数理解析研究所講究録
第1910巻 2014年 125-133

高次元小標本における混合データの幾何学的表現とクラスター分析への応用

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

高次元データ解析の理論と方法論を考えるとき，高次元小標本データの幾何学的表現に目を向けることは重要である．Hall et al. [6]，Ahn et al. [1]，Jung et al. [8]，Yata and Aoshima [13] は，標本数 n を固定して次元数 p を $p \to \infty$ としたときの，高次元小標本データの幾何学的表現を研究した．Hall et al. [6] はデータ点の挙動を幾何学的に捉え，Ahn et al. [1] は標本共分散行列の幾何学的表現を導出し，Jung et al. [8] は固有ベクトルについて幾何学的表現を論じた．これらの先行研究は，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ている．一方，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 [3] は，高次元小標本データの幾何学的表現に基づいて各種統計的推測の統計量を構築し，それらの漸近正規性を証明し，さらに標本数を設計することで，高次元統計的推測の精度保証に至るまでの一連の基礎理論と方法論を築いた．なお，高次元データの統計的推測について，青嶋・矢田 [4, 5] は重要となる基礎理論に詳しい解説を与えている．

本論文は，2つクラスからなる混合分布を考え，‘$p \to \infty$，n は固定’の枠組みでクラスター分析を考える．2つの分布を π_1, π_2 と名付け，それぞれ平均 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$ と，共分散行列 $\boldsymbol{\Sigma}_1$，$\boldsymbol{\Sigma}_2$ をもつと仮定する．各 $i\,(=1,2)$ について，適当な直交行列 $\boldsymbol{H}_i$ で $\boldsymbol{\Sigma}_i$ を $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i \boldsymbol{H}_i^T$, $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1}, ..., \lambda_{ip})$ と分解する．ただし，$\lambda_{i1} \geq \cdots \geq \lambda_{ip} (\geq 0)$ とする．そのとき，データ $\boldsymbol{x}_j$ が $\boldsymbol{x}_j \in \pi_i$ であれば，$\boldsymbol{x}_j - \boldsymbol{\mu}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i^{1/2} \boldsymbol{y}_{ij}$ について $\boldsymbol{y}_{ij} = (y_{i1j}, ..., y_{ipj})^T$ の成分は4次モーメントが一様有界と仮定する．いま，デー

タは，p.d.f.(もしくは，p.f.)

$$f(\boldsymbol{x}) = \varepsilon_1 f_1(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \varepsilon_2 f_2(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2),\ \varepsilon_1 + \varepsilon_2 = 1\ (\varepsilon_i > 0) \tag{1}$$

をもつ混合分布からの標本であるとみなす．ただし，$f_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ は π_i の p.d.f.(もしくは，p.f.) とする．この母集団から n 個のデータを無作為に抽出し，データ行列 $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ を定義する．そのとき，$E(\boldsymbol{x}_i) = \varepsilon_1\boldsymbol{\mu}_1 + \varepsilon_2\boldsymbol{\mu}_2\ (= \boldsymbol{\mu})$，$\mathrm{Var}(\boldsymbol{x}_i) = \varepsilon_1\boldsymbol{\Sigma}_1 + \varepsilon_2\boldsymbol{\Sigma}_2 + \varepsilon_1\varepsilon_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\ (= \boldsymbol{\Sigma})$ である．ここで，$\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ とおく．$\boldsymbol{\Sigma}$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_p (\geq 0)$ とし，適当な直交行列 $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_p]$ で $\boldsymbol{\Sigma}$ を $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_p)$ と分解する．そのとき $\boldsymbol{X} - [\boldsymbol{\mu}, ..., \boldsymbol{\mu}] = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Z}$ とおき，$\boldsymbol{Z} = (z_{ij})$ と表記する．

標本共分散行列を $\boldsymbol{S} = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^T = (n-1)^{-1}\sum_{i=1}^n(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$ とする．ここで，$\overline{\boldsymbol{X}} = [\bar{\boldsymbol{x}}, ..., \bar{\boldsymbol{x}}]$，$\bar{\boldsymbol{x}} = \sum_{j=1}^n \boldsymbol{x}_j/n$ である．そのとき，$\boldsymbol{S}_D = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^T(\boldsymbol{X} - \overline{\boldsymbol{X}})$ を $\boldsymbol{S}$ と正の固有値を共有する双対標本共分散行列という．$\boldsymbol{S}_D$ の固有値を $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{n-1}$ とし，$\hat{\lambda}_j$ に対する固有ベクトルを $\hat{\boldsymbol{u}}_j$ として，スペクトル分解を $\boldsymbol{S}_D = \sum_{j=1}^{n-1}\hat{\lambda}_j\hat{\boldsymbol{u}}_j\hat{\boldsymbol{u}}_j^T$ で表す．Yata and Aoshima [13] と Ishii et al. [7] は，もしも $\boldsymbol{S}_D$ がある正則条件を満たせば，次のような幾何学的表現が得られることを示した．

$$\frac{n-1}{\mathrm{tr}(\boldsymbol{\Sigma})}\boldsymbol{S}_D \xrightarrow{P} \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n^T, \quad p \to \infty. \tag{2}$$

ここで，$\boldsymbol{I}_n$ は n 次単位行列であり，$\boldsymbol{1}_n = (1, ..., 1)^T$ である．(2) は，$p \to \infty$ のとき，$\boldsymbol{S}_D$ の固有ベクトル $\hat{\boldsymbol{u}}_i$, $i = 1, ..., n-1$ は $\boldsymbol{1}_n$ と直交するものの方向は一意に定まらず，一方，固有値は定まるものの互いの差異がなくなることを意味する．混合分布 (1) は，(2) を導くための正則条件を満たさない．そこで，本論文は，混合分布 (1) がもつ幾何学的表現を明らかにする．さらに，幾何学的表現に基づくクラスター分析法を提案し，その性能を理論的かつ数値的に示す．

2　混合分布の幾何学的表現とクラスター分析

高次元小標本データのクラスター分析について，Liu et al. [10] は母集団に正規分布を仮定して考え，Ahn et al. [2] は maximal data piling によるクラスター分析を考えた．一方で，Yata and Aoshima [12] は主成分分析 (PCA) に基づくクラスター分析を考えた．例えば，Yata and Aoshima [12] は，幾つかの正則条件のもと，第 1 主成分スコア $s_{1j}\ (= \sqrt{\lambda_1}z_{1j})$, $j = 1, ..., n$ について $p \to \infty$ のとき

$$\frac{s_{1j}}{\sqrt{\lambda_1}} = \begin{cases} \sqrt{\varepsilon_2/\varepsilon_1} + o_P(1), & \boldsymbol{x}_j \in \pi_1 \text{ のとき}, \\ -\sqrt{\varepsilon_1/\varepsilon_2} + o_P(1), & \boldsymbol{x}_j \in \pi_2 \text{ のとき} \end{cases} \tag{3}$$

なることを示した．つまり，第 1 主成分スコアを精度よく推定できれば，その符号から高次元データを分類することができる．

本節では，高次元小標本における幾何学的表現の観点から，クラスター分析を考える．まず，$A_i=\{j|\ \boldsymbol{x}_j\in\pi_i,\ j=1,...,n\}$, $i=1,2$ とおき，$n_i=\#A_i$ とおく．ここで，$\#S$ は集合 S の要素の個数を表し，$n_1+n_2=n$ となる．さらに，各 j で $\boldsymbol{x}_j\in\pi_i$ ならば $r_j=(-1)^{i+1}n_{i'}/n$ $(i'\neq i)$ とおく．次のような条件を考える：

(A-i) $$\frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{\mathrm{tr}(\boldsymbol{\Sigma}_i)^2}=\frac{\sum_{s=1}^p\lambda_{is}^2}{(\sum_{s=1}^p\lambda_{is})^2}\to 0,\quad p\to\infty,\ i=1,2;$$

(A-ii) $$\frac{\sum_{r,s}^p\lambda_{ir}\lambda_{is}E\{(y_{irk}^2-1)(y_{isk}^2-1)|\boldsymbol{x}_k\in\pi_i\}}{(\sum_{s=1}^p\lambda_{is})^2}\to 0,\quad p\to\infty,\ i=1,2.$$

注意 1. (A-i) は，"$\lambda_{i1}/\mathrm{tr}(\boldsymbol{\Sigma}_i)\to 0,\ p\to\infty,\ i=1,2$" なる固有値に関する条件と同値である．各 π_i が正規分布に従うならば，(A-i) のもと (A-ii) が成り立つ．

このとき，次の定理が成り立つ．

定理 1. (A-i) と (A-ii) を満たすと仮定する．もし，$\lim_{p\to\infty}\Delta/\mathrm{tr}(\boldsymbol{\Sigma})\to c\ (\geq 0)$ かつ $\lim_{p\to\infty}\{\mathrm{tr}(\boldsymbol{\Sigma}_1)-\mathrm{tr}(\boldsymbol{\Sigma}_2)\}/\mathrm{tr}(\boldsymbol{\Sigma})=0$ ならば，次が成り立つ．

$$\frac{n-1}{\mathrm{tr}(\boldsymbol{\Sigma})}\boldsymbol{S}_D\xrightarrow{P}(1-\varepsilon_1\varepsilon_2 c)\Big(\boldsymbol{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\Big)+c\boldsymbol{r}\boldsymbol{r}^T,\quad p\to\infty.$$

ただし，$\boldsymbol{r}=(r_1,...,r_n)^T$ である．

上記の幾何学的表現から，以下の結果を得る．

定理 2. (A-i) と (A-ii) を満たすと仮定する．もし，$\liminf_{p\to\infty}\Delta/\mathrm{tr}(\boldsymbol{\Sigma})>0$ かつ $\lim_{p\to\infty}\{\mathrm{tr}(\boldsymbol{\Sigma}_1)-\mathrm{tr}(\boldsymbol{\Sigma}_2)\}/\mathrm{tr}(\boldsymbol{\Sigma})=0$ ならば，次が成り立つ．

$$\sqrt{\frac{n_1n_2}{n}}\hat{\boldsymbol{u}}_1\xrightarrow{P}\boldsymbol{r},\quad p\to\infty.$$

定理 2 で得られた幾何学的表現は，(2) とは異なり，$p\to\infty$ のとき $\boldsymbol{S}_D$ の (第一) 固有ベクトル $\hat{\boldsymbol{u}}_1$ の方向が一意に定まることを意味する．さらに，極限値 $\boldsymbol{r}$ の各成分の符号がデータを識別する．つまり，$\hat{\boldsymbol{u}}_1$ の各成分の正負によって混合データを分類する方法が考えられる．実際に，次の系が成り立つ．

系 1. $\hat{\boldsymbol{u}}_1=(\hat{u}_{11},...,\hat{u}_{1n})^T$ とおく．(A-i) と (A-ii) を満たすと仮定する．もし，$\liminf_{p\to\infty}\Delta/\mathrm{tr}(\boldsymbol{\Sigma})>0$ かつ $\lim_{p\to\infty}\{\mathrm{tr}(\boldsymbol{\Sigma}_1)-\mathrm{tr}(\boldsymbol{\Sigma}_2)\}/\mathrm{tr}(\boldsymbol{\Sigma})=0$ ならば，$\boldsymbol{x}_j\in\pi_i$ $(n_i>0;\ i=1,2),\ j=1,...,n$ について，$p\to\infty$ のとき次が成り立つ．

$$\hat{u}_{1j}=\begin{cases}\sqrt{n_2/(n_1n)}+o_P(1), & \boldsymbol{x}_j\in\pi_1\text{ のとき},\\ -\sqrt{n_1/(n_2n)}+o_P(1), & \boldsymbol{x}_j\in\pi_2\text{ のとき}.\end{cases}$$

注意. 従来型 PCA の第一主成分スコア s_{1j} の推定量は $\hat{u}_{1j}\sqrt{\hat{\lambda}_1 n}$ $(=\hat{s}_{1j}$ とおく$)$ で与えられる. 系 1 より, もし (A-i) と (A-ii) を満たし, $\liminf_{p\to\infty}\Delta/\mathrm{tr}(\boldsymbol{\Sigma})>0$ かつ $\lim_{p\to\infty}\{\mathrm{tr}(\boldsymbol{\Sigma}_1)-\mathrm{tr}(\boldsymbol{\Sigma}_2)\}/\mathrm{tr}(\boldsymbol{\Sigma})=0$ ならば, $\boldsymbol{x}_j\in\pi_i$ $(n_i>0;\ i=1,2),\ j=1,...,n$ について, $p\to\infty$ のとき次が成り立つ.

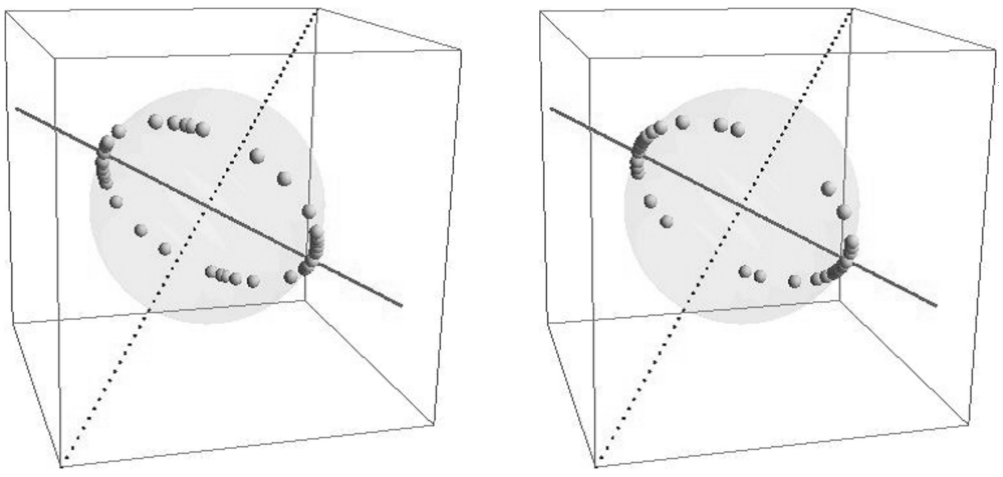
$$\frac{\hat{s}_{1j}}{\sqrt{\hat{\lambda}_1}}=\hat{u}_{1j}\sqrt{n}=\begin{cases}\sqrt{n_2/n_1}+o_P(1), & \boldsymbol{x}_j\in\pi_1\ \text{のとき},\\ -\sqrt{n_1/n_2}+o_P(1), & \boldsymbol{x}_j\in\pi_2\ \text{のとき}.\end{cases}\tag{4}$$

これは, (3) の一致推定になっている. なお, 従来型 PCA による主成分スコアの高次元小標本における漸近的性質は, Yata and Aoshima [11, 14] を参照のこと.

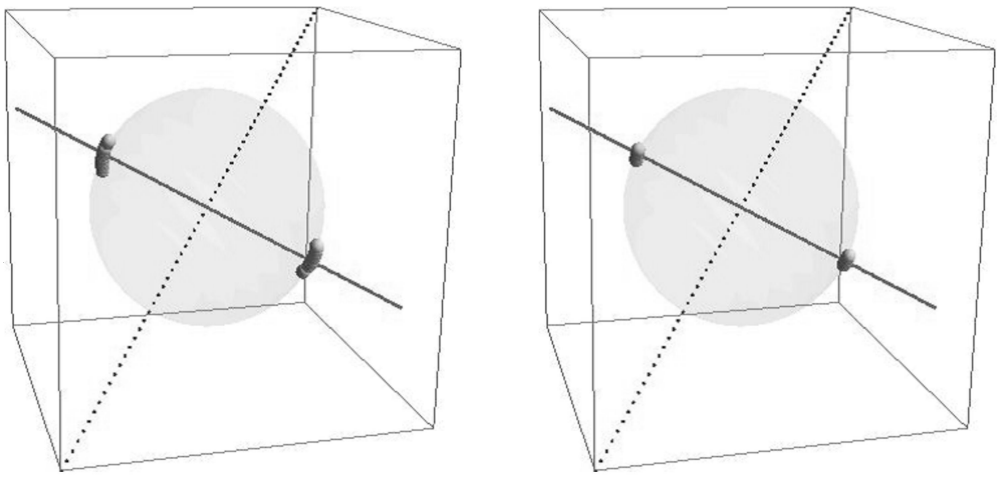
注意. $\lim_{p\to\infty}\{\mathrm{tr}(\boldsymbol{\Sigma}_1)-\mathrm{tr}(\boldsymbol{\Sigma}_2)\}/\mathrm{tr}(\boldsymbol{\Sigma})=0$ を満たさない場合については, Kurishita et al. [9] が $p\to\infty$ のとき $\mathrm{tr}(\boldsymbol{\Sigma}_1)$ と $\mathrm{tr}(\boldsymbol{\Sigma}_2)$ の差異を利用したクラスター分析法を与えた.

3 シミュレーション

定理 2 で与えた幾何学的表現を確認する. 混合分布 (1) を以下のように設定した. $\pi_i: N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i),\ i=1,2$ とし, $\boldsymbol{\mu}_1=\boldsymbol{0},\ \boldsymbol{\mu}_2=(1,...,1)^T,\ \boldsymbol{\Sigma}_1=(0.3^{|i-j|^{1/3}}),$ $\boldsymbol{\Sigma}_2=\boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B},\ \boldsymbol{B}=\mathrm{diag}[\{0.5+1/(p+1)\}^{1/2},...,\{0.5+p/(p+1)\}^{1/2}]$ とした. 混合比率は, $\varepsilon_1=1/3,\ \varepsilon_2=2/3$ とした. このとき, (A-i), (A-ii) と条件 $\liminf_{p\to\infty}\Delta/\mathrm{tr}(\boldsymbol{\Sigma})>0$ と $\lim_{p\to\infty}\{\mathrm{tr}(\boldsymbol{\Sigma}_1)-\mathrm{tr}(\boldsymbol{\Sigma}_2)\}/\mathrm{tr}(\boldsymbol{\Sigma})=0$ を満たす. 標本数は $n=3$ とした. 混合分布からのデータは, $\boldsymbol{x}_1\in\pi_1,\ \boldsymbol{x}_2,\boldsymbol{x}_3\in\pi_2$ $(n_1=1,\ n_2=2)$ とし, 各 π_i からランダムに標本を発生させた. 次元数が $p=4,\ 40,\ 400,\ 4000$ の各場合で, $\pm\hat{\boldsymbol{u}}_1$ の対を独立に 20 組発生させて, $\hat{\boldsymbol{u}}_1$ の出力結果を ● で表示した. 図 1(a)-(d) は, その結果を纏めたものである.



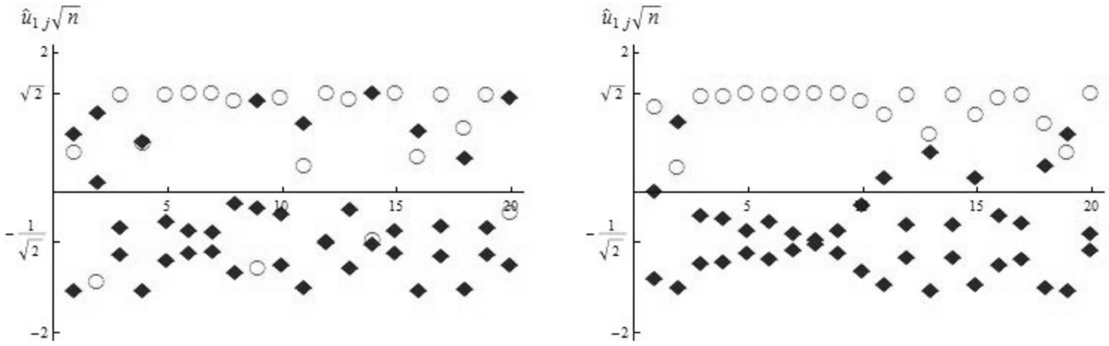
(a) $p=4$　　(b) $p=40$



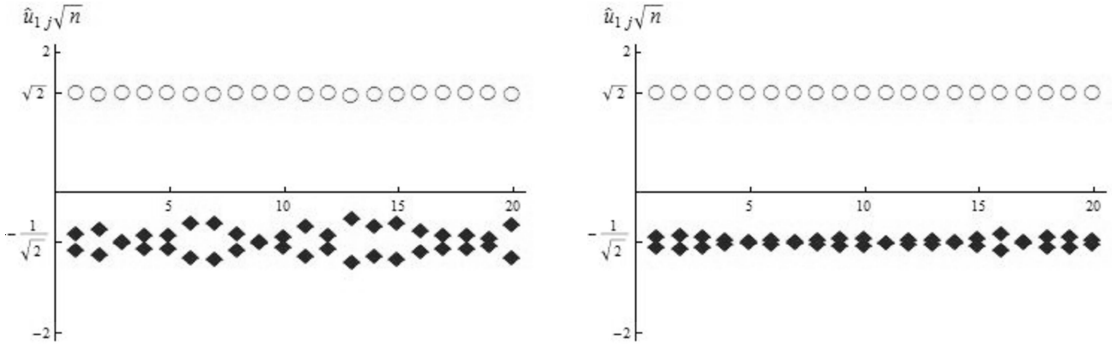
(c) $p = 400$　　(d) $p = 4000$

図 1. $\pi_i : N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$ の混合分布からの 20 組の $\pm\hat{\boldsymbol{u}}_1$ (●) の幾何学的表現．球は半径 1 の 3 次元球を表し，点線は $\mathbf{1}_n = (1, 1, 1)^T$，実線は $\boldsymbol{r} = (2/3, -1/3, -1/3)^T$ を表す．

次元数が増えるにつれ $\hat{\boldsymbol{u}}_1$ が $\boldsymbol{r}$ に収束していく様子が見てとれる．すなわち，第一主成分スコアの正負をもって混合データを分類することが可能になる．上記のシミュレーションで得た $\hat{u}_{1j}\sqrt{n}$, $j = 1, 2, 3$ を図 2(a)-(d) にプロットした．$\hat{u}_{1j}\sqrt{n}$, $j = 1, 2, 3$ の値を縦にプロットし，20 組の結果を横に並べた．○は $\boldsymbol{x}_j \in \pi_1$ の場合を，◆は $\boldsymbol{x}_j \in \pi_2$ の場合を表す．(4) の理論結果の通り，次元数が増えるにつれて，$\boldsymbol{x}_j \in \pi_1$ の場合は $\sqrt{n_2/n_1}$ $(= \sqrt{2})$ に収束し，$\boldsymbol{x}_j \in \pi_2$ の場合は $-\sqrt{n_1/n_2}$ $(= -1/\sqrt{2})$ に収束し，両者は完全に分離していく様子が見てとれる．



(a) $p = 4$　　(b) $p = 40$



(c) $p = 400$ (d) $p = 4000$

図 2. 図 1 のシミュレーションで得た $\hat{u}_{1j}\sqrt{n},\ j = 1, 2, 3$ の 20 組のプロット．$\hat{u}_{1j}\sqrt{n},\ j = 1, 2, 3$ の値を縦にプロットし，20 組の結果を横に並べた．○は $\boldsymbol{x}_j \in \pi_1$ の場合を表し，◆は $\boldsymbol{x}_j \in \pi_2$ の場合を表す．

A 付録

定理 1 の証明． (A-ii) のもと次が成り立つ．

$$\begin{aligned}\mathrm{Var}\{||\boldsymbol{x}_k - \boldsymbol{\mu}_i||^2 - \mathrm{tr}(\boldsymbol{\Sigma}_i)|\boldsymbol{x}_k \in \pi_i\} &= \sum_{r,s}^{p} \lambda_{ir}\lambda_{is} E\{(y_{irk}^2 - 1)(y_{isk}^2 - 1)|\boldsymbol{x}_k \in \pi_i\} \\ &= o\{\mathrm{tr}(\boldsymbol{\Sigma}_i)^2\},\ i = 1, 2. \end{aligned} \tag{5}$$

さらに，$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \leq \Delta\lambda_{i1} \leq \Delta\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2},\ i = 1, 2$ なることに注意すると，(A-i) のもと次が成り立つ．

$$\begin{aligned}\mathrm{Var}\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{x}_k - \boldsymbol{\mu}_i)|\boldsymbol{x}_k \in \pi_i\} &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &\leq \Delta\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/2} = o\{\Delta\mathrm{tr}(\boldsymbol{\Sigma}_i)\},\ i = 1, 2. \end{aligned} \tag{6}$$

いま，$\eta_i = n_i/n,\ i = 1, 2$ とおく．そのとき，$\boldsymbol{x}_k - \eta_1\boldsymbol{\mu}_1 - \eta_2\boldsymbol{\mu}_2 = \boldsymbol{x}_k - \boldsymbol{\mu}_i + (-1)^{i+1}(1 - \eta_i)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ と書ける．ここで，$\{\Delta\mathrm{tr}(\boldsymbol{\Sigma}_i)\}^{1/2} \leq \{\Delta + \mathrm{tr}(\boldsymbol{\Sigma}_i)\}/2$ と $\limsup_{p\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}_i)/\mathrm{tr}(\boldsymbol{\Sigma}) < \infty,\ i = 1, 2$ なることに注意する．(5) と (6) とチェビシェフの不等式から，$\boldsymbol{x}_k \in \pi_i$ のとき，(A-i) と (A-ii) のもと次が成り立つ．

$$||\boldsymbol{x}_k - \eta_1\boldsymbol{\mu}_1 - \eta_2\boldsymbol{\mu}_1||^2 = \mathrm{tr}(\boldsymbol{\Sigma}_i) + (1 - \eta_i)^2\Delta + o_p\{\Delta + \mathrm{tr}(\boldsymbol{\Sigma})\}. \tag{7}$$

さらに，$k \neq k'$ のとき次が成り立つ．

$$\mathrm{Var}\{(\boldsymbol{x}_k - \boldsymbol{\mu}_i)^T(\boldsymbol{x}_{k'} - \boldsymbol{\mu}_{i'})|\boldsymbol{x}_k \in \pi_i, \boldsymbol{x}_{k'} \in \pi_{i'}\} = \mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_{i'}). \tag{8}$$

このとき，$\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) \leq \{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\mathrm{tr}(\boldsymbol{\Sigma}_2^2)\}^{1/2} \leq \{\mathrm{tr}(\boldsymbol{\Sigma}_1^2) + \mathrm{tr}(\boldsymbol{\Sigma}_2^2)\}/2$ に注意する．(6) と (8) とチェビシェフの不等式から，$\boldsymbol{x}_k \in \pi_i$ と $\boldsymbol{x}_{k'} \in \pi_{i'}\ (k \neq k')$ のとき，(A-i) と

(A-ii) のもと次が成り立つ.

$$(\boldsymbol{x}_k - \eta_1\boldsymbol{\mu}_1 - \eta_2\boldsymbol{\mu}_2)^T(\boldsymbol{x}_{k'} - \eta_1\boldsymbol{\mu}_1 - \eta_2\boldsymbol{\mu}_2)$$
$$= (-1)^{i+i'}(1-\eta_i)(1-\eta_{i'})\Delta + o_p\{\Delta + \mathrm{tr}(\boldsymbol{\Sigma})\}. \tag{9}$$

ここで, $\boldsymbol{\mu}_0 = \eta_1\boldsymbol{\mu}_1 + \eta_2\boldsymbol{\mu}_2$ とおく. (7) と (9) より, $\lim_{p\to\infty}\Delta/\mathrm{tr}(\boldsymbol{\Sigma}) \to c\ (\geq 0)$ かつ $\lim_{p\to\infty}\{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}/\mathrm{tr}(\boldsymbol{\Sigma}) = 0$ ならば, (A-i) と (A-ii) のもと次が成り立つ.

$$\frac{(\boldsymbol{X} - [\boldsymbol{\mu}_0, ..., \boldsymbol{\mu}_0])^T(\boldsymbol{X} - [\boldsymbol{\mu}_0, ..., \boldsymbol{\mu}_0])}{\mathrm{tr}(\boldsymbol{\Sigma})} \xrightarrow{P} (1 - \varepsilon_1\varepsilon_2 c)\boldsymbol{I}_n + c\boldsymbol{r}\boldsymbol{r}^T. \tag{10}$$

ここで, $(\boldsymbol{X} - [\boldsymbol{\mu}_0, ..., \boldsymbol{\mu}_0])(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^T/n) = \boldsymbol{X} - \overline{\boldsymbol{X}}$ かつ $\mathbf{1}_n^T\boldsymbol{r} = 0$ なることに注意する. (10) より, $\lim_{p\to\infty}\Delta/\mathrm{tr}(\boldsymbol{\Sigma}) \to c\ (\geq 0)$ かつ $\lim_{p\to\infty}\{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}/\mathrm{tr}(\boldsymbol{\Sigma}) = 0$ ならば, (A-i) と (A-ii) のもと次が成り立つ.

$$\frac{n-1}{\mathrm{tr}(\boldsymbol{\Sigma})}\boldsymbol{S}_D = \frac{(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^T/n)(\boldsymbol{X} - [\boldsymbol{\mu}_0, ..., \boldsymbol{\mu}_0])^T(\boldsymbol{X} - [\boldsymbol{\mu}_0, ..., \boldsymbol{\mu}_0])(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^T/n)}{\mathrm{tr}(\boldsymbol{\Sigma})}$$
$$\xrightarrow{P} (1 - \varepsilon_1\varepsilon_2 c)(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^T/n) + c\boldsymbol{r}\boldsymbol{r}^T.$$

それゆえ, 結果を得る. □

定理 2 の証明. 定理 1 の証明と同様にして, $\liminf_{p\to\infty}\Delta/\mathrm{tr}(\boldsymbol{\Sigma}) > 0$ かつ $\lim_{p\to\infty}\{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}/\mathrm{tr}(\boldsymbol{\Sigma}) = 0$ ならば, (A-i) と (A-ii) のもと次が成り立つ.

$$\frac{n-1}{\Delta}\boldsymbol{S}_D = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1)}{\Delta}(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n^T/n) + \boldsymbol{r}\boldsymbol{r}^T + \boldsymbol{o}_p(1).$$

ここで, $\hat{\boldsymbol{u}}_1^T\mathbf{1}_n = 0\ (i = 1, ..., n-1)$ なることに注意すれば, $\boldsymbol{r} \neq 0$ のとき $\hat{\boldsymbol{u}}_1$ と $\boldsymbol{r}$ の向きは一致する. それゆえ, $||\boldsymbol{r}|| = \sqrt{n_1 n_2/n}$ に注意すれば, 結果を得る. □

系 1 の証明. $n_i > 0,\ i = 1, 2$ のとき $\boldsymbol{r} \neq 0$ となるので, 定理 2 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 26540010 研究代表者: 青嶋 誠「ビッグデータの統計学：理論の開拓と 3V への挑戦」, 若手研究 (B) 26800078 研究代表者: 矢田 和善「高次元漸近理論の統一的研究」から研究助成を受けています.

参考文献

[1] Ahn, J., Marron, J. S., Muller, K. M. and Chi Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. Biometrika, **94**, 760-766.

[2] Ahn, J., Lee, M. H. and Yoon, Y. J. (2012). Clustering high dimension, low sample size data using the maximal data piling distance. Statist. Sinica, **22**, 443-464.

[3] Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data. Sequential Anal. (*Editor's special invited paper*), **30**, 356-399.

[4] 青嶋　誠, 矢田和善 (2013a). 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] 青嶋　誠, 矢田和善 (2013b). 論説：高次元小標本における統計的推測. 数学, **65**, 225-247.

[6] Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. J. R. Stat. Soc. Ser. B Stat. Methodol., **67**, 427-444.

[7] Ishii, A., Yata, K. and Aoshima, M. (2014). Asymptotic distribution of the largest eigenvalue via geometric representations of high-dimension, low-sample-size data. Sri Lankan J. Appl. Statist., Special Issue: Modern Statistical Methodologies in the Cutting Edge of Science (ed. Mukhopadhyay, N.), accepted.

[8] Jung, S., Sen, A. and Marron, J. S. (2012). Boundary behavior in high dimension, low sample size asymptotics of PCA. J. Multivariate Anal., **109**, 190-203.

[9] Kurishita, K., Yata K. and Aoshima, M. (2012). Cluster analysis for high-dimensional data. Proceedings of Statistical Inference for High-Dimensional Data and Its Applications (ed. Sugiyama, K.), 11-20.

[10] Liu, Y. Hayes, D. N., Nobel, A. and Marron, J. S. (2008). Statistical significance of clustering for high-dimension, low-sample size data. J. Amer. Statist. Assoc., **103**, 1281-1293.

[11] Yata, K. and Aoshima, M. (2009). PCA consistency for non-Gaussian data in high dimension, low sample size context. Comm. Statist. Theory Methods, Special Issue Honoring Zacks, S. (ed. Mukhopadhyay, N.), **38**, 2634-2652.

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

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

[14] Yata, K. and Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings. J. Multivariate Anal., **122**, 334-354.