高次元現象の統計数理

青嶋　誠　(筑波大数理物質)*

1. はじめに

ゲノム科学・情報工学・金融工学などの現代科学の一つの特徴は，データがもつ次元数の膨大さにある．例えば，DNA マイクロアレイによる遺伝子発現データを見ると，次元数が数万にものぼる一方で，標本数は 100 にも満たないという事例が多い．高次元データは，高次元になるほど標本を入手することが難しく，一般にデータの次元数 p と標本数 n には

$$p \gg n \text{ (もしくは，} p > n\text{)}$$

といった大小関係がある．これを強調して，高次元小標本データとよぶこともある．2005 年頃から，次世代シーケンサーによる超高次元ゲノムデータも統計解析の対象となり，例えば Wang et al. [13] は，次元数が 300 万を超え，標本数は 40 というゲノムデータを扱っている．近年，高次元小標本データの解析は，様々な分野で需要が高まっている．本講演は，次世代シーケンサーによる超高次元ゲノムデータをはじめ，天文宇宙学における天体の分光マッピングによる波長データ ([11]) や，材料科学における結晶構造データなど，高次元小標本データの解析例を幾つか紹介する．

高次元データは，豊富な情報を有するものの，それが巨大なノイズに埋もれているために見つけ難い．伝統的な多変量解析法では，高次元データの解析に精度を保証することができない．次元数が標本数を超えると，次元の呪いとして様々な高次元現象が現れ，多変量解析の理論と方法論の成立を阻むのである．特に非スパース性という現象に陥ると，データの潜在情報は高次元において膨張するノイズに埋没してしまい，推測の強不一致という最悪な結果を招くことにもなる．

高次元データの解析には，高次元データ特有の理論と方法論が必要になる．著者の研究グループは，高次元統計解析と非スパースモデリングという新しい理論と方法論を考案した．本講演は，様々な高次元現象を説明し，たった数十程度の小標本であっても高速かつ高精度に解析するための，高次元現象の統計数理を紹介する．

2. 高次元データの非スパース性

1990 年代後半，高次元データを解析するための方法論として，機械学習が目覚ましく発展した．スタンフォード大学の Donoho 教授や Tibshirani 教授らは，スパースモデリングを発展させた．これは，高次元データの母共分散行列にスパース性を仮定して，高次元データの特徴を少数の変数や標本で捉えようとするもので，L1 正則化法や圧縮センシングがよく知られる ([7]). 近年，多層ニューラルネットワークを用いた深層学習が盛んに研究されている．これもスパース性を仮定して，データは大標本であること

本研究は，筑波大学の矢田和善准教授との共同研究である．本研究は，科学研究費補助金 基盤研究 (A) 20H00576　研究代表者: 青嶋 誠「大規模複雑データの理論と方法論の革新的展開」，および，学術研究助成基金助成金 挑戦的研究 (萌芽) 22K19769　研究代表者: 青嶋 誠「テンソル構造をもつ巨大データの統計的圧縮技術の開発」から助成を受けたものである．
2010 Mathematics Subject Classification: 62H25, 62H30
キーワード：高次元統計解析, 非スパースモデリング

* e-mail: aoshima@math.tsukuba.ac.jp
web: https://www.math.tsukuba.ac.jp/~aoshima-lab/jp/

が前提になっている．ところで，高次元データの母共分散行列に，スパース性は仮定できるのだろうか．

2.1. 高次元共分散行列の非スパース性

p次非負定値対称行列$\mathbf{\Sigma}$をp次元データの母共分散行列とし，その固有値を$\lambda_1 \geq \cdots \geq \lambda_p\ (\geq 0)$とする．次の3つの高次元データセットについて，それぞれの固有値を上から10個求めてみる．第1のデータセットは，Takeuchi et al. [11]による天体の分光マッピングによる波長データで，標本数は231地点，スペクトルは1971次元である．第2のデータセットは，von Roemeling et al. [12]による2群からなる淡明細胞型腎細胞癌(ccRCC)のマイクロアレイデータで，ccRCC群は71サンプル，健常群 (Normal) は72サンプルで，2群とも遺伝子数(プローブ数)が54675次元である．第3のデータセットは，Wang et al. [13]の次世代シーケンサーによる2群からなる加齢黄斑変性 (AMD) のゲノムデータで，AMD群は21サンプル，健常群(Normal)は19サンプルで，2群ともゲノム領域数が3095656次元である．図1は，寄与率$\lambda_j/\mathrm{tr}(\mathbf{\Sigma})$について，Yata and Aoshima [15]のノイズ掃き出し法による推定値$\tilde{\lambda}_j/\mathrm{tr}(\boldsymbol{S})$をプロットしたものである[1]．どの高次元データも，最初の幾つかの固有値が飛び抜けて大きく，他の固有値は小さなものであることが見てとれる．特に，第1固有値は強くスパイクしており，一般に$\mathrm{tr}(\mathbf{\Sigma})/p \to c\ (>0)$となることを考えれば，$\lambda_1$は$O(p)$のスピードで発散することが分かる．

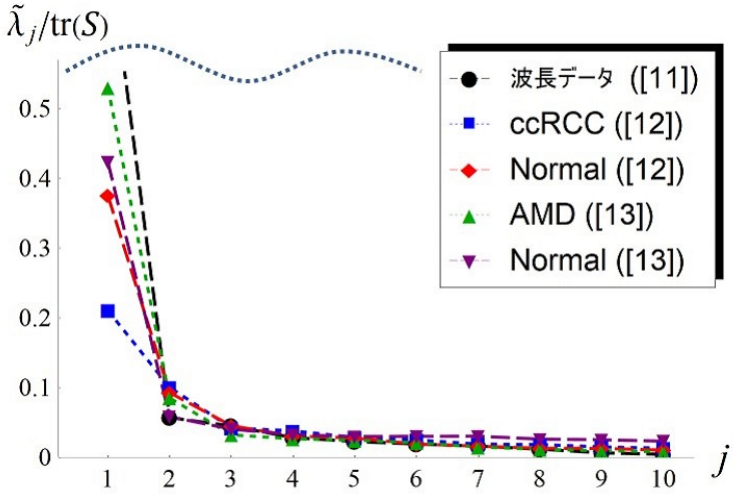

図 *1.* 天体の分光マッピングによる波長データ ($p = 1971$)，2つの群のマイクロアレイデータ ($p = 54675$)，2つの群の次世代シーケンサーによるゲノムデータ ($p = 3095656$). それぞれの高次元データについて，第10寄与率まで推定値をプロットした．

高次元データの母共分散行列の固有値に対して，Johnstone [9] や Paul [10] 等は，ランダム行列理論から標本固有値の漸近的性質を導出した[2]．特に，スタンフォード大学の Johnstone 教授は，Tracy-Widom 分布に基づく標本固有値の極限分布を与えた ([9]).

---

[1] $\tilde{\lambda}_j$ は (4.2) 式で与えられるもの，$\boldsymbol{S}$ は標本共分散行列である．後述の (A-i) と (A-ii) のもと，$\tilde{\lambda}_j/\mathrm{tr}(\boldsymbol{S}) = \lambda_j\{1 + o_P(1)\}/\mathrm{tr}(\mathbf{\Sigma})$ $(p \to \infty,\ n \to \infty)$ なる一致性をもつ．

[2] これらの研究は，次元数pと標本数nが$n/p \to c\ (>0)$なる場合を考えている．高次元小標本データ $(n/p \to 0)$ には対応していないことに注意する．

そのとき，$\boldsymbol{\Sigma}$の固有値には，次のようなスパイクモデルが仮定されていた．

$$\begin{aligned} &\lambda_1, ..., \lambda_t \text{は1よりも大きく，次元数} p \text{には依存しない定数} \\ &\lambda_{t+1} = \cdots = \lambda_p = 1 \end{aligned} \tag{2.1}$$

(2.1)のもとでは，$\|\boldsymbol{\Sigma} - \boldsymbol{I}_p\|_F^2 = \sum_{s=1}^{t}(\lambda_s - 1)^2 = O(1)\ (p \to \infty)$となり[3]，$p \times p$行列$\boldsymbol{\Sigma}$と$p$次単位行列$\boldsymbol{I}_p$の距離が有界になる．つまり，(2.1)は，$\boldsymbol{\Sigma}$の非対角成分の殆どが零である，といった非常にスパースな共分散構造を想定していることになる．スパースモデリングでは，(2.1)と同様な「λ_1は次元数pに関して有界である」という条件が仮定される．図1でも見たが，固有値が次元数pに依存しないという仮定は，実データがもつ固有値構造から大きく乖離する．それでは，実データの固有値構造に飛び抜けて大きな固有値が現れる原因は何か．主たる原因は，高次元データを構成する変数間の相関にある．実データに対する$\boldsymbol{\Sigma}$は，非対角成分の多くが非零となり，非スパースな共分散構造をもつのである．困ったことに，非スパースな共分散構造は高次元において巨大なノイズを生み，潜在情報を埋もれさせてしまう．これが，高次元データの解析を困難にしている．このような高次元データに，ランダム行列理論・スパースモデリング・深層学習などを，そのまま使用することは望ましくない．そこで，高次元データが非スパースな共分散構造をもつことを考慮して，新たに開発された方法論が非スパースモデリングである．

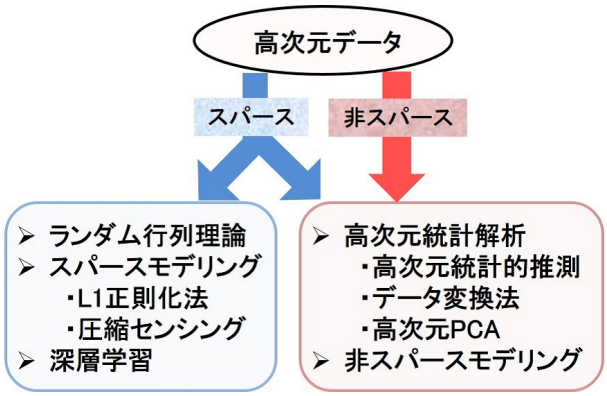

図 2. 高次元データの解析: スパース性と非スパース性.

2.2. 強スパイク固有値モデル

高次元データの母共分散行列$\boldsymbol{\Sigma}$は，次元数の増加とともに行列のサイズが大きくなり，図1で見たように最初の幾つかの固有値は次元数に依存して大きなものとなる．特に，最大固有値λ_1は，次元数の冪関数として捉えることが自然である．Aoshima and Yata [4]は，高次元データの固有値モデルを2つに分類した．一つ目は，強スパイク固有値モデル (strongly spiked eigenvalue (SSE) モデル) とよばれ，次のように定義される．

$$\liminf_{p\to\infty} \left\{ \frac{\lambda_1^2}{\mathrm{tr}(\boldsymbol{\Sigma}^2)} \right\} > 0 \tag{2.2}$$

いま一つは，弱スパイク固有値モデル (NSSE モデル) とよばれ，次のように定義される．

$$\frac{\lambda_1^2}{\mathrm{tr}(\boldsymbol{\Sigma}^2)} \to 0\ \ (p \to \infty) \tag{2.3}$$

[3] $\|\cdot\|_F$はフロベニウスノルムを表す.

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．簡単のため，以下のような固有値構造を考えてみる．

$$\lambda_s = c_s p^{\alpha_s}\ (s = 1, ..., t), \quad \lambda_s = c_s\ (s = t+1, ..., p) \tag{2.4}$$

ここで，$c_s > 0,\ \alpha_1 \geq \cdots \geq \alpha_t > 0$ は次元数 p に依存しない未知の実数，t は次元数 p に依存しない未知の自然数である．(2.4) は，$\alpha_1 \geq 0.5$ のとき強スパイク固有値モデル (2.2) の一つとなり，$\alpha_1 < 0.5$ のとき弱スパイク固有値モデル (2.3) の一つとなる．

後述するが，弱スパイク固有値モデルにおいては種々の統計量に高次元漸近正規性という望ましい性質が得られるが，強スパイク固有値モデルでは高次元漸近正規性が成り立たない．それどころか，強くスパイクする固有値が一つでもあると，それが巨大なノイズとなって潜在情報を覆い隠してしまう．そのとき，強スパイク固有値モデルは，非対角成分の多くが非零といった非スパースな $\boldsymbol{\Sigma}$ を想定したものになる．(2.2) は母共分散行列に関する境界条件になっており，SSE モデルは非スパース性に対応し，NSSE モデルはスパース性に対応する．

3. 高次元漸近正規性

平均が p 次ベクトル $\boldsymbol{\mu}$，共分散行列が $\boldsymbol{\Sigma}$ の母集団から，$n\ (\geq 3)$ 個のデータベクトル $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ を無作為に抽出する．それらを並べて，$p \times n$ のデータ行列を $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ 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．そのとき，$\boldsymbol{X} - [\boldsymbol{\mu}, ..., \boldsymbol{\mu}] = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Z}$ となる $\boldsymbol{Z}$ を $\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_p]^T$，$\boldsymbol{z}_s = (z_{1s}, ..., z_{ns})^T$ と表記する．$\boldsymbol{Z}$ の成分は4次モーメントが一様有界であると仮定し，さらに，必要な箇所で，適宜，次を仮定する．

(A-i) $E(z_{\ell s}^2 z_{\ell t}^2) = 1$，$E(z_{\ell s} z_{\ell t} z_{\ell s'}) = 0$，$E(z_{\ell s} z_{\ell t} z_{\ell s'} z_{\ell t'}) = 0$，$s \neq t, s', t'$

(A-i) は，母集団の分布について，正規分布を緩めた仮定になっている．

母平均 $\boldsymbol{\mu}$ に関する推測を考える．そのとき，共分散行列 $\boldsymbol{\Sigma}$ はノイズとなり，前述の SSE モデルと NSSE モデルはノイズモデルになる．標本平均ベクトル $\overline{\boldsymbol{x}} = \sum_{\ell=1}^n \boldsymbol{x}_\ell / n$ について，平均 2 乗誤差は $E(\|\overline{\boldsymbol{x}} - \boldsymbol{\mu}\|^2) = \mathrm{tr}(\boldsymbol{\Sigma})/n$ となる．ここで，$\|\cdot\|$ はユークリッドノルムである．一般に $\mathrm{tr}(\boldsymbol{\Sigma})/p \to c\ (> 0)$ となることを考えれば，高次元小標本 $(p/n \to \infty)$ において

$$E(\|\overline{\boldsymbol{x}} - \boldsymbol{\mu}\|^2) \to \infty$$

となり，平均2乗誤差は発散してしまう．すなわち，次元数の増加とともに標本平均に含まれるノイズが巨大化し，$\overline{\boldsymbol{x}}$ で $\boldsymbol{\mu}$ を推定することは困難になる．ところで，$\|\overline{\boldsymbol{x}} - \boldsymbol{\mu}\|^2$ の分散は，(A-i) のもと $p \to \infty$ のとき

$$\mathrm{Var}(\|\overline{\boldsymbol{x}} - \boldsymbol{\mu}\|^2) = 2\frac{\mathrm{tr}(\boldsymbol{\Sigma}^2)(n-1)}{n^3} + \sum_{s=1}^p \frac{\lambda_s^2 \mathrm{Var}(z_{\ell s}^2)}{n^3} = 2\frac{\mathrm{tr}(\boldsymbol{\Sigma}^2)}{n^2} + O\left(\frac{\mathrm{tr}(\boldsymbol{\Sigma}^2)}{n^3}\right)$$

となる．さらに，球形条件[4]“$\mathrm{tr}(\boldsymbol{\Sigma}^2)/\{\mathrm{tr}(\boldsymbol{\Sigma})\}^2 \to 0\ \ (p \to \infty)$”を満たすと仮定すると，$p \to \infty$ のとき (n は固定したままでも) 次が成り立つ．

$$\|\overline{\boldsymbol{x}} - \boldsymbol{\mu}\| = \sqrt{\mathrm{tr}(\boldsymbol{\Sigma})/n}\{1 + o_P(1)\}$$

[4] $\mathrm{tr}(\boldsymbol{\Sigma}^2)/\{\mathrm{tr}(\boldsymbol{\Sigma})\}^2 \leq \lambda_1/\mathrm{tr}(\boldsymbol{\Sigma}) \leq \lambda_1/\{\mathrm{tr}(\boldsymbol{\Sigma}^2)\}^{1/2}$ に注意すれば，NSSE モデル (2.3) は球形条件を満たす．

すなわち，高次元小標本 $(p/n \to \infty)$ において，$\overline{\boldsymbol{x}}$ は中心 $\boldsymbol{\mu}$，半径 $\sqrt{\mathrm{tr}(\boldsymbol{\Sigma})/n}$ の球面に収束する．図 3 は，単位行列 $\boldsymbol{I}_p$ を共分散行列にもつ p 次元正規分布 $N_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$ について，大きさ $n = 3$ の無作為標本による標本平均ベクトルを 200 回発生させ，$\overline{\boldsymbol{x}} - \boldsymbol{\mu}$ を固有空間 $(\hat{\boldsymbol{h}}_1, \hat{\boldsymbol{h}}_2, \hat{\boldsymbol{h}}_3$ で張られる空間) にプロットしたものである．ここで，$\hat{\boldsymbol{h}}_j$ は標本共分散行列の第 j 固有ベクトルを表す．次元数が $p = 4$ のとき (左図) は，半径 $\sqrt{p/n} = \sqrt{4/3}$ の球の周辺に点が散らばっている．一方，次元数が $p = 1000$ のとき (右図) は，ノイズが膨張して半径 $\sqrt{p/n} = \sqrt{1000/3}$ の球の表面に点が集中し，球面集中現象が見てとれる．

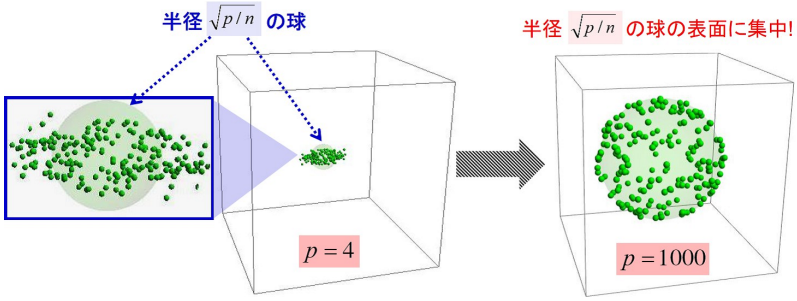

図 3. 高次元小標本における $\overline{\boldsymbol{x}}$ の球面集中現象．

球面での $\overline{\boldsymbol{x}}$ の確率変動について，Aoshima and Yata [1, 3] は，次のような高次元漸近正規性を証明した．

定理 1 ([3]). *(A-i)* を仮定する．*NSSE* モデルのもとで，$p \to \infty$, $n \to \infty$ のとき次が成り立つ[5]．

$$\frac{\|\overline{\boldsymbol{x}} - \boldsymbol{\mu}\|^2 - tr(\boldsymbol{\Sigma})/n}{\sqrt{2tr(\boldsymbol{\Sigma}^2)/n^2}} \xrightarrow{\mathcal{L}} N(0,1)$$

ここで，$\xrightarrow{\mathcal{L}}$ は分布収束を表す．

定理 1 は，高次元において，$\overline{\boldsymbol{x}}$ が中心 $\boldsymbol{\mu}$，半径 $\sqrt{\mathrm{tr}(\boldsymbol{\Sigma})/n}$ の球面の周りに正規分布で変動することを意味する．このことから，NSSE モデルであれば，球面付近の微小な確率変動を捉えることで統計的推測の精度を保証することが可能となる．しかしながら，SSE モデルの場合，これが上手くいかない．強くスパイクする (非スパースな) ノイズが悪さをして，高次元漸近正規性は成立しない．いま，$p = 1000$ のとき，NSSE モデルとして $\boldsymbol{\Sigma} = \mathrm{diag}(p^{1/3}, 1, ..., 1)$ を，SSE モデルとして $\boldsymbol{\Sigma} = \mathrm{diag}(p, 1, ..., 1)$ を考えてみる．図 4 は，p 次元正規分布 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ について，大きさ $n = 10$ の無作為標本による標本平均ベクトルをもとに $U = \{\|\overline{\boldsymbol{x}} - \boldsymbol{\mu}\|^2 - \mathrm{tr}(\boldsymbol{\Sigma})/n\}/\sqrt{2\mathrm{tr}(\boldsymbol{\Sigma}^2)/n^2}$ を計算し，これを 2000 回発生させてヒストグラムを作成したものである．定理 1 の通り，NSSE モデルにおけるヒストグラムの形状は $N(0,1)$ に近いことが確認できる．一方，SSE モデルにおいては，高次元漸近正規性が成立しない．さらに，SSE モデルにおいては $\mathrm{tr}(\boldsymbol{\Sigma}^2) = O(\lambda_1^2)$ となるので，図 1 のように最大固有値が飛び抜けている場合，$\|\overline{\boldsymbol{x}} - \boldsymbol{\mu}\|^2$ の分散が非常に大きくなる．つまり，SSE モデルが当てはまる状況では，高次元漸近正規性が壊れるだけでなく，推測の精度も著しく悪くなるのである．このことから，NSSE モデルを想定して (つまり，ノイズのスパース性を想定して) 構築された統計的推測法は，SSE

[5] 本講演で $n \to \infty$ と表記するとき，$n = \log p$ 程度で十分である．つまり，高次元小標本であっても，結果は保証される．

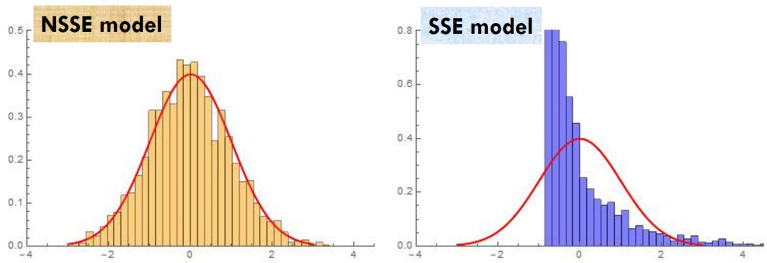

図 4. NSSE モデル $\boldsymbol{\Sigma} = \mathrm{diag}(p^{1/3}, 1, ..., 1)$ と SSE モデル $\boldsymbol{\Sigma} = \mathrm{diag}(p, 1, ..., 1)$ の場合の，球面付近の $\overline{\boldsymbol{x}}$ の漸近的挙動．U $(p = 1000, n = 10)$ を 2000 回発生させ，ヒストグラムを作成した．実線は $N(0, 1)$ の確率密度関数を表す．

モデルに対しては精度を保証することができず，それゆえ SSE モデルに使うわけにはいかない．それでは，SSE モデルも考慮した統計的推測法を構築することはできないか．Aoshima and Yata [4] は，SSE モデルの場合は自動的に NSSE モデルに変換して推測を行うデータ変換法を考案した．次節でこれを説明するが，表記を簡単にするために，ここで SSE モデルを次のように簡略化しておく．

(A-ii) (次元数 p に依存しない) ある自然数 k に対して，

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

(ii) $\liminf_{p\to\infty} \dfrac{\lambda_k^2}{\Psi_k} > 0$ かつ $\dfrac{\lambda_{k+1}^2}{\Psi_{k+1}} \to 0, p \to \infty$ (ただし，$\Psi_j = \sum_{s=j}^{p} \lambda_s^2$ である)

(2.4) の固有値構造であれば，$k(< t)$ に対して $\alpha_k \geq 0.5 > \alpha_{k+1}$ 且つ $1 \leq s < s' \leq k$ に対して $c_s \neq c_{s'}$ ならば，条件 (A-ii) を満たす．(A-ii) は，図 1 の実データ解析をヒントに，最初の k 個の固有値が飛び抜けて大きいことを表現した SSE モデルになっている．

4. データ変換法

Aoshima and Yata [4] は，次のようなデータ変換を考えた[6]．

$$\boldsymbol{x}_{\ell *} = \boldsymbol{A}\boldsymbol{x}_{\ell} \ \ (\ell = 1, ..., n) \quad (\text{ただし，} \boldsymbol{A} = \textstyle\sum_{s=k+1}^{p} \boldsymbol{h}_s \boldsymbol{h}_s^T \ (k \geq 0) \text{である}) \tag{4.1}$$

変換後のデータについて，$\mathrm{Var}(\boldsymbol{x}_{\ell *}) = \sum_{s=k+1}^{p} \lambda_s \boldsymbol{h}_s \boldsymbol{h}_s^T$ $(= \boldsymbol{\Sigma}_*$ とおく) となるので，$\lambda_{\max}(\boldsymbol{\Sigma}_*)$ を $\boldsymbol{\Sigma}_*$ の最大固有値とすれば，SSE モデル (A-ii) の条件 (ii) から

$$\frac{\{\lambda_{\max}(\boldsymbol{\Sigma}_*)\}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_*^2)} = \frac{\lambda_{k+1}^2}{\Psi_{k+1}} \to 0, \ \ p \to \infty$$

が成立する．つまり，変換後のデータ $\boldsymbol{x}_{\ell *}$ には，NSSE モデルが当てはまることになる．データ変換に使う正射影行列は $\boldsymbol{A} = \boldsymbol{I}_p - \sum_{s=1}^{k} \boldsymbol{h}_s \boldsymbol{h}_s^T$ と書けるので，強くスパイクする固有空間を如何に高精度に推定するかが鍵となる．本節では，高次元主成分分析 (PCA) として知られるノイズ掃き出し法を用いて固有値・固有ベクトルを推定する．

標本共分散行列

$$\boldsymbol{S} = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^T \quad (\text{ただし，} \overline{\boldsymbol{X}} = [\overline{\boldsymbol{x}}, ..., \overline{\boldsymbol{x}}] \text{である})$$

[6] 強くスパイクする固有空間の個数 k は未知である．k の推定は，Aoshima and Yata [4] の S2.2 節を参照のこと．なお，$k = 0$ の場合は恒等変換となり，データは変換せずとも NSSE モデルが当てはまる．

について，$\boldsymbol{S}$の固有値を$\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p\,(\geq 0)$，各固有値$\hat{\lambda}_j$に対する固有ベクトルを$\hat{\boldsymbol{h}}_j$とする．ただし，$\hat{\boldsymbol{h}}_1, ..., \hat{\boldsymbol{h}}_p$は$p$次元実ベクトル空間の正規直交基底をなすとする．$\boldsymbol{S}$の固有値分解は$\boldsymbol{S} = \sum_{s=1}^{p} \hat{\lambda}_s \hat{\boldsymbol{h}}_s \hat{\boldsymbol{h}}_s^T$と書ける．$\boldsymbol{S}_D = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^T(\boldsymbol{X} - \overline{\boldsymbol{X}})$とおくと，$n$次の対称行列$\boldsymbol{S}_D$は$\boldsymbol{S}$と正の固有値を共有する．$\boldsymbol{S}_D$の固有値を$\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{n-1}\,(\geq 0)$，各$\hat{\lambda}_j$に対する固有ベクトルを$\hat{\boldsymbol{u}}_j = (\hat{u}_{1j}, ..., \hat{u}_{nj})^T$とする．ただし，$\hat{\boldsymbol{u}}_1, ..., \hat{\boldsymbol{u}}_{n-1}$は互いに正規直交とする．$\boldsymbol{S}_D$の固有値分解は$\boldsymbol{S}_D = \sum_{s=1}^{n-1} \hat{\lambda}_s \hat{\boldsymbol{u}}_s \hat{\boldsymbol{u}}_s^T$と書ける．$\boldsymbol{S}$と$\boldsymbol{S}_D$の固有ベクトルには，次の関係がある．

$$\hat{\boldsymbol{h}}_j = \{(n-1)\hat{\lambda}_j\}^{-1/2}(\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_j \quad (j = 1, ..., n-1)$$

Yata and Aoshima [14]は，固有値構造が(2.4)の場合に，$\hat{\lambda}_j$, $\hat{\boldsymbol{h}}_j$, さらに主成分スコアについて，推定量の一致性(と不一致性)を明らかにした．Aoshima and Yata [4]と Yata and Aoshima [16]は，一般のSSEモデルの場合にその理論を拡張している．SSEモデル(A-ii)を使って$\hat{\lambda}_j$と$\hat{\boldsymbol{h}}_j$について簡単に述べると，次のようになる．

定理 2 ([4]). *(A-i)*を仮定する．*SSE*モデル*(A-ii)*のもとで，各$j\,(\leq k)$について，$p \to \infty$, $n \to \infty$のとき次が成り立つ．

$$\hat{\lambda}_j/\lambda_j = 1 + \delta_j + O_P(n^{-1/2}), \quad (\hat{\boldsymbol{h}}_j^T \boldsymbol{h}_j)^2 = (1+\delta_j)^{-1} + O_P(n^{-1/2})$$

ただし，$\delta_j = \lambda_j^{-1} \sum_{s=k+1}^{p} \lambda_s/(n-1)$である．

上記の通り，$\hat{\lambda}_j$と$\hat{\boldsymbol{h}}_j$は，δ_jというノイズをもつ．(2.4)の固有値構造であれば，$\alpha_j = 0.5$に対して$n = o(p^{0.5})$だとすると$\delta_j \to \infty$となり，$\lambda_j/\hat{\lambda}_j = o_P(1)$や$(\hat{\boldsymbol{h}}_j^T \boldsymbol{h}_j)^2 = o_P(1)$といった強不一致性を起こす．この問題を解決するために，Yata and Aoshima [15, 16]はノイズ掃き出し法という高次元におけるPCAを考案し，次のような固有値の推定量を与えた．

$$\tilde{\lambda}_j = \hat{\lambda}_j - \frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{s=1}^{j} \hat{\lambda}_s}{n-1-j} \quad (j = 1, ..., n-2) \tag{4.2}$$

固有ベクトルの推定量は次のようになる．

$$\tilde{\boldsymbol{h}}_j = \{(n-1)\tilde{\lambda}_j\}^{-1/2}(\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_j \quad (j = 1, ..., n-2) \tag{4.3}$$

Aoshima and Yata [4]と Yata and Aoshima [16]による次の定理は，$\delta_j \to \infty$となる場合にも，ノイズ掃き出し法による推定量が一致性をもつことを示している．

定理 3 ([4]). *(A-i)*を仮定する．*SSE*モデル*(A-ii)*のもとで，各$j\,(\leq k)$について，$p \to \infty$, $n \to \infty$のとき次が成り立つ．

$$\tilde{\lambda}_j/\lambda_j = 1 + O_P(n^{-1/2}), \quad (\tilde{\boldsymbol{h}}_j^T \boldsymbol{h}_j)^2 = 1 + O_P(n^{-1})$$

データ変換(4.1)を使うためには，$\boldsymbol{A}\boldsymbol{x}_\ell$を構成する$\boldsymbol{h}_j^T \boldsymbol{x}_\ell\,(= x_{\ell j}$とおく$), j = 1, ..., k$を推定する必要がある．Aoshima and Yata [4]は，$\hat{\boldsymbol{h}}_j^T \boldsymbol{x}_\ell$や$\tilde{\boldsymbol{h}}_j^T \boldsymbol{x}_\ell$といった推定は非常に大きなバイアスが生むことを指摘し，次のような推定を考えた．

$$\tilde{x}_{\ell j} = \tilde{\boldsymbol{h}}_{\ell j}^T \boldsymbol{x}_\ell \quad (j = 1, ..., k) \tag{4.4}$$

ただし，

$$\tilde{\boldsymbol{h}}_{\ell j} = (n-1)^{1/2}(\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{u}}_{\ell j}/\{(n-2)\tilde{\lambda}_j^{1/2}\},$$
$$\hat{\boldsymbol{u}}_{\ell j} = (\hat{u}_{1j}, ..., \hat{u}_{\ell-1 j}, -\hat{u}_{\ell j}/(n-1), \hat{u}_{\ell+1 j}, ..., \hat{u}_{nj})^T$$

ここで，$\sum_{\ell=1}^{n} \tilde{\boldsymbol{h}}_{\ell j}/n = \tilde{\boldsymbol{h}}_j$となることに注意する．

5. 高次元2標本検定とデータ解析

母集団が2つあり，各母集団π_i $(i=1,2)$は平均にp次ベクトル$\boldsymbol{\mu}_i$，共分散行列にp次非負定値対称行列$\boldsymbol{\Sigma}_i$ $(=\sum_{s=1}^{p}\lambda_{is}\boldsymbol{h}_{is}\boldsymbol{h}_{is}^T)$をもつとする．次の検定を考える[7]．

$$H_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2 \quad \text{vs.} \quad H_1:\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2 \tag{5.1}$$

各母集団から，n_i (≥ 3)個のp次データベクトル$\boldsymbol{x}_{i1},...,\boldsymbol{x}_{in_i}$を無作為に抽出し，データ行列を$\boldsymbol{X}_i=[\boldsymbol{x}_{i1},...,\boldsymbol{x}_{in_i}]$とする．高次元小標本では標本共分散行列の逆行列が存在しないので，従来のホテリングのT^2-統計量は使えない．高次元2標本検定では，次の統計量が使われる．

$$T=\|\overline{\boldsymbol{x}}_1-\overline{\boldsymbol{x}}_2\|^2-\sum_{i=1}^{2}\frac{\mathrm{tr}(\boldsymbol{S}_i)}{n_i}=2\sum_{i=1}^{2}\frac{\sum_{\ell<\ell'}^{n_i}\boldsymbol{x}_{i\ell}^T\boldsymbol{x}_{i\ell'}}{n_i(n_i-1)}-2\frac{\sum_{\ell=1}^{n_1}\sum_{\ell'=1}^{n_2}\boldsymbol{x}_{1\ell}^T\boldsymbol{x}_{2\ell'}}{n_1n_2} \tag{5.2}$$

ただし，$\overline{\boldsymbol{x}}_i=\sum_{\ell=1}^{n_i}\boldsymbol{x}_{i\ell}/n_i$，$\overline{\boldsymbol{X}}_i=[\overline{\boldsymbol{x}}_i,...,\overline{\boldsymbol{x}}_i]$, $\boldsymbol{S}_i=(n_i-1)^{-1}(\boldsymbol{X}_i-\overline{\boldsymbol{X}}_i)(\boldsymbol{X}_i-\overline{\boldsymbol{X}}_i)^T$である．ここで，$E(T)=\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2$となり, H_0のもとで

$$\mathrm{Var}(T)=2\sum_{i=1}^{2}\frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{n_i(n_i-1)}+\frac{4}{n_1n_2}\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\quad(=K\text{とおく})$$

となることに注意する．2つの母集団がともにNSSEモデルのとき，Tに関する高次元漸近正規性がChen and Qin [6]とAoshima and Yata [1, 3]によって示され，それぞれ別の検定手法が与えられた．ところが，2つの母集団のどちらか一方でもSSEモデルのとき，3節の図4のように，Tの高次元漸近正規性は成立しない．さらに，2つの母集団がともにSSEモデルのとき，$K=O(\lambda_{11}^2/n_1^2+\lambda_{21}^2/n_2^2)$となり，各母集団の最大固有値の影響で$T$の分散は非常に大きくなる．つまり，SSEモデルが当てはまる状況では，Tの高次元漸近正規性が壊れるだけでなく，検定の精度も著しく悪くなる．

Aoshima and Yata [4]は，SSEモデルも考慮し，データ変換法を使って高次元2標本検定を次のように与えた．各母集団のデータを，(4.1)と同様にデータ変換する．

$$\boldsymbol{x}_{i\ell *}=\boldsymbol{A}_i\boldsymbol{x}_{i\ell}\ \ (\ell=1,...,n_i)\quad(\text{ただし，}\boldsymbol{A}_i=\textstyle\sum_{s=k_i+1}^{p}\boldsymbol{h}_{is}\boldsymbol{h}_{is}^T\ (k_i\geq 0)\text{である})$$

変換後のデータを使うと，(5.2)式のTは次のようになる．

$$\begin{aligned}T_*=&2\sum_{i=1}^{2}\sum_{\ell<\ell'}^{n_i}\frac{\boldsymbol{x}_{i\ell}^T\boldsymbol{A}_i\boldsymbol{x}_{i\ell'}}{n_i(n_i-1)}-2\frac{\sum_{\ell=1}^{n_1}\sum_{\ell'=1}^{n_2}\boldsymbol{x}_{1\ell}^T\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{x}_{2\ell'}}{n_1n_2}\\=&2\sum_{i=1}^{2}\frac{\sum_{\ell<\ell'}^{n_i}(\boldsymbol{x}_{i\ell}^T\boldsymbol{x}_{i\ell'}-\sum_{j=1}^{k_i}x_{i\ell j}x_{i\ell' j})}{n_i(n_i-1)}\\&-2\frac{\sum_{\ell=1}^{n_1}\sum_{\ell'=1}^{n_2}(\boldsymbol{x}_{1\ell}-\sum_{j=1}^{k_1}x_{1\ell j}\boldsymbol{h}_{1j})^T(\boldsymbol{x}_{2\ell'}-\sum_{j=1}^{k_2}x_{2\ell' j}\boldsymbol{h}_{2j})}{n_1n_2}\end{aligned}$$

[7] 本節は，SSEモデルも考慮した高次元統計的推測を説明するために，基本的で本質的な高次元2標本検定を扱う．高次元判別分析などは，[5, 8]を参照のこと．なお，いわば基礎編となるNSSEモデルの高次元統計的推測については，青嶋・矢田 [2]が入門書となる．

ただし，$x_{i\ell j} = \boldsymbol{h}_{ij}^T \boldsymbol{x}_{i\ell}$ である．ここで，$E(T_*) = \|\boldsymbol{A}_1\boldsymbol{\mu}_1 - \boldsymbol{A}_2\boldsymbol{\mu}_2\|^2$ (=Δ_* とおく) となり，H_0 のもとで

$$\mathrm{Var}(T_*) = 2\sum_{i=1}^{2} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{i*}^2)}{n_i(n_i-1)} + \frac{4}{n_1 n_2}\mathrm{tr}(\boldsymbol{\Sigma}_{1*}\boldsymbol{\Sigma}_{2*}) \quad (= K_* \text{とおく})$$

となることに注意する．ただし，$\boldsymbol{\Sigma}_{i*} = \sum_{s=k_i+1}^{p} \lambda_{is}\boldsymbol{h}_{is}\boldsymbol{h}_{is}^T$ である．Aoshima and Yata [4] は，T_* に関して高次元漸近正規性を証明した．さらに，Aoshima and Yata [4] は，T_* を構成する $\boldsymbol{h}_{ij}$ と $x_{i\ell j}$ を (4.3) 式の $\tilde{\boldsymbol{h}}_{ij}$ と (4.4) 式の $\tilde{x}_{i\ell j}$ で置き換えた T_* の推定量 $\widehat{T}_*$ について，次の結果を与えた．

定理 4 ([4]). *2* つの母集団に *(A-i)* を仮定する．*SSE* モデルに *(A-ii)* を仮定する．さらに，$v = \min\{p, n_1, n_2\}$ とおき，$\limsup_{v\to\infty} \Delta_*^2/K_* < \infty$ を仮定する．適当な正則条件[8] のもと，$v \to \infty$ のとき次が成り立つ．

$$(\widehat{T}_* - \Delta_*)/\sqrt{\widehat{K}_*} \xrightarrow{\mathcal{L}} N(0,1)$$

ただし，$\widehat{K}_*$ は *Aoshima and Yata [4]* の *5* 節で与えられる K_* の一致推定量である．

仮説 (5.1) は，次のように検定される．

$$\widehat{T}_*/\sqrt{\widehat{K}_*} > z_\alpha \Longrightarrow \text{有意水準}\,\alpha\,\text{で}\,H_0\,\text{を棄却} \tag{5.3}$$

ただし，z_α は $N(0,1)$ の上側 α 点を表す．検定方式 (5.3) の第 1 種の過誤と検出力については，Aoshima and Yata [4] の定理 6 を参照のこと．

図 1 で紹介した Wang et al. [13] の次世代シーケンサーによる AMD ゲノムデータを解析する．ゲノム領域数が $3095656\ (= p)$ 次元，AMD 群が 21$(= n_1)$ サンプル，健常群 (Normal) が 19$(= n_2)$ サンプルであった．図 1 で見たように，両群とも SSE モデルが当てはまる．仮説 (5.1) を有意水準 $\alpha = 0.05$ で検定する．検定方式 (5.3) の検定統計量を計算すると[9]，$\widehat{T}_*/\sqrt{\widehat{K}_*} = 23.44 > z_{0.05} = 1.65$ となり，有意水準 5% で H_0 は棄却される．SSE モデル (非スパース性) を考慮し，データ変換法を使って非スパースなノイズを除去したことで潜在情報が浮き彫りとなり，2 群間の平均の差異が検出できた．

ちなみに，両群にスパース性を仮定して NSSE モデルのもとで検定統計量を計算すると，$T/\sqrt{\widehat{K}} = 1.01 < z_{0.05}$ となり，H_0 は棄却されない．しかし，この結論は統計的に保証されるものではない．なぜならば，NSSE モデル (スパース性の仮定) が妥当ではなく，検定統計量に高次元漸近正規性が成立していないからである．$K = O(\lambda_{11}^2/n_1^2 + \lambda_{21}^2/n_2^2)$ に注意すれば，T の潜在情報である $E(T) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ が非スパースなノイズとなる最大固有値に押しつぶされ，2 群間の平均の差異を検出できなかったことが分かる．

次に，AMD 群と健常群 (Normal) の差異の原因となる変数 (ゲノム領域) を絞り込む．いま，$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = (\mu_{o1}, ..., \mu_{op})^T$，$\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 = (\overline{x}_{o1}, ..., \overline{x}_{op})^T$ とおく．有意な変数の集合を $D = \{j\,|\,|\mu_{oj}| > 0\}$ とおく．$\overline{x}_{o1}, ..., \overline{x}_{op}$ を $|\overline{x}_{o(1)}| \geq \cdots \geq |\overline{x}_{o(p)}|$ と絶対値の大きい順に並び替え，m 個の変数からなる $\widehat{D}_m = \{j\,|\,|\overline{x}_{oj}| \geq |\overline{x}_{o(m)}|\}$ で D を推定する．適当な m と

[8] Aoshima and Yata [4] の定理 5 を参照のこと．

[9] k_i の推定については，[4] の S2.2 節を参照のこと．このデータセットは，$(\hat{k}_1, \hat{k}_2) = (7, 7)$ であった．

適当な正則条件のもと，$P(D \subseteq \widehat{D}_m) \to 1 \ (p \to \infty)$ となる[10]．しかし，この変数選択はスパース性を当てにしたものである．

そこで，非スパース性を考慮し，データ変換法を使って $\overline{\boldsymbol{x}}_{i*} = \boldsymbol{A}_i \overline{\boldsymbol{x}}_i$ とする．(4.3)式の $\tilde{\boldsymbol{h}}_{ij}$ と (4.4)式の $\tilde{x}_{i\ell j}$ を用いて，$\overline{\boldsymbol{x}}_{i*}$ を

$$\overline{\tilde{\boldsymbol{x}}}_{i*} = \overline{\boldsymbol{x}}_i - \sum_{j=1}^{k_i} \overline{\tilde{x}}_{ij} \tilde{\boldsymbol{h}}_{ij} \quad \text{(ただし，} \overline{\tilde{x}}_{ij} = \textstyle\sum_{\ell=1}^{n_i} \tilde{x}_{i\ell j}/n_i \text{ である)}$$

で推定する．$\overline{\tilde{\boldsymbol{x}}}_{1*} - \overline{\tilde{\boldsymbol{x}}}_{2*} = (\overline{\tilde{x}}_{o1}, ..., \overline{\tilde{x}}_{op})^T$ とし，$\overline{\tilde{x}}_{o1}, ..., \overline{\tilde{x}}_{op}$ を $|\overline{\tilde{x}}_{o(1)}| \geq \cdots \geq |\overline{\tilde{x}}_{o(p)}|$ と並び替え，m 個の変数からなる $\widetilde{D}_m = \{j \mid |\overline{\tilde{x}}_{oj}| \geq |\overline{\tilde{x}}_{o(m)}|\}$ で D を推定する[11]．

結果を見てみよう．$\widehat{D}_m$ と $\widetilde{D}_m$ を用いて，3095656 個の変数から $m = 10,\ 100,\ 1000$ 個の変数をそれぞれ選択した．変数選択の精度を高次元 PCA で確認する．$\boldsymbol{x}_{ij}$ において，$\widehat{D}_m$ の変数のみを利用した m 次元データを $\widehat{\boldsymbol{x}}_{ij}$，$\widetilde{D}_m$ の変数のみを利用した m 次元データを $\widetilde{\boldsymbol{x}}_{ij}$ とし，AMD 群と健常群 (Normal) を混合させたデータセットを $\widehat{\boldsymbol{X}} = (\widehat{\boldsymbol{x}}_{11}, ..., \widehat{\boldsymbol{x}}_{1n_1}, \widehat{\boldsymbol{x}}_{21}, ..., \widehat{\boldsymbol{x}}_{2n_2})$，$\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{x}}_{11}, ..., \widetilde{\boldsymbol{x}}_{1n_1}, \widetilde{\boldsymbol{x}}_{21}, ..., \widetilde{\boldsymbol{x}}_{2n_2})$ とおく．Yata and Aoshima [17] は，規準化した主成分スコアを用いることで，高次元混合データのクラスタリングが可能であることを証明した．2 群間の差異に有意な変数を正しく選択できていれば，理論的には，主成分スコアを使って 2 群を精度良く分離できるはずである．そこで，$\widehat{\boldsymbol{X}}$ と $\widetilde{\boldsymbol{X}}$ の各々について，$n_1 + n_2 = 40$ 個のサンプルに対する（規準化した）第 1 主成分スコアと第 2 主成分スコアを計算し，図 5 にプロットした．データ変換法を使った $\widetilde{D}_m$ による変数選択は，2 群を精度よく分離できていることが見てとれる．データ変換法を使って非スパースなノイズを除去することで，3095656 個の変数から 2 群の差異に有意な 100 個の変数を絞り込めたのである．

非スパースモデリングの一幕をお見せした．データ解析に要した時間は，モバイル PC で 5 分程度である．高い計算処理能力と大標本を前提とする深層学習が，経済的にも技術的にも環境的にも敷居の高いテクノロジーであるのに対して，高次元統計解析による非スパースモデリングは，いわば庶民的な町工場の技術である．非スパースモデリングを使うことで，非スパースなノイズの影響を取り除くことができ，超高次元データもたった数十程度の小標本で処理でき，モバイル PC でも高速かつ高精度に解析することができるのである．

参考文献

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

[2] 青嶋 誠，矢田和善 (2019). 高次元の統計学，共立出版，東京.

[3] Aoshima, M. and Yata, K. (2015). Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions, *Meth. Comput. Appl. Probab.*, **17**, 419–439.

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

[5] Aoshima, M. and Yata, K. (2019). Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models, *Ann. Inst. Statist. Math.*, **71**, 473–503.

[6] Chen, S.X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing, *Ann. Statist.*, **38**, 808–835.

[10] 例えば，Aoshima and Yata [1] の 5 節を参照のこと．

[11] $\widetilde{D}_m$ は $\widehat{D}_m$ よりも緩い条件で一致性をもつ．

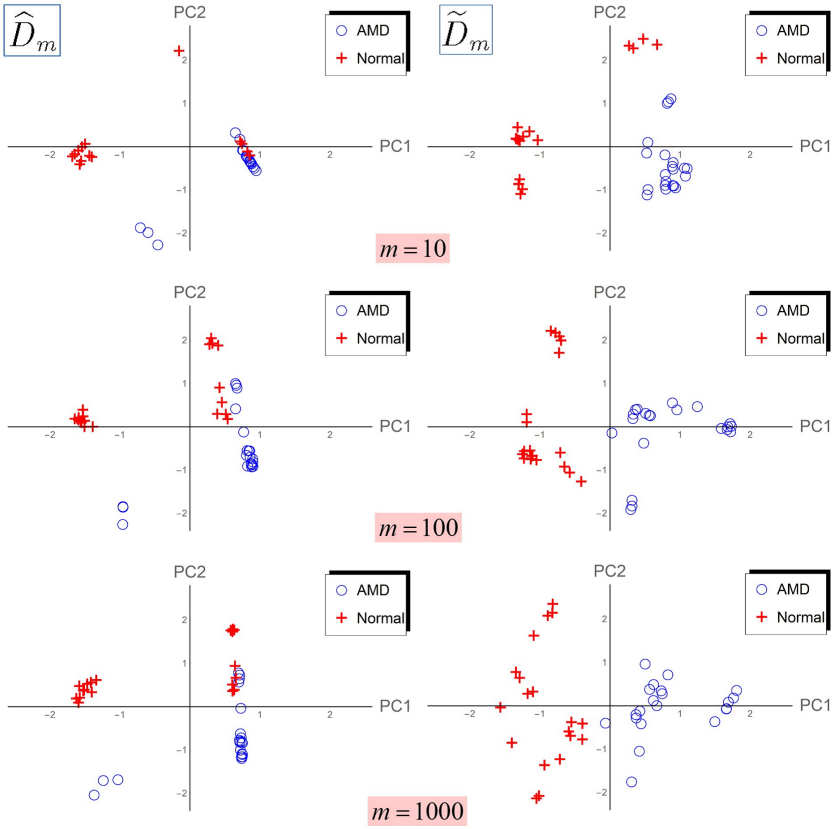

図 5. $m = 10, 100, 1000$ 個の変数を選択した AMD 群と健常群 (Normal) の混合データに対する第 1 と第 2 の主成分スコア. $\widehat{D}_m$: データ変換なし (左), $\widetilde{D}_m$: データ変換法 (右).

[7] Donoho, D.L. (2006). Compressed sensing, IEEE Trans. Information Theory, **52**, 1289–1306.

[8] Ishii, A., Yata, K. and Aoshima, M. (2022). Geometric classifiers for high-dimensional noisy data, *J. Multivariate Anal.* (*Editor's invited paper*), **188**, 104850.

[9] Johnstone, I.M. (2001). On the distribution of the largest eigenvalue in principal components analysis, *Ann. Statist.*, **29**, 295–327.

[10] Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, *Statistica Sinica*, **17**, 1617–1642.

[11] Takeuchi, T, Yata, K., Aoshima, M., et al. (2022). High dimensional statistical analysis and its application to ALMA map of NGC 253, *arXiv preprint*, arXiv:2203.04535.

[12] von Roemeling, C., Radisky, D., Marlow, L., et al. (2014). Neuronal pentraxin 2 supports clear cell renal cell carcinoma by activating the AMPA-selective glutamate receptor-4, *Cancer Res.*, **74**, 4796–4810.

[13] Wang, J., Zibetti, C., Shang, P., et al. (2018). ATAC-seq analysis reveals a widespread decrease of chromatin accessibility in age-related macular degeneration, *Nat. Commun.*, **9**, 1364.

[14] 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.

[15] 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.

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

[17] Yata, K. and Aoshima, M. (2020). Geometric consistency of principal component scores for high-dimensional mixture models and its application, *Scand. J. Stat.*, **47**, 899–921.