強スパイク固有値モデルにおける高次元統計的推測

石井　晶　(東京理大理工)

1. はじめに

情報化の進展に伴い，データの次元数が標本数よりも遥かに大きい高次元小標本(HDLSS)データの解析が益々重要になっている．例えば，ゲノム科学，情報工学，金融工学などで見られるように，データの次元数 (p) は優に数万を超え，それに対して，標本数 (n) は数十程度という状況が多々ある．高次元データは，1つの標本に対して多くの情報が得られる一方で，重要な潜在情報が巨大なノイズ空間に埋もれてしまう．それゆえ，巨大なノイズ空間から，いかに潜在情報を抜き出せるか，ということが重要である．このような高次元小標本データに対して，従来の多変量解析の理論や方法論を用いることは出来ない．高次元小標本データの特徴を生かした，全く新しい理論と方法論の構築が必要である．

いま，高次元データの共分散行列として p 次の正定値対称行列 $\mathbf{\Sigma}$ を考える．$\mathbf{\Sigma}$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_p (> 0)$ とする．高次元データに対して理論を構築する際に鍵となるのは，その固有空間の構造である．スパース推定の枠組みでは，固有値に関して「λ_1 が次元数 p に関して有界である」という条件がよく仮定される．しかしながら，高次元データの共分散行列は次元数と共に行列のサイズは大きくなり，その固有値は次元数に依存した，次元数の関数として捉えることが自然である．実際，青嶋 [2]，青嶋・矢田 [4, 5] 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 [10] 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．

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

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．

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

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．簡単のため，高次元データにおける固有値モデルの1つである一般化スパイクモデルを例にとる．一般化スパイクモデルとは，以下のような固有値構造をもつものである．

$$\lambda_s = c_s p^{\alpha_s} \ (s = 1, \cdots, t) \ \text{ かつ } \ \lambda_s = c_s \ (s = t+1, \cdots, p) \tag{1.3}$$

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．$\alpha_1 \geq 1/2$ のとき，(1.3) は強スパイク固有値モデル (1.1) の1つとなり，$\alpha_1 < 1/2$ のとき，弱スパイク固有値モデル (1.2) の1つとなる．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，二標本検定や判別分析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 [9] 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．そればかりでなく，強スパイクしている固有値が巨大なノイズとなり，潜在空間の情報を埋もれさせてしまう場合もある．強スパイクする固有値を生み出す原因としては，変数間の相関，異常値の混入，データの混合などが挙げられる．

実際に，2つの有名な遺伝子発現データを用いて，高次元データの固有値構造を見てみる．ノイズ掃き出し法を用いて，最大固有値 ($\tilde{\lambda}_1$) から第5固有値 ($\tilde{\lambda}_5$) までを推定し，表1にその数値を纏め，それらを図1にプロットした．なお，固有値の推定は，データを基準化した後に行った．1つ目は Singh et al. [35] で与えられた遺伝子数 $p = 12625$ の前立腺がんデータで，前立腺がん患者 (Prostate cancer) の52標本と非腫瘍者 (Non-tumor) の50標本という2つのクラスを含んでいる．2つ目は Borovecki et al. [16] で与えられた遺伝子数 $p = 22283$ のハンチントン病データで，ハンチントン病患者 (Huntington's disease) の17標本と正常者 (Normal) の14標本という2つのクラスを含んでいる．ノイズ掃き出し法については，本稿の第2節で紹介する．表1や図1からも，最初の数個の固有値は大きい値をとり，特に最大固有値は強くスパイクしている様子が見てとれる．

表1. ノイズ掃き出し法による2つの遺伝子発現データの第5固有値までの推定値．

	$\tilde{\lambda}_1$	$\tilde{\lambda}_2$	$\tilde{\lambda}_3$	$\tilde{\lambda}_4$	$\tilde{\lambda}_5$
前立腺がん患者 (Singh et al. [35])	7192	586	433	323	237
非腫瘍者 (Singh et al. [35])	8510	503	381	227	200
ハンチントン病患者 (Borovecki et al. [16])	18033	649	387	189	101
正常者 (Borovecki et al. [16])	15590	1095	633	417	226

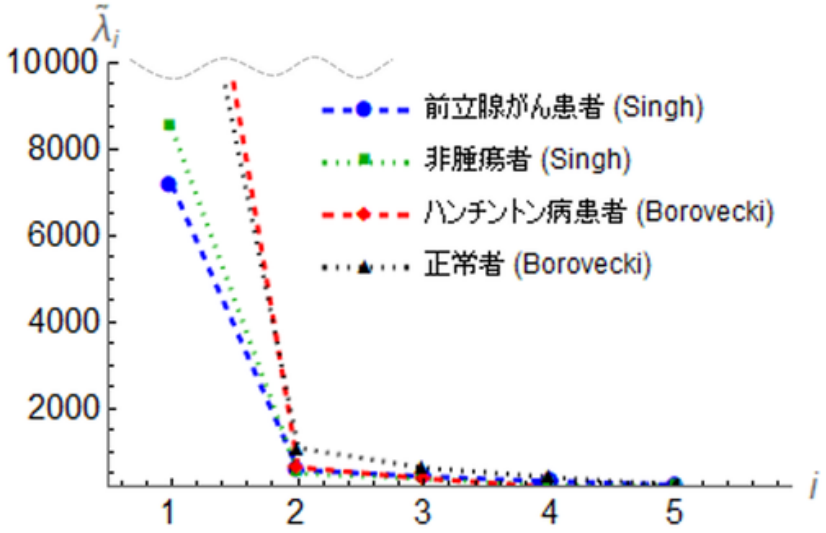

図1. 2つの遺伝子発現データに対する固有値のスパイク状況．

本稿の第2節では，Ishii, Yata and Aoshima [25, 26] で与えた強スパイク固有値モデルにおける固有空間の推定についての結果を纏める．第3節では，Ishii, Yata and Aoshima

[26, 28] で与えた高次元平均ベクトルについて，一標本検定の結果を，第4節では，Ishii [22, 23]，Ishii, Yata and Aoshima [28] で与えた二標本検定の結果を纏める．第5節では，Ishii, Yata and Aoshima [26, 27] で与えた高次元共分散行列の同等性検定について結果を纏める．最後に，第6節では，Ishii [24]，Ishii, Yata and Aoshima [29] で与えた高次元判別分析についての結果を纏める．

2. 強スパイク固有値モデルにおける固有空間の推定

高次元データの共分散行列に対し，固有値の理論的研究で重要な結果が与えられたのは，2000年以降である．Johnstone [30] や Paul [33] 等は標本固有値の漸近的性質を導出した．これらの研究は，データの次元数 p と標本数 n が $n/p \to c\,(>0)$ 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や類する条件を仮定していた．しかしながら，高次元小標本の枠組みでは n を p と同程度には仮定できない．さらに，Johnstone [30] らは，$\mathbf{\Sigma}$ の固有値に次のようなスパイクモデルを仮定していた．

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

スパイクモデル (2.1) において，$\lambda_1, ..., \lambda_t$ が推定の対象となる潜在的な固有空間である．しかしながら，図1でも示したように，「固有値が次元数 p に依存しない」という上記の固有値モデルは，実データがもつ固有値構造からかけ離れている．スパイクモデル (2.1) において，λ_s の $t+1$ 番目以降は，$\lambda_s\ (s = 1, ..., t)$ を推定する際のノイズである．しかしながら，すべてのノイズが等しいという設定は，数学的な扱い易さゆえであり，実際には厳しい制約である．このような背景から，Yata and Aoshima [38] において，一般化スパイクモデル (1.3) が考えられた．Yata and Aoshima [38] は，一般化スパイクモデルのもと，従来型PCA が高次元小標本に対して不適解を起こすことを示した．それに対して，Yata and Aoshima [39] と Yata and Aoshima [40] は，「クロスデータ行列法」と「ノイズ掃き出し法」という2つの方法論を提案し，それに基づく新しいPCA を提唱した．高次元小標本データに対するPCA の一致性等の詳細は，Aoshima et al. [13] を参照のこと．さらに，Ishii, Yata and Aoshima [25, 26] では，強スパイク固有値モデルに着目し，ノイズ掃き出し法による固有空間の推定量の一致性を「$p \to \infty$ だが，n は固定」という枠組みへ拡張した．本節では，Ishii, Yata and Aoshima [25, 26] の結果に基づき，強スパイク固有値モデルにおける固有空間の推定を扱う．

平均が $\boldsymbol{\mu}$，共分散行列が $\mathbf{\Sigma}$ 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 $\mathbf{\Sigma} = \boldsymbol{H}\mathbf{\Lambda}\boldsymbol{H}^T, \mathbf{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_p)$ と分解する．そのとき，$\boldsymbol{Z} = \mathbf{\Lambda}^{-1/2}\boldsymbol{H}^T(\boldsymbol{X} - [\boldsymbol{\mu}, ..., \boldsymbol{\mu}])$ 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．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．$\boldsymbol{S}_n$ 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．一方で，$n \times n$ 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．$\boldsymbol{S}_D$ の $\hat{\lambda}_j$ に対応する固有ベクトルを $\hat{\boldsymbol{u}}_j$ とし，その固有値分解を

$$\boldsymbol{S}_D = \sum_{j=1}^{n-1} \hat{\lambda}_j \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^T \tag{2.2}$$

とする．ここで，

$$\hat{\boldsymbol{h}}_j = \frac{\boldsymbol{X} - \overline{\boldsymbol{X}}}{\sqrt{(n-1)\hat{\lambda}_j}} \hat{\boldsymbol{u}}_j \quad (j = 1, ..., n-1) \tag{2.3}$$

となることに注意する．(2.2) 式と (2.3) 式より，$p \times p$ 行列 $\boldsymbol{S}_n$ の固有値と固有ベクトルは，$n \times n$ 行列 $\boldsymbol{S}_D$ の固有値と固有ベクトルを使って計算できる．よって，高次元小標本データにおいて，双対標本共分散行列を用いることで，計算コストを大幅に削減することができる．いま，最大固有値に対し，次の条件を考える．

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

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

(A-i) は，「$\lambda_2/\lambda_1 \to 0$ $(p \to \infty)$」を意味し，最大固有値が特に強くスパイクしている強スパイク固有値モデルである．本稿では，(A-i) を「単一強スパイク固有値モデル」と呼ぶことにする．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 χ_{n-1}^2 に従うことに注意する．

2.1. 高次元固有値推定量の一致性と漸近分布

Ishii, Yata and Aoshima [25, 26] では，「$p \to \infty$ 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った．Yata and Aoshima [40] が与えていなかった高次元小標本における幾何学的表現を与え，さらに，ノイズ掃き出し法による最大固有値の推定量について漸近分布を導出した．Ishii, Yata and Aoshima [25, 26] では，高次元小標本における幾何学的表現に基づき，標本固有値 $\hat{\lambda}_1$ の不一致性を次のように与えた．

補題 2.1 (Ishii, Yata and Aoshima [25, 26])**.** (A-i) と (A-ii) を仮定する．そのとき，「$p \to \infty$ だが，n は固定」もしくは，「$p \to \infty$ かつ $n \to \infty$」の枠組みで，次が成り立つ．

$$\frac{\hat{\lambda}_1}{\lambda_1} - \frac{\kappa}{\lambda_1(n-1)} = z_{o1}^2 + o_P(1)$$

ここで，$\kappa = \mathrm{tr}(\boldsymbol{\Sigma}) - \lambda_1 = \sum_{j=2}^{p} \lambda_j$ である．

補題 2.1 は，標本固有値が $\kappa/(n-1)$ なる大きさのノイズを含むことを意味している．よって，ノイズの大きさを見積もり，標本固有値から掃き出せば，最大固有値の漸近的な性質が得られることを示唆している．このようなアイディアで開発された方法論

がノイズ掃き出し法である．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．

$$\tilde{\lambda}_i = \hat{\lambda}_i - \frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{j=1}^{i} \hat{\lambda}_j}{n-1-i} \quad (i = 1, ..., n-2) \tag{2.4}$$

ここで，(2.4) 式の第 2 項は $\kappa/(n-1)$ の推定量である．Yata and Aoshima [40, 41] は，「$p \to \infty$ かつ $n \to \infty$」の枠組みで，$\tilde{\lambda}_i$ の一致性について議論した．その一方で，Ishii, Yata and Aoshima [25, 26] 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 [25, 26])**.** (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 より，$n = 5$ 程度の高次元小標本データであっても，$\tilde{\lambda}_1$ は漸近分布をもつ．さらに，Ishii, Yata and Aoshima [26] は，定理 2.1 より，最大固有値の寄与率の信頼区間を構築した．また，Ishii [23] は，(A-ii) を満たさない場合について考察し，クロスデータ行列法を用いた最大固有値推定量の漸近分布を導出した．

2.2. 高次元固有ベクトル推定の一致性

Ishii, Yata and Aoshima [26] では，標本固有ベクトル $\hat{\boldsymbol{h}}_1$ の不一致性を次のように示した．

補題 2.2 (Ishii, Yata and Aoshima [26])**.** (A-i) と (A-ii) を仮定する．そのとき，「$p \to \infty$ だが，n は固定」もしくは，「$p \to \infty$ かつ $n \to \infty$」の枠組みで，次が成り立つ．

$$\hat{\boldsymbol{h}}_1^T \boldsymbol{h}_1 = \left(1 + \frac{\kappa}{\lambda_1 z_{o1}^2}\right)^{-1/2} + o_P(1)$$

つまり，標本固有ベクトルも $\kappa/(n-1)$ なる大きさのノイズを含む．ノイズ掃き出し法による固有ベクトルの推定は，(2.3) 式にある標本固有値 $\hat{\lambda}_j$ を (2.4) 式の $\tilde{\lambda}_j$ で置き換えた

$$\tilde{\boldsymbol{h}}_j = \frac{\boldsymbol{X} - \overline{\boldsymbol{X}}}{\sqrt{(n-1)\tilde{\lambda}_j}}\hat{\boldsymbol{u}}_j \quad (j = 1, ..., n-1) \tag{2.5}$$

で与えられる．Yata and Aoshima [40, 41] は，「$p \to \infty$ かつ $n \to \infty$」の枠組みで，$\tilde{\boldsymbol{h}}_j$ の一致性について議論している．一方で，Ishii, Yata and Aoshima [26] は，「$p \to \infty$ だが，n は固定」の枠組みでも，$\tilde{\boldsymbol{h}}_1$ の一致性を次のように与えた．

定理 2.2 (Ishii, Yata and Aoshima [26])**.** (A-i) と (A-ii) を仮定する．「$p \to \infty$ だが，n は固定」もしくは，「$p \to \infty$ かつ $n \to \infty$」の枠組みで，次が成り立つ．

$$\tilde{\boldsymbol{h}}_1^T \boldsymbol{h}_1 = 1 + o_P(1) \tag{2.6}$$

なお，(2.6) 式は，$\|\tilde{\boldsymbol{h}}_1\| = \sqrt{\hat{\lambda}_1/\tilde{\lambda}_1} \geq 1$ なので，$\|\tilde{\boldsymbol{h}}_1 - \boldsymbol{h}_1\| = o_P(1)$ と同値ではないことに注意する．ここで，$\|\cdot\|$ はユークリッドノルムを表す．しかしながら，(2.6) 式の意味での一致性が，以降の節で扱う高次元統計的推測で重要となる．

3. 高次元平均ベクトルの一標本検定

母集団に関する設定は，第2節と同じとする．高次元平均ベクトルに対して，次の検定問題を考える．

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0 \tag{3.1}$$

ここで，$\boldsymbol{\mu}_0$ は $\boldsymbol{\mu}_0 = \boldsymbol{0}$ など，既知のベクトルである．本節では，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 [19, 20] や Srivastava [36] 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 [14] は次の検定統計量を議論した．

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

$E(T_{\text{BS}}) = \|\boldsymbol{\mu}\|^2$ となることに注意する．Bai and Saranadasa [14] 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_{BS} に関する漸近正規性を示した．しかしながら，T_{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，Ishii, Yata and Aoshima [26] では，T_{BS} を単一強スパイク固有値モデルのもとで修正し，新たな検定手法を提案した．一方，Ishii, Yata and Aoshima [28] 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．

3.1. 単一強スパイク固有値モデルにおける一標本検定

Ishii, Yata and Aoshima [26] 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_{BS} に対し，次の結果を得た．

補題 3.1 (Ishii, Yata and Aoshima [26])**.** (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．

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

補題3.1 をもとに，Ishii, Yata and Aoshima [26] では，次の検定統計量を提案した．

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

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 3.1 (Ishii, Yata and Aoshima [26])**.** (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_{ν_1,ν_2} は自由度 (ν_1, ν_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 (3.1) 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 (ν_1, ν_2) の F 分布の上側 α 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 3.1 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 α に抑えた一標本検定を行うことができる．

3.2. データ変換を利用した一標本検定

Ishii, Yata and Aoshima [28] では，強スパイク固有値モデルのもと，「$p \to \infty$ かつ $n \to \infty$」の枠組みで新たな検定手法を提案した．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，Ψ_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 [10] 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 [10] 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$)$,

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

$\lambda_{\max}(\boldsymbol{\Sigma}_*)$ を $\boldsymbol{\Sigma}_*$ の最大固有値とすると，$\text{tr}(\boldsymbol{\Sigma}_*^2) = \Psi_{k+1}$ 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，

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

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．

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

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．$E(T_{\text{DT}}) = \Delta_*$ であり，$\text{Var}(T_{\text{DT}}) = K_*$ とおくと，

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

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

このとき，次の定理が得られる．

定理 3.2 (Ishii, Yata and Aoshima [28])**.** (A-iv) から (A-vi) のもと，$m \to \infty$ で次が成り立つ．

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

したがって，Ishii, Yata and Aoshima [28] では，データ変換を用いて，次の検定統計量を構築した．

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

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

図 1 からも見てとれるように，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 (3.1) 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 3.3 (Ishii, Yata and Aoshima [28])**.** (A-iv) から (A-viii) のもと，$m \to \infty$ で次が成り立つ．

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

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 [28] では，定理 3.3 より，予め設定した $\alpha \in (0, 1/2)$ に対し，検定問題 (3.1) に対する検定ルールを次のように与えた．

$$H_0 \text{ を棄却する} \iff \frac{\widehat{T}_{\text{DT}}}{\widehat{K}_{1*}^{1/2}} > z_\alpha \tag{3.2}$$

ここで，z_α は $N(0,1)$ の上側 α 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 (3.2) 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 3.4 (Ishii, Yata and Aoshima [28])**.** (A-iv), (A-v), (A-vii), (A-viii) を仮定する．$m \to \infty$ で次が成り立つ．

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

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

4. 高次元平均ベクトルの二標本検定

母集団が 2 つあると仮定する．各母集団 π_i は，平均に p 次のベクトル $\boldsymbol{\mu}_i$，共分散行列に p 次の正定値対称行列 $\boldsymbol{\Sigma}_i$ をもつとする．母集団 π_i から，n_i 個の p 次元データベクトル $\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}$ を無作為に抽出して，$p \times n_i$ データ行列 $\boldsymbol{X}_i = [\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}]$ を定義する．ただし，$p > n_i$ である．$\boldsymbol{\Sigma}_i$ の固有値を $\lambda_{1(i)} \geq \cdots \geq \lambda_{p(i)} (> 0)$ とし，適当な直交行列 $\boldsymbol{H}_i = [\boldsymbol{h}_{1(i)}, ..., \boldsymbol{h}_{p(i)}]$ で $\boldsymbol{\Sigma}_i$ を $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i \boldsymbol{H}_i^T$, $\boldsymbol{\Lambda}_i = \text{diag}(\lambda_{1(i)}, ..., \lambda_{p(i)})$ と分解する．ただし，$\liminf_{p\to\infty} \lambda_{1(i)}/\lambda_{2(i)} > 1$ を仮定する．$\boldsymbol{Z}_i = \boldsymbol{\Lambda}_i^{-1/2} \boldsymbol{H}_i^T (\boldsymbol{X}_i - [\boldsymbol{\mu}_i, ..., \boldsymbol{\mu}_i])$ とおき，$\boldsymbol{Z}_i = [\boldsymbol{z}_{1(i)}, ..., \boldsymbol{z}_{p(i)}]^T$，$\boldsymbol{z}_{s(i)} = (z_{s1(i)}, ..., z_{sn_i(i)})^T$ $(s = 1, ..., p)$ と表記する．第 2 節と同様に，各 $s = 1, ..., p$ に対し，$z_{os(i)}^2 = \sum_{k=1}^{n_i} (z_{sk(i)} - \bar{z}_{s(i)})^2/(n_i - 1)$，$\bar{z}_{s(i)} = n_i^{-1} \sum_{k=1}^{n_i} z_{sk(i)}$ とおく．各母集団に対し，標本共分散行列を $\boldsymbol{S}_{in_i} = (n_i - 1)^{-1} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)^T$ と定義する．ただし，$\overline{\boldsymbol{X}}_i = [\overline{\boldsymbol{x}}_{in_i}, ..., \overline{\boldsymbol{x}}_{in_i}]$，$\overline{\boldsymbol{x}}_{in_i} = n_i^{-1} \sum_{k=1}^{n_i} \boldsymbol{x}_{ik}$ である．

いま，次の検定問題を考える．

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 \tag{4.1}$$

高次元データに対する二標本検定として有名なものは，次の検定統計量である．

$$T_n = \|\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}\|^2 - \sum_{i=1}^{2} \text{tr}(\boldsymbol{S}_{in_i})/n_i$$

Bai and Saranadasa [14]，Chen and Qin [18]， Aoshima and Yata [3, 7] は，T_n の漸近正規性を弱スパイク固有値モデルのもとで与え，それに基づき，検定手法を与えた．しかしながら，T_n の漸近的な挙動は，第 3 節と同様に，弱スパイク固有値モデルに非常に敏感である．そこで，Ishii [22, 23] では，単一強スパイク固有値モデルのもと，Ishii, Yata and Aoshima [28] では，強スパイク固有値モデルのもと，新たな検定手法を与えた．

4.1. 単一強スパイク固有値モデルにおける二標本検定

Ishii [22, 23] では，T_n を単一強スパイク固有値モデルのもとで再評価することにより，新たな検定手法を提案した．ここでは，主に Ishii [22] による結果を紹介する．第 1 固有空間について，次の仮定をおく．

(C-i) $\dfrac{\lambda_{1(1)}}{\lambda_{1(2)}} = 1 + o(1),\ \boldsymbol{h}_{1(1)}^T \boldsymbol{h}_{1(2)} = 1 + o(1)\ \ (p \to \infty)$

ここで，(C-i) は，2 母集団が第 1 固有空間を共有していることを意味する．実際のデータ解析では，仮定 (C-i) の妥当性を確認する必要がある．Ishii, Yata and Aoshima [26] では，その確認方法として，第 1 固有空間の同等性検定手法を与えた．これについては，第 5 節で述べる．

$n_{\min} = \min\{n_1, n_2\}$ とする．このとき，T_n は単一強スパイク固有値モデルのもとで，次のように評価される．

補題 4.1 (Ishii [22])**.** 各母集団に (A-i) を仮定する．また，(C-i) を仮定する．H_0 のもと，「$p \to \infty$ だが，n_i は固定」もしくは，「$p \to \infty$ かつ $n_{\min} \to \infty$」の枠組みで，次が成り立つ．

$$\frac{T_n}{\lambda_{1(1)}} = (\bar{z}_{1(1)} - \bar{z}_{1(2)})^2 - \sum_{i=1}^{2} \frac{z_{o1(i)}^2}{n_i} + o_P(n_{\min}^{-1}) \tag{4.2}$$

(4.2) 式において，$\lambda_{1(1)}$ は未知のため，推定する必要がある．その推定について，Ishii [22] ではノイズ掃き出し法を適用し，新たな検定統計量の構築と漸近分布の導出をした．$u_n = (1/n_1 + 1/n_2)^{-1}$ とする．Ishii [22] では，以下の検定統計量を与えた．

$$F_a = u_n \frac{T_n + \sum_{i=1}^{2} \tilde{\lambda}_{1(i)}/n_i}{\sum_{i=1}^{2}(n_i - 1)\tilde{\lambda}_{1(i)}}$$

ここで，$\tilde{\lambda}_{1(i)}$ は，ノイズ掃き出し法を用いた $\lambda_{1(i)}$ の推定量である．更に，F_a の漸近帰無分布を次のように導出した．

定理 4.1 (Ishii [22])**.** (A-i) から (A-iii) を各母集団に仮定する．また，(C-i) を仮定する．H_0 のもと，$p \to \infty$ で次の結果が得られる．

$$F_a \Rightarrow \begin{cases} F_{1,\nu} & (\nu \text{ が固定のとき}) \\ \\ \chi_1^2 & (\nu \to \infty \text{ のとき}) \end{cases}$$

ここで，$\nu = n_1 + n_2 - 2$ である．

注意 4.1. (A-ii) が満たされない場合について，Ishii [23] では，クロスデータ行列法を用いた検定手法を与えた．

予め設定した $\alpha \in (0, 1/2)$ に対し，検定問題 (4.1) について，検定ルールを以下のように与えた．

$$H_0 \text{ を棄却する} \iff F_a \geq F_{1,\nu}(\alpha)$$

よって，定理 4.1 と同じ条件のもと，第 1 種の過誤は「$p \to \infty$ だが，n_i は固定」の枠組みで，漸近的に α となる．

検出力についても以下のような結果を与えた．いま，$\boldsymbol{\mu}_{12} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ とする．

定理 4.2 (Ishii [22])**.** (A-i) から (A-iii) を各母集団に仮定する．(C-i)，$\dfrac{n_{\min} \boldsymbol{\mu}_{12}^T \boldsymbol{\Sigma}_i \boldsymbol{\mu}_{12}}{\lambda_{1(1)}^2} = o(1)$ のもと，「$p \to \infty$ かつ $\nu \to \infty$」の枠組みで次が成り立つ．

$$\text{Power}\,(F_a) = 1 - C_{\chi_1^2}\left(\chi_1^2(\alpha) - u_n \frac{\|\boldsymbol{\mu}_{12}\|^2}{\lambda_{1(1)}}\right) + o(1)$$

ここで，$C_{\chi^2_1}(\cdot)$ は自由度 1 のカイ二乗分布の累積分布関数，$\chi^2_1(\alpha)$ は自由度 1 のカイ二乗分布の上側 α 点である．

系 4.1 (Ishii [22]). $u_n\lambda_{1(1)}^{-1}\|\boldsymbol{\mu}_{12}\|^2 \to \infty$ $(p \to \infty)$ を仮定する．このとき，定理 4.2 と同じ仮定のもと，「$p \to \infty$ だが，ν は固定」の枠組みで，次が成り立つ．

$$\text{Power}\,(F_a) = 1 + o(1)$$

4.2. データ変換を用いた二標本検定

Aoshima and Yata [10] は，データ変換を用いて，強スパイク固有値モデルにおける一般的な二標本検定を考案した．Ishii, Yata and Aoshima [28] では，データ変換を用いた一標本検定の応用として，二標本検定，更には多標本問題も扱った．いま，2 母集団において，標本数のバランスが取れている状況「$n_1/n_2 \to 1\,(n_1, n_2 \to \infty)$」を考える．$n_1 \le n_2$ とする．次のように，二標本問題を一標本問題に落とし込む．

$$\boldsymbol{x}_j = \boldsymbol{x}_{1j} - \boldsymbol{x}_{2j} \quad (j = 1, ..., n_1) \tag{4.3}$$

$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$ とする．通常，2 母集団のうちのどちらかが強スパイク固有値モデルであれば，上記の $\boldsymbol{x}_j$ 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，$\text{tr}(\boldsymbol{\Sigma}_1^2) \ge \text{tr}(\boldsymbol{\Sigma}_2^2)$ かつ $\liminf_{p\to\infty} \lambda^2_{\max}(\boldsymbol{\Sigma}_1)/\text{tr}(\boldsymbol{\Sigma}_1^2) > 0$ であれば，

$$\liminf_{p\to\infty} \frac{\lambda^2_{\max}(\boldsymbol{\Sigma})}{\text{tr}(\boldsymbol{\Sigma}^2)} \ge \liminf_{p\to\infty} \frac{\lambda^2_{\max}(\boldsymbol{\Sigma}_1)}{4\text{tr}(\boldsymbol{\Sigma}_1^2)} > 0$$

となり，$\boldsymbol{x}_j$ は強スパイク固有値モデルとなる．よって，$\boldsymbol{x}_j\,(j = 1, ..., n_1)$ に検定手法 (3.2) を適用することで，検定問題 (4.1) を考えることができる．

また，Ishii, Yata and Aoshima [28] では，2 母集団が独立でない場合にも上記の検定手法を応用した．例えば，遺伝子発現データ解析では，同一患者からの正常な細胞とがんの細胞などを解析対象とすることがある．そのような場合，母集団間の独立性を仮定した通常の二標本検定を適用することはできない．しかしながら，(4.3) 式でデータを変換した後に，検定手法 (3.2) を適用すれば，同一の患者からの標本であっても，正常細胞とがん細胞との差異を検出することができる．詳細は，Ishii, Yata and Aoshima [28] の第 5.2 節を参照のこと．

更に，多標本問題として，次の検定問題を扱った．

$$H_0 : \sum_{i=1}^{g} b_i\boldsymbol{\mu}_i = \boldsymbol{0} \quad \text{vs.} \quad H_1 : \sum_{i=1}^{g} b_i\boldsymbol{\mu}_i \neq \boldsymbol{0} \tag{4.4}$$

ここで，$b_i\,(i = 1, 2, ..., g)$ は，次元数 p に依存せず，0 でない既知の定数である．Bennett [15] や Anderson [1] で与えられている以下の変換を考える．

$$\boldsymbol{x}_j = b_1\boldsymbol{x}_{1j} + \sum_{i=2}^{g} b_i\sqrt{\frac{n_1}{n_i}}\Big(\boldsymbol{x}_{ij} - \frac{1}{n_1}\sum_{j=1}^{n_1}\boldsymbol{x}_{ij} + \frac{1}{\sqrt{n_1 n_i}}\sum_{j'=1}^{n_i}\boldsymbol{x}_{ij'}\Big), \quad j = 1, ..., n_1 \tag{4.5}$$

ここで，$n_1 = \min\{n_1, ..., n_g\}$ とする．このとき，次のようになる．

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

いま，$i=1,...,g$ に対し，$\pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ を仮定する．すると，$\boldsymbol{x}_1,...,\boldsymbol{x}_{n_1}$ 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. [32] は，弱スパイク固有値モデルのもとで，変換 (4.5) を用いた検定手法を提案した．Ishii, Yata and Aoshima [28] は，強スパイク固有値モデルのもとで検定手法を与えた．通常，どれか 1 つの母集団が強スパイク固有値モデルであれば，(4.5) 式の $\boldsymbol{x}_j$ も強スパイク固有値モデルとなることに注意する．よって，$\boldsymbol{x}_j\ (j=1,...,n_1)$ に検定手法 (3.2) を適用することで検定問題 (4.4) を考えることができる．なお，高次元データに対して母集団の正規性を仮定することは厳しいものである．Ishii, Yata and Aoshima [28] では，正規性が崩れたとき，検定手法 (3.2) の頑健性についても議論している．詳細は，Ishii, Yata and Aoshima [28] の第 4.2 節を参照のこと．

5. 共分散行列の同等性検定

2 母集団に対し，共分散行列の同等性検定を考える．母集団に関する設定は，第 4 節と同じものとする．高次元データの共分散行列に対し，次の同等性検定を考える．

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

Schott [34] は「$p/n_i \to c_i \in [0,\infty)$」という枠組みで，フロベニウスノルムをもとにした検定統計量を提案した．Srivastava and Yanagihara [37] は，ムーア・ペンローズ型の一般逆行列を用いた検定統計量を考案した．Aoshima and Yata [3] は $\mathrm{tr}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)$ をもとにした検定統計量を構築し，予め設定した第 1 種の過誤と検出力をもつような標本数決定問題を議論した．上記の先行研究は，いずれも弱スパイク固有値モデルのもと，「$p \to \infty$ かつ $n_i \to \infty$」の枠組みで理論を展開している．

本節では，Ishii, Yata and Aoshima [26, 27] で与えた共分散行列の同等性検定を紹介する．Ishii, Yata and Aoshima [26] では，単一強スパイク固有値モデルのもと，「$p \to \infty$ だが，n_i は固定」の枠組みで新たな検定手法を提案した．Ishii, Yata and Aoshima [26] では，強スパイク固有値モデルにおいて汎用性が高く，「$p \to \infty$ かつ $n_i \to \infty$」の枠組みで新たな検定手法を提案した．

5.1. 単一強スパイク固有値モデルにおける共分散行列の同等性検定

Ishii, Yata and Aoshima [26] では，単一強スパイク固有値モデルのもとで，新たな高次元共分散行列の同等性検定法を考案した．Ishii, Yata and Aoshima [26] では，はじめに，第 1 固有空間の同等性検定に関する検定手法を与えた．次の検定問題を考える．

$$H_0 : (\lambda_{1(1)}, \boldsymbol{h}_{1(1)}) = (\lambda_{1(2)}, \boldsymbol{h}_{1(2)}) \quad \text{vs.} \quad H_1 : (\lambda_{1(1)}, \boldsymbol{h}_{1(1)}) \neq (\lambda_{1(2)}, \boldsymbol{h}_{1(2)}) \tag{5.2}$$

$\nu_1 = n_1 - 1$ と $\nu_2 = n_2 - 1$ とする．このとき，定理 2.1 より，固有値に関して次の結果が得られる．

系 5.1 (Ishii, Yata and Aoshima [26])**.** (A-i) から (A-iii) を各母集団に仮定する．このとき，「$p \to \infty$ だが，n_i は固定」の枠組みで，以下が成り立つ．

$$\frac{\tilde{\lambda}_{1(1)}/\lambda_{1(1)}}{\tilde{\lambda}_{1(2)}/\lambda_{1(2)}} \Rightarrow F_{\nu_1,\nu_2}$$

(2.5) 式より，$\tilde{\boldsymbol{h}}_{1(i)}$ をノイズ掃き出し法を用いた $\boldsymbol{h}_{1(i)}$ の推定量とする．そのとき，固有ベクトルに関して，次の結果を得る．

補題 5.1 (Ishii, Yata and Aoshima [26])**.** (A-i) と (A-ii) を各母集団に仮定する．このとき，「$p \to \infty$ だが，n_i は固定」または，「$p \to \infty$ かつ $n_i \to \infty$」のもと，次が成り立つ．

$$|\tilde{\boldsymbol{h}}_{1(1)}^T \tilde{\boldsymbol{h}}_{1(2)}| = |\boldsymbol{h}_{1(1)}^T \boldsymbol{h}_{1(2)}| + o_P(1)$$

補題 5.1 より，2 母集団間で，第 1 固有ベクトルは内積の意味で一致性をもつ．この結果を用いて，Ishii, Yata and Aoshima [26] では，次のような検定統計量を提案した．

$$F_1 = \frac{\tilde{\lambda}_{1(1)}}{\tilde{\lambda}_{1(2)}} \tilde{h}_*$$

ここで，$\tilde{h} = \max\{|\tilde{\boldsymbol{h}}_{1(1)}^T \tilde{\boldsymbol{h}}_{1(2)}|, |\tilde{\boldsymbol{h}}_{1(1)}^T \tilde{\boldsymbol{h}}_{1(2)}|^{-1}\}$ とし，

$$\tilde{h}_* = \begin{cases} \tilde{h} & (\tilde{\lambda}_{1(1)} \geq \tilde{\lambda}_{1(2)} \text{のとき}) \\ \tilde{h}^{-1} & (\text{その他}) \end{cases}$$

である．F_1 の漸近帰無分布として以下を得る．

定理 5.1 (Ishii, Yata and Aoshima [26])**.** (A-i) から (A-iii) を各母集団に仮定する．H_0 のもと，「$p \to \infty$ だが，n_i は固定」の枠組みで，以下が成り立つ．

$$F_1 \Rightarrow F_{\nu_1,\nu_2}$$

よって，予め設定した $\alpha \in (0, 1/2)$ に対し，検定問題 (5.2) について，検定ルールを次のように与えた．

$$H_0 \text{を棄却する} \Longleftrightarrow F_1 \notin [\{F_{\nu_2,\nu_1}(\alpha/2)\}^{-1}, F_{\nu_1,\nu_2}(\alpha/2)] \tag{5.3}$$

すると，定理 5.1 と同じ仮定のもと，「$p \to \infty$ だが，n_i は固定」の枠組みで，以下が成り立つ．

$$\text{Size } (F_1) \ = \alpha + o(1)$$

第 4 節で述べた二標本検定において，F_1 を用いて，(C-i) の妥当性を確認できる．

以上の結果を踏まえ，検定問題 (5.1) を考える．$\kappa_i = \sum_{j=2}^{p} \lambda_{j(i)}$ $(i = 1, 2)$ とする．ノイズ掃き出し法を用いて，κ_i を $\tilde{\kappa}_i = \text{tr}(\boldsymbol{S}_{D(i)}) - \tilde{\lambda}_{1(i)}$ で推定する．ここで，$\boldsymbol{S}_{D(i)}$ は，母集団 π_i の双対標本共分散行列である．(A-i) と (A-ii) のもとで，$\tilde{\kappa}_i$ は「$p \to \infty$ だが，n_i は固定」の枠組みで，κ_i の一致推定量となる．いま，$\tilde{\gamma} = \max\{\tilde{\kappa}_1/\tilde{\kappa}_2, \tilde{\kappa}_2/\tilde{\kappa}_1\}$ とする．Ishii, Yata and Aoshima [26] では次の検定統計量を構築した．

$$F_2 = \frac{\tilde{\lambda}_{1(1)}}{\tilde{\lambda}_{1(2)}} \tilde{h}_* \tilde{\gamma}_* \ (= F_1 \tilde{\gamma}_*)$$

ここで，

$$\tilde{\gamma}_* = \begin{cases} \tilde{\gamma} & (\tilde{\lambda}_{1(1)} \geq \tilde{\lambda}_{1(2)} \text{のとき}) \\ \tilde{\gamma}^{-1} & (\text{その他}) \end{cases}$$

このとき，F_2 の漸近帰無分布について，次の結果を得た．

定理 5.2 (Ishii, Yata and Aoshima [26])**.** (A-i) から (A-iii) を各母集団に仮定する．H_0 のもと，「$p \to \infty$ だが，n_i は固定」の枠組みで次が成り立つ．

$$F_2 \Rightarrow F_{\nu_1,\nu_2}$$

したがって，F_1 の検定ルール (5.3) を F_2 に適用することで，検定問題 (5.1) に対して検定を行う．すると，「$p \to \infty$ だが，n は固定」の枠組みで，第 1 種の過誤が漸近的に α となる．検定統計量 F_2 の構築について，そのアイディアは，高次元固有空間を第 1 固有空間とそれ以降の固有空間とで分割し，差異を検出することである．このアイディアは，Ishii, Yata and Aoshima [27] で与えた検定手法でも鍵となる．

5.2. 強スパイク固有値モデルにおける共分散行列の同等性検定

Ishii, Yata and Aoshima [27] では，強スパイク固有値モデルにおいて非常に汎用性の高い検定手法を与えた．Li and Chen [31] は，$\mathrm{tr}\{(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^2\} = \Delta_v$ に基づいた検定統計量を次のように提案した．

$$T_{\mathrm{LC}} = W_1 + W_2 - 2\mathrm{tr}(\boldsymbol{S}_{1n_1}\boldsymbol{S}_{2n_2})$$

ここで，$W_i = \{n_i(n_i-1)\}^{-1}\sum_{s\neq t}(\boldsymbol{x}_{is}^T\boldsymbol{x}_{it})^2 - 2\{n_i(n_i-1)(n_i-2)\}^{-1}\sum_{s\neq t\neq v}\boldsymbol{x}_{is}^T\boldsymbol{x}_{it}\boldsymbol{x}_{it}^T\boldsymbol{x}_{iv} + \{n_i(n_i-1)(n_i-2)(n_i-3)\}^{-1}\sum_{s\neq t\neq v\neq w}\boldsymbol{x}_{is}^T\boldsymbol{x}_{it}\boldsymbol{x}_{iv}^T\boldsymbol{x}_{iw}$ である．そのとき，$E(T_{\mathrm{LC}}) = \Delta_v$ となる．Li and Chen [31] は，弱スパイク固有値モデルといくつかの正則条件を仮定し，帰無仮説のもと，検定統計量の漸近正規性を示した．しかしながら，T_{LC} の漸近正規性は弱スパイク固有値モデルに対して非常に敏感である．そこで，Ishii, Yata and Aoshima [27] では，Δ_v をもとにした全く新しい検定手法を提案した．はじめに，単一強スパイク固有値モデルのもと，次の結果を得た．

$$\Delta_v = (\lambda_{1(1)} - \lambda_{1(2)})^2 + 2\lambda_{1(1)}\lambda_{1(2)}\{1 - (\boldsymbol{h}_{1(1)}^T\boldsymbol{h}_{1(2)})^2\} + o(\lambda_{1(1)}^2 + \lambda_{1(2)}^2) \tag{5.4}$$

(5.4) 式の $\lambda_{1(i)}$, $\boldsymbol{h}_{1(i)}$ をノイズ掃き出し法で推定し，単一強スパイク固有値モデルのもとで，次の統計量を考える．

$$V = \frac{(\tilde{\lambda}_{1(1)} - \tilde{\lambda}_{1(2)})^2 + 2\tilde{\lambda}_{1(1)}\tilde{\lambda}_{1(2)}\{1 - (\tilde{\boldsymbol{h}}_{1(1)}^T\tilde{\boldsymbol{h}}_{1(2)})^2\}}{\sum_{i=1}^2 2\tilde{\lambda}_{1(i)}^2/(n_i-1)}$$

$q = \min\{p, n_1, n_2\}$ とおく．新たに与えた統計量 V に対し，単一強スパイク固有値モデル，いくつかの正則条件と H_0 のもと，$q \to \infty$ で次が成り立つ．

$$V \Rightarrow \chi_1^2$$

次に，単一強スパイク固有値モデルから強スパイク固有値モデルへの拡張を行う．いま，V の第 2 項 $1 - (\tilde{\boldsymbol{h}}_{1(1)}^T\tilde{\boldsymbol{h}}_{1(2)})^2$ に注目する．帰無仮説のもと，第 2 項目は 0 に収束するべき量である．しかしながら，$\boldsymbol{h}_{1(i)}$ の推定にノイズ掃き出し法を単純に用いると，強スパイク固有値モデルのもとではバイアスが生じる．その結果，強スパイク固有値モデルに対する検定統計量として V を用いると，第 1 種の過誤を抑えられない．そこで，Ishii, Yata and Aoshima [27] では，V の第 2 項のバイアス補正を行い，強スパイク固有値モデルに対する検定統計量として，次を与えた．

$$T_{\mathrm{NR}} = \frac{(\tilde{\lambda}_{1(1)} - \tilde{\lambda}_{1(2)})^2 + 2\tilde{\lambda}_{1(1)}\tilde{\lambda}_{1(2)}\{1 - \min\{1, (\tilde{\boldsymbol{h}}_{1(1)}^T\tilde{\boldsymbol{h}}_{1(2)})^2\}\}^{1+\hat{\eta}}}{\sum_{i=1}^2 2\tilde{\lambda}_{1(i)}^2/(n_i-1)}$$

ここで，$\hat{\eta} = \hat{\Psi}_{2(1)}^{1/2}/\tilde{\lambda}_{1(1)} + \hat{\Psi}_{2(2)}^{1/2}/\tilde{\lambda}_{1(2)}$ であり，$\hat{\Psi}_{2(i)}$ は，$\Psi_{2(i)} = \sum_{s=2}^{p} \lambda_{s(i)}^2$ のクロスデータ行列法を用いた推定量である．また，検定統計量 T_{NR} の漸近帰無分布は，強スパイク固有値モデルと適当な正則条件のもと，自由度 1 のカイ二乗分布となる．

しかしながら，T_{NR} はその形からも見てとれるように，第 1 固有空間の同等性検定に対する検定統計量となっている．Ishii, Yata and Aoshima [27] では，2 番目以降の固有空間の差異も検出するため，$\kappa_i = \sum_{s=2}^{p} \lambda_{s(i)}$ と $\Psi_{2(i)}$ に着目し，差異を取り出した．$\kappa_\star = (\kappa_1/\kappa_2 + \kappa_2/\kappa_1)/2$，$\psi_\star = (\Psi_{2(1)}/\Psi_{2(2)} + \Psi_{2(2)}/\Psi_{2(1)})/2$ とおくと，両者は何れも 1 以上となり，帰無仮説のもとで 1 となる量であることに注意する．ここで，$\kappa_\star$，$\psi_\star$ は未知なので，ノイズ掃き出し法とクロスデータ行列法を用いて得られた一致推定量を $\tilde{\kappa}_\star$，$\tilde{\psi}_\star$ とする．以上より，Ishii, Yata and Aoshima [27] では，次のような検定統計量を構築した．

$$T_{\text{IYA}} = \tilde{\kappa}_\star \tilde{\psi}_\star T_{\text{NR}}$$

定理 5.3 (Ishii, Yata and Aoshima [27])**.** 各母集団に (1.1)，(A-iii)，(A-iv) を仮定する．H_0 のもと，$q \to \infty$ で次が成り立つ．

$$T_{\text{IYA}} \Rightarrow \chi_1^2$$

よって，予め設定した $\alpha \in (0, 1/2)$ に対し，検定問題 (5.1) について検定ルールを次のように与えた．

$$H_0 \text{ を棄却する} \Longleftrightarrow T_{\text{IYA}} > \chi_1^2(\alpha)$$

上記の検定ルールを用いると，定理 5.3 と同じ条件のもと，$q \to \infty$ のとき，第 1 種の過誤が漸近的に α となる．さらに，T_{IYA} の検出力について，以下の結果を与えた．

定理 5.4 (Ishii, Yata and Aoshima [27])**.** 各母集団に (1.1)，(A-iii)，(A-iv) を仮定する．$q \to \infty$ で以下が成り立つ．

$$\text{Power}\,(T_{\text{IYA}}) \geq 1 - G_{\omega^2}\left(\kappa_\star^{-1}\psi_\star^{-1}\chi_1^2(\alpha)\right) + o(1)$$

ここで，$\omega = (\lambda_{1(1)} - \lambda_{1(2)})/(2\lambda_{1(1)}^2/n_1 + 2\lambda_{1(2)}^2/n_2)^{1/2}$，$G_{\omega^2}(\cdot)$ は非心率 ω^2，自由度が 1 の非心カイ二乗分布の累積分布関数を表す．

系 5.2 (Ishii, Yata and Aoshima [27])**.** 各母集団に (1.1) と (A-iv) を仮定する．$\limsup_{p\to\infty} |\boldsymbol{h}_{1(1)}^T \boldsymbol{h}_{1(2)}| < 1$ もしくは，$\liminf_{p\to\infty} |\lambda_{1(1)}/\lambda_{1(2)}| > 0$ のもと，$q \to \infty$ で $\text{Power}\,(T_{\text{IYA}}) = 1 + o(1)$ となる．

6. データ変換を用いた高次元判別分析

高次元データの 2 群の判別を考える．各母集団 π_i の設定は，第 4 節と同じとする．母集団 π_i から，$n_i\ (\geq 4)$ 個の訓練データ $\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}$ を無作為に抽出する．判別対象の個体を $\boldsymbol{x}_0$ とし，訓練データとは互いに独立であるとする．π_1 に属する $\boldsymbol{x}_0$ を π_2 に誤判別する確率を $e(1)$，π_2 に属する $\boldsymbol{x}_0$ を π_1 に誤判別する確率を $e(2)$ と表記する．

高次元小標本データに対しては，フィッシャーの線形判別方式や 2 次判別方式を適用することはできない．高次元データに対する判別分析手法は多々提案されている．特に，2 群に共分散行列の同等性を仮定しない場合，Dudoit et al. [21] は標本共分散行列の対角成分のみを用いた判別関数，Chan and Hall [17]，Aoshima and Yata [6] はユークリッド距離に基づく線形判別関数，Aoshima and Yata [3, 8] は高次元小標本の幾何学的

表現に基づく2次判別関数を与えた．また，Aoshima and Yata [12] は，高次元データの特徴に基づく2次判別関数のクラスを考え，誤判別確率に関する一致性と漸近正規性が成り立つ条件を導出し，高次元における最適性を議論した．上記の先行研究は，2群がともに弱スパイク固有値モデルのもと，判別方式の精度を保証した．しかしながら，2群のどちらかが強スパイク固有値モデルをもつ場合，強スパイクする固有値がノイズとなり，判別の精度が低下する．詳細は青嶋 [2] を参照のこと．

強スパイク固有値モデルに対して，Aoshima and Yata [11] は，第3節で与えたデータ変換を用いて，Aoshima and Yata [6] による線形判別方式を改良した．また，「$p \to \infty$ かつ $n_i \to \infty$」の枠組みで，高い精度保証を与えた．さらに，Ishii [24] では，データ変換による線形判別方式を「$p \to \infty$ だが，n_i は固定」の枠組みへ拡張した．Ishii, Yata and Aoshima [29] では，Aoshima and Yata [3, 8] による2次判別方式に対し，データ変換を用いた新たな高次元2次判別方式を与えた．

本節では，「$p \to \infty$ だが，n_i は固定」の枠組みで，Ishii [24] と Ishii, Yata and Aoshima [29] による判別方式を解説する．簡単のため，$\boldsymbol{h}_{1(1)} = \boldsymbol{h}_{1(2)}$ と仮定する．

6.1. データ変換を用いた高次元線形判別方式

Aoshima and Yata [6] は，$\Delta_m = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ に基づく判別分析「Distance-Based Discriminant Analysis (DBDA)」を考え，判別関数を次のように与えた．

$$D(\boldsymbol{x}_0) = \Big(\boldsymbol{x}_0 - \frac{\overline{\boldsymbol{x}}_{1n_1} + \overline{\boldsymbol{x}}_{2n_2}}{2}\Big)^T (\overline{\boldsymbol{x}}_{2n_2} - \overline{\boldsymbol{x}}_{1n_1}) - \frac{\mathrm{tr}(\boldsymbol{S}_{1n_1})}{2n_1} + \frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{2n_2} \tag{6.1}$$

判別方式は，$D(\boldsymbol{x}_0) < 0$ のとき $\boldsymbol{x}_0 \in \pi_1$，$D(\boldsymbol{x}_0) \geq 0$ のとき $\boldsymbol{x}_0 \in \pi_2$ である．いま，次の条件を考える．

$$\frac{\max_{j=1,2} \mathrm{tr}(\boldsymbol{\Sigma}_j^2)}{\Delta_m^2} \to 0 \ \ (p \to \infty) \tag{6.2}$$

DBDA は条件 (6.2) のもと，「$p \to \infty$ だが，n_i は固定」の枠組みで，次の一致性をもつ．

$$e(1) \to 0, \quad e(2) \to 0 \tag{6.3}$$

つまり，DBDA は，条件 (6.2) のもと，高次元で完全分類を達成する．ところで，条件 (6.2) は，強スパイク固有値モデルのもと，以下を意味する．

$$\frac{\max_{j=1,2} \{\lambda_{\max}(\boldsymbol{\Sigma}_j)\}^2}{\Delta_m^2} \to 0 \ \ (p \to \infty)$$

つまり，Δ_m が強スパイクするノイズより有意に大きくならなければならない．よって，強スパイク固有値モデルの場合，条件 (6.2) は満たされ難く，判別関数 (6.1) の性能は低下する．そこで，Ishii [24] は，Aoshima and Yata [11] と同様に，データ変換によって，判別関数 (6.1) から強スパイクするノイズを除去することを考えた．さらに，強スパイク固有値モデルのもと，「$p \to \infty$ だが，n_i は固定」の枠組みで高い精度保証を与えた．Ishii [24] では，次のような判別関数を与えた．

$$\begin{aligned} D_{\mathrm{DT}}(\boldsymbol{x}_0) =& \|\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}\|^2 - \{(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1})^T \tilde{\boldsymbol{h}}_{1(2)}\}^2 - \|\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}\|^2 + \{(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2})^T \tilde{\boldsymbol{h}}_{1(1)}\}^2 \\ &+ \{\boldsymbol{x}_0^T \tilde{\boldsymbol{h}}_{1(2)}\}^2 - \{\boldsymbol{x}_0^T \tilde{\boldsymbol{h}}_{1(1)}\}^2 + \sum_{i=1}^{2} (-1)^i \frac{\mathrm{tr}(\boldsymbol{S}_{in_i}) - \tilde{\lambda}_{1(i)}}{n_i} \end{aligned} \tag{6.4}$$

判別方式は，$D_{\mathrm{DT}}(\boldsymbol{x}_0) < 0$ のとき $\boldsymbol{x}_0 \in \pi_1$，$D_{\mathrm{DT}}(\boldsymbol{x}_0) \geq 0$ のとき $\boldsymbol{x}_0 \in \pi_2$ である．判別関数 (6.4) で工夫している点は，訓練データに対するデータ変換の適用方法である．例えば，$\overline{\boldsymbol{x}}_{1n_1}$ には，第 1 群の訓練データを含まない，第 2 群の第 1 固有ベクトルの推定量 $\tilde{\boldsymbol{h}}_{1(2)}$ をかけている．これにより，データ変換を用いる際に生じるバイアスを避けることができる．ここで，次の条件を仮定する．

(C-ii) $\boldsymbol{\mu}_{i'}^T \boldsymbol{\Sigma}_i \boldsymbol{\mu}_{i'} = o\big(\lambda_{1(i)} \min\big\{\Delta_m^2/\lambda_{1(\max)}, \Delta_m\big\}\big)$ $(p \to \infty;\, i = 1, 2;\, i' = 1, 2)$

(C-iii) $\dfrac{\max_{j=1,2} \Psi_{2(j)}}{\Delta_m^2} \to 0$ $(p \to \infty)$

ただし，$\lambda_{1(\max)} = \max\{\lambda_{1(1)}, \lambda_{1(2)}\}$ である．強スパイク固有値モデルのもとで，(C-iii) は条件 (6.2) よりも緩い仮定であることに注意する．Ishii [24] は，次の結果を与えた．

定理 6.1 (Ishii [24])**.** (A-i) と (A-iv) を各母集団に仮定する．さらに，(C-ii) と (C-iii) を仮定する．判別関数 (6.4) は「$p \to \infty$ だが，n_i は固定」の枠組みで，一致性 (6.3) を与える．

6.2. データ変換を用いた高次元 2 次判別方式

Aoshima and Yata [3, 6] は，高次元小標本の幾何学的表現に基づく判別分析「Geometrical Quadratic Discriminant Analysis (GQDA)」を考え，判別関数を次のように与えた．

$$G(\boldsymbol{x}_0) = \frac{p\|\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}\|^2}{\mathrm{tr}(\boldsymbol{S}_{1n_1})} - \frac{p\|\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}\|^2}{\mathrm{tr}(\boldsymbol{S}_{2n_2})} - \frac{p}{n_1} + \frac{p}{n_2} - p \log\left\{\frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{\mathrm{tr}(\boldsymbol{S}_{1n_1})}\right\} \tag{6.5}$$

GQDA は，$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ であっても，$\boldsymbol{\Sigma}_1$ と $\boldsymbol{\Sigma}_2$ の差異を用いることで，一致性 (6.3) をもつ．詳細は，青嶋・矢田 [4, 5] を参照のこと．しかしながら，第 6.1 節と同様に，強スパイク固有値モデルの場合，判別関数 (6.5) の性能は低下する．そこで，Ishii, Yata and Aoshima [29] は，データ変換を用いて，高次元 2 次判別方式を次のように与えた．

$$\begin{aligned} &G_{\mathrm{DT}}(\boldsymbol{x}_0) \\ &= p\frac{\|\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}\|^2 - \{(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1})^T \tilde{\boldsymbol{h}}_{1(2)}\}^2}{\mathrm{tr}(\boldsymbol{S}_{1n_1}) - \tilde{\lambda}_{1(1)}} - p\frac{\|\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}\|^2 - \{(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2})^T \tilde{\boldsymbol{h}}_{1(1)}\}^2}{\mathrm{tr}(\boldsymbol{S}_{2n_2}) - \tilde{\lambda}_{1(2)}} \\ &\quad - \frac{p}{n_1} + \frac{p}{n_2} - p \log\left\{\frac{\mathrm{tr}(\boldsymbol{S}_{2n_2}) - \tilde{\lambda}_{1(2)}}{\mathrm{tr}(\boldsymbol{S}_{1n_1}) - \tilde{\lambda}_{1(1)}}\right\} \end{aligned} \tag{6.6}$$

判別方式は，$G_{\mathrm{DT}}(\boldsymbol{x}_0) < 0$ のとき $\boldsymbol{x}_0 \in \pi_1$，$G_{\mathrm{DT}}(\boldsymbol{x}_0) \geq 0$ のとき $\boldsymbol{x}_0 \in \pi_2$ である．いま，

$$\Delta_{v\star} = \Delta_m + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^2}{2 \max_{i=1,2} \mathrm{tr}(\boldsymbol{\Sigma}_i)}$$

とおく．ここで，次の条件を仮定する．

(C-iv) $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{h}_{1(i)} = o(\min\{\Delta_{v\star}^{1/2},\, \Delta_{v\star}/\lambda_{1(\max)}^{1/2}\})$ $(p \to \infty;\, i = 1, 2)$

(C-v) $\dfrac{\max_{j=1,2} \Psi_{2(j)}}{\Delta_{v\star}^2} \to 0$ $(p \to \infty)$

そのとき，Ishii, Yata and Aoshima [29] は，次の結果を与えた．

定理 6.2 (Ishii, Yata and Aoshima [29])**.** (A-i) と (A-iv) を各母集団に仮定する．さらに，(C-iv) と (C-v) を仮定する．判別関数 (6.6) は「$p \to \infty$ だが，n_i は固定」の枠組みで，一致性 (6.3) をもつ．

したがって，強スパイク固有値モデルにおいて，判別関数 (6.6) は高い精度保証を与える．Ishii, Yata and Aoshima [29] は，判別関数 (6.4) と (6.6) の性能を第 1 節で紹介した 2 つの遺伝子発現データにより検証した．ここでは，比較の対象として，Dudoit et al. [21] による線形判別方式 (DLDA) 及び 2 次判別方式 (DQDA) を用いる．1 つ抜き交差確認法 (Leave-One-Out Cross Validation: LOOCV) によって誤判別の割合を計算した．Singh et al. [35] の前立腺がんデータ (π_1: Prostate cancer，π_2: Non-tumor) と，ハンチントン病データ (π_1: Huntington's disease，π_2: Normal) に対し，各群における誤判別の割合 $\bar{e}_1$, $\bar{e}_2$ と，その平均値 $\bar{e} = (\bar{e}_1 + \bar{e}_2)/2$ を表 2 に纏めた．

表 2. LOOCV による誤判別の割合

前立腺がんデータ (Singh et al. [35], $p = 12625$)				
	$G_{\mathrm{DT}}(\boldsymbol{x}_0)$	$D_{\mathrm{DT}}(\boldsymbol{x}_0)$	DLDA	DQDA
$\bar{e}(1)$	0.26	0.08	0.34	0.36
$\bar{e}(2)$	0.25	0.173	0.404	0.365
$\bar{e}$	0.255	0.127	0.372	0.363
ハンチントン病データ (Borovecki et al. [16], $p = 22283$)				
	$G_{\mathrm{DT}}(\boldsymbol{x}_0)$	$D_{\mathrm{DT}}(\boldsymbol{x}_0)$	DLDA	DQDA
$\bar{e}(1)$	0.059	0.235	0.118	0.235
$\bar{e}(2)$	0.	0.071	0.143	0.071
$\bar{e}$	0.03	0.153	0.131	0.153

表 1 や図 1 で見たように，上記 2 つの高次元小標本データは，強スパイク固有値モデルをもつ．強スパイクする固有空間の構造を生かした判別関数 (6.4) と (6.6) は，良い結果を与えている．特に，2 つ目のハンチントン病データは，2 群間の共分散行列に差異のあるデータであり，判別関数 (6.6) は，ノイズとなる強スパイクする固有値を取り除いた上で，残りの固有空間の差異の情報を上手く汲み取ることができたと考えられる．

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

[4] 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.

[5] 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.

[6] Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data, *Annals of the Institute of Statistical Mathematics*, **66**, 983-1010.

[7] Aoshima, M. and Yata, K. (2015). Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions, *Methodology and Computing in Applied Probability*, **17**, 419-439.

[8] Aoshima, M. and Yata, K. (2015). Geometric classifier for multiclass, high-dimensional data, *Sequential Analysis, Special Issue: Celebrating Seventy Years of Charles Stein's 1945 Seminal Paper on Two-Stage Sampling*, **34**, 279-294.

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

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

[11] Aoshima, M. and Yata, K. (2018). Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models, *Annals of the Institute of Statistical Mathematics*, in press (doi:10.1007/s10463-018-0655-z).

[12] Aoshima, M. and Yata, K. (2018). High-dimensional quadratic classifiers in non-sparse settings, *Methodology and Computing in Applied Probability*, in press (doi:10.1007/s11009-018-9646-z).

[13] Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H. and Marron, J. S. (2018). A survey of high dimension low sample size asymptotics, *Australian and New Zealand Journal of Statistics, Special Issue: in Honour of Peter Gavin Hall*, **60**, 4-19.

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

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

[16] Borovecki, F., Lovrecic, L., Zhou, J., Jeong, H., Then, F., Rosas, H.D., Hersch, S.M., Hogarth, P., Bouzou, B., Jensen, R.V. and Krainc, D. (2005). Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 11023-11028.

[17] Chan, Y.-B. and Hall, P. (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings, *Biometrika*, **96**, 469-478.

[18] Chen, S.X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing, *The Annals of Statistics*, **38**, 808-835.

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

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

[21] Dudoit, S., Fridlyand, J. and Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, **97**, 77-87.

[22] Ishii, A. (2017). A two-sample test for high-dimension, low-sample-size data under the strongly spiked eigenvalue model, *Hiroshima Mathematical Journal*, **47**, 273-288.

[23] Ishii, A. (2017). A high-dimensional two-sample test for non-Gaussian data under a strongly spiked eigenvalue model, *Journal of the Japan Statistical Society*, **47**, 273-291.

[24] Ishii, A. (2019). A classifier under the strongly spiked eigenvalue model in high-dimension, low-sample-size context, *Communications in Statistics. Theory and Methods*, in press.

[25] 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 Journal of Applied Statistics, Special Issue: Modern Statistical Methodologies in the Cutting Edge of Science (ed. Mukhopadhyay, N.)* , **5**, 81-94.

[26] 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.

[27] Ishii, A., Yata, K., Aoshima, M. (2018). Equality tests of high-dimensional covariance matrices under the strongly spiked eigenvalue model, *Journal of Statistical Planning and Inference*, revised.

[28] 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.

[29] Ishii, A., Yata, K., Aoshima, M. (2019). A quadratic classifier for high-dimension, low-sample-size data under the strongly spiked eigenvalue model, submitted.

[30] Johnstone, I.M. (2001). On the distribution of the largest eigenvalue in principal components analysis, *The Annals of Statistics*, **29**, 295-327.

[31] Li, J. and. Chen, S.X. (2012). Two sample tests for high-dimensional covariance matrices, *The Annals of Statistics*, **40**, 908-940.

[32] 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.

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

[34] Schott, J.R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes, *Computational Statistics and Data Analysis*, **51**, 6535-6542.

[35] Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R. and Sellers, W.R. (2002). Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, **1**, 203-209.

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

[37] Srivastava, M.S. and Yanagihara, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension, *Journal of Multivariate Analysis*, **101**, 1319-1329.

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

[39] 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.

[40] 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.

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