} \hat{\mathbf{u}}_{oj} \hat{\mathbf{u}}_{oj}^T$ とする。次のような球形条件を考える。

$$(A-i) \quad \frac{\text{tr}(\boldsymbol{\Sigma}^2)}{\text{tr}(\boldsymbol{\Sigma})^2} = \frac{\sum_{s=1}^p \lambda_s^2}{(\sum_{s=1}^p \lambda_s)^2} \rightarrow 0, \quad p \rightarrow \infty.$$

この条件の詳細は、Jung and Marron [8] を参照されたい。そのとき、Ahn et al. [1] と Jung and Marron [8] は、 \mathbf{X} に正規分布もしくは \mathbf{Z} に ρ -mixing を仮定して

$$\frac{n}{\text{tr}(\boldsymbol{\Sigma})} \mathbf{S}_{oD} \xrightarrow{P} \mathbf{I}_n, \quad p \rightarrow \infty \quad (1)$$

を示した。ただし、 I_n は n 次単位行列である。幾何学的表現 (1) により、 \mathbf{X} が正規分布もしくは \mathbf{Z} が ρ -mixing の場合に、 $p \rightarrow \infty$ のとき次が成り立つ。

$$\frac{n\hat{\lambda}_{oi}}{\text{tr}(\boldsymbol{\Sigma})} = \frac{n}{\text{tr}(\boldsymbol{\Sigma})} \hat{\mathbf{u}}_{oi}^T \mathbf{S}_{oD} \hat{\mathbf{u}}_{oi} \xrightarrow{P} 1, \quad i = 1, \dots, n. \quad (2)$$

注意 1 非正規分布かつ非 ρ -mixing の場合の幾何学的表現については、Yata and Aoshima [13], 青嶋・矢田 [4] を参照のこと。

上記の幾何学的表現を精密に評価すれば、次の定理を得る。

定理 1 ([13]) もしも \mathbf{Z} が条件

$$(A\text{-ii}) \quad \frac{\sum_{r,s}^p \lambda_r \lambda_s E(z_{rk}^2 - 1)(z_{sk}^2 - 1)}{(\sum_{s=1}^p \lambda_s)^2} \rightarrow 0, \quad p \rightarrow \infty$$

を満たすならば、(A-i) のもと $p \rightarrow \infty$ のとき (1) と (2) が成り立つ。

2.2 $\boldsymbol{\mu} \neq \mathbf{0}$ のとき

次に、 $\boldsymbol{\mu} \neq \mathbf{0}$ の場合を考える。いま、標本共分散行列を $\mathbf{S} = (n-1)^{-1}(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^T = (n-1)^{-1} \sum_{j=1}^n (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^T$ とする。ここで、 $\bar{\mathbf{X}} = [\bar{\mathbf{x}}, \dots, \bar{\mathbf{x}}]$, $\bar{\mathbf{x}} = \sum_{j=1}^n \mathbf{x}_j / n$ である。双対な標本共分散行列は $\mathbf{S}_D = (n-1)^{-1}(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{X} - \bar{\mathbf{X}})$ となる。2.1 節と同様に、 \mathbf{S}_D の固有値を $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_{n-1} (\geq 0)$ とし、 $\hat{\lambda}_j$ に対する固有ベクトルを $\hat{\mathbf{u}}_j$ として、スペクトル分解を $\mathbf{S}_D = \sum_{j=1}^{n-1} \hat{\lambda}_j \hat{\mathbf{u}}_j \hat{\mathbf{u}}_j^T$ とする。ここで、 $\text{rank}(\mathbf{S}_D) \leq n-1$ に注意する。そのとき、次の幾何学的表現を得る。

定理 2 もしも \mathbf{Z} が (A-i) と (A-ii) を満たせば、 $p \rightarrow \infty$ のとき次が成り立つ。

$$\frac{n-1}{\text{tr}(\boldsymbol{\Sigma})} \mathbf{S}_D \xrightarrow{P} \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T.$$

ここで、 $\mathbf{1}_n = (1, \dots, 1)^T$ である。

系 1 もしも \mathbf{Z} が (A-i) と (A-ii) を満たせば、 $p \rightarrow \infty$ のとき次が成り立つ。

$$\frac{(n-1)\hat{\lambda}_i}{\text{tr}(\boldsymbol{\Sigma})} = \frac{n-1}{\text{tr}(\boldsymbol{\Sigma})} \hat{\mathbf{u}}_i^T \mathbf{S}_D \hat{\mathbf{u}}_i \xrightarrow{P} 1, \quad i = 1, \dots, n-1.$$

定理 2 と系 1 の幾何学的表現は \mathbf{S}_{oD} のものとは若干異なっているが、固有値が確率的に定まるという意味では、同様の性質をもつといえる。

3 固有値の推定量と漸近分布

Yata and Aoshima [13] は, (1) で与えられる幾何学的表現に着目して, ノイズ掃き出し法という固有値推定法を提案した. 定理 2 に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.

$$\hat{\lambda}_i = \hat{\lambda}_i - \frac{\text{tr}(\mathbf{S}_D) - \sum_{j=1}^i \hat{\lambda}_j}{n-1-i} \quad (i = 1, \dots, n-2). \quad (3)$$

(3) の第 2 項がノイズの掃き出しを意味している. この推定量は, $p \rightarrow \infty$ かつ $n \rightarrow \infty$ のとき, Yata and Aoshima [13] によって一致性と漸近分布が導かれている. 本論文では, $p \rightarrow \infty$ だが n は固定という枠組みで, 推定量の漸近分布を導出する. 最大固有値 λ_1 に, 次の条件を仮定する.

$$(A\text{-iii}) \quad \frac{\text{tr}(\Sigma^2) - \lambda_1^2}{\lambda_1^2} = \frac{\sum_{s=2}^p \lambda_s^2}{\lambda_1^2} \rightarrow 0, \quad p \rightarrow \infty,$$

$$(A\text{-iv}) \quad \frac{\sum_{r,s \geq 2}^p \lambda_r \lambda_s E(z_{rk}^2 - 1)(z_{sk}^2 - 1)}{\lambda_1^2} \rightarrow 0, \quad p \rightarrow \infty.$$

いま, $\bar{z}_j = n^{-1} \sum_{k=1}^n z_{jk}$, ($j = 1, \dots, p$) とし, $\mathbf{z}_{oj} = \mathbf{z}_j - (\bar{z}_j, \dots, \bar{z}_j)^T$ ($j = 1, \dots, p$) とおくと,

$$(n-1)\mathbf{S}_D = \lambda_1 \mathbf{z}_{o1} \mathbf{z}_{o1}^T + \sum_{j=2}^p \lambda_j \mathbf{z}_{oj} \mathbf{z}_{oj}^T$$

と書ける. このとき, 系 1 から, 第 2 項について (A-iii) と (A-iv) のもと $p \rightarrow \infty$ のとき次が成り立つ.

$$\hat{\mathbf{u}}_i^T \frac{\sum_{j=2}^p \lambda_j \mathbf{z}_{oj} \mathbf{z}_{oj}^T}{\lambda_1} \hat{\mathbf{u}}_i = \frac{\text{tr}(\Sigma) - \lambda_1}{\lambda_1} + o_p(1), \quad i = 1, \dots, n-1.$$

それゆえ,

$$\frac{\hat{\lambda}_1}{\lambda_1} = \hat{\mathbf{u}}_1^T \frac{\mathbf{S}_D}{\lambda_1} \hat{\mathbf{u}}_1 = (\hat{\mathbf{u}}_1^T \mathbf{z}_{o1} / \sqrt{n-1})^2 + \frac{\text{tr}(\Sigma) - \lambda_1}{\lambda_1(n-1)} + o_p(1)$$

が主張できる. ここで, $\mathbf{z}_{o1}^T \mathbf{1}_n = 0$ に注意する. そのとき, $\mathbf{z}_{o1} \neq \mathbf{0}$ ならば, (A-iii) と (A-iv) のもと $p \rightarrow \infty$ で次が成り立つ.

$$\frac{\hat{\lambda}_1}{\lambda_1} = \hat{\mathbf{u}}_1^T \frac{\mathbf{S}_D}{\lambda_1} \hat{\mathbf{u}}_1 = \|\mathbf{z}_{o1} / \sqrt{n-1}\|^2 + \frac{\text{tr}(\Sigma) - \lambda_1}{\lambda_1(n-1)} + o_p(1). \quad (4)$$

(3) において $i = 1$ のときの第 2 項は, (4) の第 2 項の一致推定量になる. 以上を纏めると次の定理を得る.

定理 3 もしも $z_{o1} \neq 0$ ならば, (A-iii) と (A-iv) のもと $p \rightarrow \infty$ のとき次が成り立つ.

$$\frac{\hat{\lambda}_1}{\lambda_1} = \|z_{o1}/\sqrt{n-1}\|^2 + o_p(1).$$

注意 2 もしも最大固有値に対して (A-iii) と (A-iv) よりも強い条件である $(\text{tr}(\Sigma) - \lambda_1)/\lambda_1 \rightarrow 0$, $p \rightarrow \infty$ を課すならば, (4) から $p \rightarrow \infty$ のとき次を得る.

$$\frac{\hat{\lambda}_1}{\lambda_1} = \|z_{o1}/\sqrt{n-1}\|^2 + o_p(1).$$

この結果と定理 3 を見比べると, ノイズ掃き出し法による最大固有値の推定量 $\hat{\lambda}_1$ は, 従来型の推定量 $\hat{\lambda}_1$ よりも緩い条件で一致性を導けることが分かる.

特に X が正規分布のとき, 次の結果を得る.

系 2 X が正規分布ならば, (A-iii) のもと $p \rightarrow \infty$ のとき

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

ここで, \Rightarrow は分布収束を表し, χ_{n-1}^2 は自由度 $n-1$ の χ^2 分布を表す.

注意 3 X が正規分布ならば, $(\text{tr}(\Sigma) - \lambda_1)/\lambda_1 \rightarrow 0$, $p \rightarrow \infty$ のもと $p \rightarrow \infty$ のとき

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

これは, Jung and Marron [8] の結果と同様の結果である.

系 2 をシミュレーション実験で検証する. 母集団は $N_p(\mu, \Sigma)$ とし, $\lambda_1 = p^{2/3}$, $\lambda_2 = \dots = \lambda_p = 1$ と設定した. 標本数は $n = 5$ とし, $p = 20, 100, 500, 2500$ の 4 つの場合を考えた. これらの設定は, (A-iii) と (A-iv) を満たすことに注意する. 図 1(a)-(d) は $(n-1)\hat{\lambda}_1/\lambda_1$ について, 図 2(a)-(d) は $(n-1)\hat{\lambda}_1/\lambda_1$ について, それぞれ 2000 回のシミュレーションによるヒストグラムを与えている. 系 2 で主張する通り, 次元数が上がるにつれて, ノイズ掃き出し法による $(n-1)\hat{\lambda}_1/\lambda_1$ のヒストグラムは χ_{n-1}^2 の分布に収束していく様子が見てとれる. 一方, 従来型の固有値推定による $(n-1)\hat{\lambda}_1/\lambda_1$ のヒストグラムは, 次元数が上がるにつれて χ_{n-1}^2 の分布から大きく乖離する. これは, 注意 3 の $(\text{tr}(\Sigma) - \lambda_1)/\lambda_1 \rightarrow 0$, $p \rightarrow \infty$ なる条件を満たさないために (4) の第 2 項がバイアスとして残り, それが次元数に対して単調増加するた

めである。このような場合にも、ノイズ掃き出し法を用いれば、最大固有値の漸近分布を得ることができるのである。

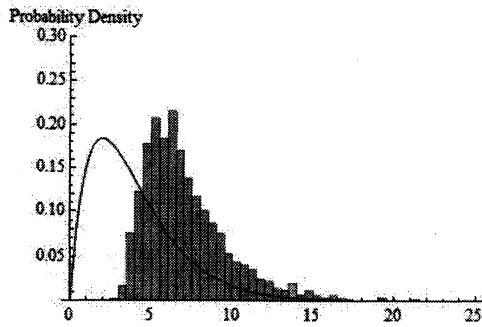
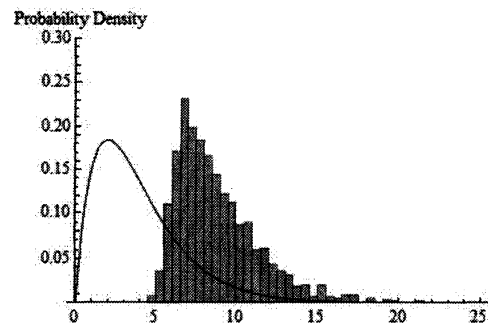
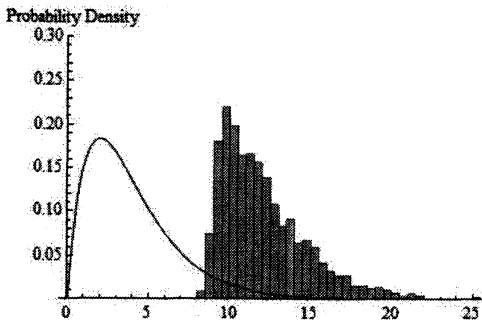
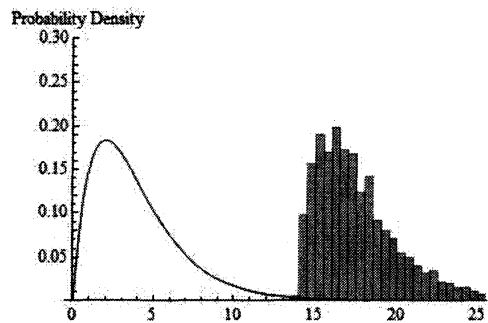
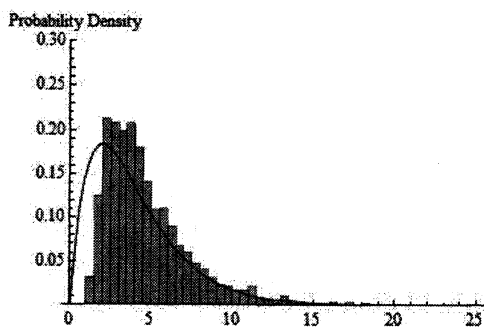
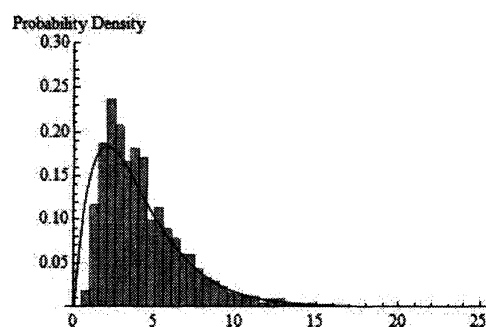
(a) $p = 20$ (b) $p = 100$ (c) $p = 500$ (d) $p = 2500$

図1 従来型の固有値推定による $(n-1)\hat{\lambda}_1/\lambda_1$ について、2000回のシミュレーションによるヒストグラム。実線は χ^2_{n-1} の密度関数を表す。

(a) $p = 20$ (b) $p = 100$

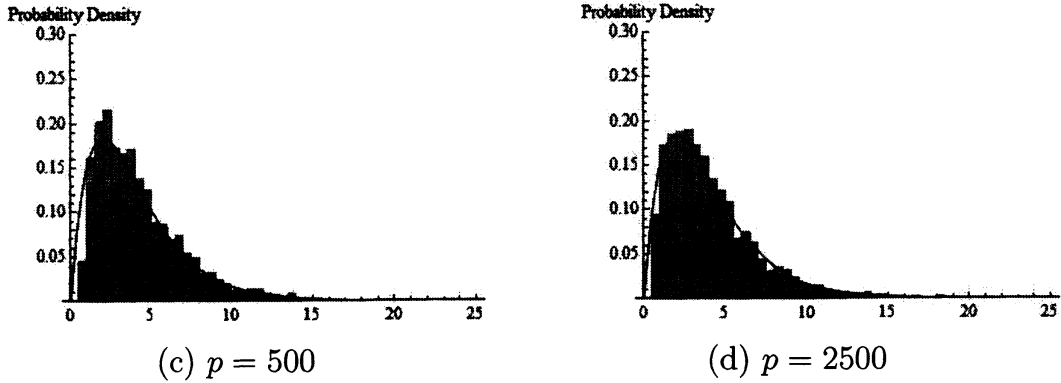


図2 ノイズ掃き出し法による $(n-1)\lambda_1/\lambda_1$ について、2000回のシミュレーションによるヒストグラム。実線は χ_{n-1}^2 の密度関数を表す。

4 付録

定理2の証明 (A-i) と (A-ii) のもとで、任意の $\tau > 0$ に対してチェビシェフの不等式を用いると、 $p \rightarrow \infty$ のとき

$$\begin{aligned}
 P\left(\left|\frac{\|\mathbf{x}_k - \boldsymbol{\mu}\|^2}{\text{tr}(\boldsymbol{\Sigma})} - 1\right| > \tau\right) &\leq \tau^{-2} \frac{\sum_{r,s}^p \lambda_r \lambda_s E\{(z_{rk}^2 - 1)(z_{sk}^2 - 1)\}}{\text{tr}(\boldsymbol{\Sigma})^2} \rightarrow 0; \\
 P\left(\left|\frac{(\mathbf{x}_k - \boldsymbol{\mu})^T (\mathbf{x}_{k'} - \boldsymbol{\mu})}{\text{tr}(\boldsymbol{\Sigma})}\right| > \tau\right) &\leq \tau^{-2} \frac{\text{tr}(\boldsymbol{\Sigma}^2)}{\text{tr}(\boldsymbol{\Sigma})^2} \rightarrow 0 \quad (k \neq k')
 \end{aligned} \tag{5}$$

となる。それゆえ、 $(\mathbf{X} - [\boldsymbol{\mu}, \dots, \boldsymbol{\mu}])^T (\mathbf{X} - [\boldsymbol{\mu}, \dots, \boldsymbol{\mu}]) / \text{tr}(\boldsymbol{\Sigma}) \xrightarrow{P} \mathbf{I}_n$ が成り立つ。さらに、(5)に注意し、 $\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|^2 / \text{tr}(\boldsymbol{\Sigma}) = \sum_{i,j} (\mathbf{x}_i - \boldsymbol{\mu})^T (\mathbf{x}_j - \boldsymbol{\mu}) / \{n^2 \text{tr}(\boldsymbol{\Sigma})\} \xrightarrow{P} 1/n$ 、 $(\bar{\mathbf{x}} - \boldsymbol{\mu})^T (\mathbf{x}_j - \boldsymbol{\mu}) / \text{tr}(\boldsymbol{\Sigma}) \xrightarrow{P} 1/n$ を得る。以上より、次が成り立つ。

$$\begin{aligned}
 &\frac{(n-1)\mathbf{S}_D}{\text{tr}(\boldsymbol{\Sigma})} \\
 &= \frac{(\mathbf{X} - [\boldsymbol{\mu}, \dots, \boldsymbol{\mu}])^T (\mathbf{X} - [\boldsymbol{\mu}, \dots, \boldsymbol{\mu}]) + (\bar{\mathbf{X}} - [\boldsymbol{\mu}, \dots, \boldsymbol{\mu}])^T (\bar{\mathbf{X}} - [\boldsymbol{\mu}, \dots, \boldsymbol{\mu}])}{\text{tr}(\boldsymbol{\Sigma})} \\
 &\quad - \frac{(\mathbf{X} - [\boldsymbol{\mu}, \dots, \boldsymbol{\mu}])^T (\bar{\mathbf{X}} - [\boldsymbol{\mu}, \dots, \boldsymbol{\mu}]) + (\bar{\mathbf{X}} - [\boldsymbol{\mu}, \dots, \boldsymbol{\mu}])^T (\mathbf{X} - [\boldsymbol{\mu}, \dots, \boldsymbol{\mu}])}{\text{tr}(\boldsymbol{\Sigma})} \\
 &\xrightarrow{P} \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T.
 \end{aligned}$$

それゆえ、結果を得る。 □

系1の証明 定理2より、漸近的に $\text{rank}(\mathbf{S}_D) = n-1$ となる。それゆえ、 $\hat{\mathbf{u}}_i^T \mathbf{1}_n = 0$ ($i = 1, \dots, n-1$) なることに注意すれば結果を得る。 □

定理 3 の証明 まず, (A-iv) のもと次が成り立つ.

$$\begin{aligned} & P\left(\left|\frac{\text{tr}(\mathbf{S}_D)}{\lambda_1} - \frac{\text{tr}(\boldsymbol{\Sigma}) - \lambda_1 + \lambda_1\|z_{o1}/\sqrt{n-1}\|^2}{\lambda_1}\right| > \tau\right) \\ &= P\left(\left|\frac{\sum_{s=2}^p \lambda_s \{ \|z_{os}\|^2 - (n-1) \}}{\lambda_1(n-1)}\right| > \tau\right) \\ &= O\left(\frac{\sum_{r,s \geq 2}^p \lambda_r \lambda_s E\{(z_{rk}^2 - 1)(z_{sk}^2 - 1)\}}{\lambda_1^2 n}\right) \rightarrow 0. \end{aligned}$$

それゆえ, $\text{tr}(\mathbf{S}_D)/\lambda_1 = (\text{tr}(\boldsymbol{\Sigma}) - \lambda_1 + \lambda_1\|z_{o1}/\sqrt{n-1}\|^2)/\lambda_1 + o_p(1)$ となり, これと (4) を合わせて, (3) の第 2 項について次が成り立つ.

$$\begin{aligned} \frac{\text{tr}(\mathbf{S}_D) - \hat{\lambda}_1}{\lambda_1(n-2)} &= \frac{\text{tr}(\boldsymbol{\Sigma}) - \lambda_1 + \lambda_1\|z_{o1}/\sqrt{n-1}\|^2 - \hat{\lambda}_1}{\lambda_1(n-2)} + o_p(1) \\ &= \frac{\text{tr}(\boldsymbol{\Sigma}) - \lambda_1}{\lambda_1(n-1)} + o_p(1). \end{aligned}$$

上記と (4) を合わせれば結果を得る. □

系 2 の証明 \mathbf{X} が正規分布ならば, z_{ik} , $i = 1, \dots, p$; $k = 1, \dots, n$ は互いに独立に $N(0, 1)$ に従う. さらに (A-iii) のもと, (A-iv) が成り立つ. そのとき, $\|z_{o1}\|^2 = \sum_{k=1}^n z_{1k}^2 - n\bar{z}_1^2$ が χ_{n-1}^2 に従うことに注意すれば, $P(\|z_{o1}\| \neq 0) = 1$ より結果を得る. □

謝辞 本研究は, 科学研究費補助金 基盤研究 (B) 22300094 研究代表者: 青嶋 誠「高次元データの理論と方法論の総合的研究」, および, 学術研究助成基金助成金 挑戦的萌芽研究 23650142 研究代表者: 青嶋 誠「高速で頑健かつ高精度な多変量統計手法の新展開」, 若手研究 (B) 23740066 研究代表者: 矢田 和善「高次元小標本の理論的体系の構築」から研究助成を受けている.

参考文献

- [1] J. Ahn, J. S. Marron, K. M. Muller and Y.-Y. Chi, The high-dimension, low-sample-size geometric representation holds under mild conditions, *Biometrika*, **94** (2007), 760-766.
- [2] M. Aoshima and K. Yata, Two-stage procedures for high-dimensional data, *Sequential Anal. (Editor's special invited paper)*, **30** (2011), 356-399.
- [3] M. Aoshima and K. Yata, Authors' responses, *Sequential Anal.*, **30** (2011), 432-440.

- [4] 青嶋 誠, 矢田和善, 論説:高次元小標本における統計的推測, 数学, **65** (2013), 印刷中.
- [5] J. Baik and J. W. Silverstein, Eigenvalues of large sample covariance matrices of spiked population models, *J. Multivariate Anal.*, **97** (2006), 1382-1408.
- [6] P. Hall, J. S. Marron and A. Neeman, Geometric representation of high dimension, low sample size data, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67** (2005), 427-444.
- [7] I. M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, *Ann. Statist.*, **29** (2001), 295-327.
- [8] S. Jung and J. S. Marron, PCA consistency in high dimension, low sample size context, *Ann. Statist.*, **37** (2009), 4104-4130.
- [9] S. Jung, A. Sen and J. S. Marron, Boundary behavior in high dimension, low sample size asymptotics of PCA, *J. Multivariate Anal.*, **109** (2012), 190-203.
- [10] D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, *Statist. Sinica*, **17** (2007), 1617-1642.
- [11] K. Yata and M. Aoshima, Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix, *J. Multivariate Anal.*, **101** (2010), 2060-2077.
- [12] K. Yata and M. Aoshima, Intrinsic dimensionality estimation of high-dimension, low sample size data with d -asymptotics, *Comm. Statist. Theory Methods*, Special Issue Honoring M. Akahira, (ed. M. Aoshima), **39** (2010), 1511-1521.
- [13] K. Yata and M. Aoshima, Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations, *J. Multivariate Anal.*, **105** (2012), 193-215.