論　説

高次元小標本における統計的推測

青　嶋　　誠
矢　田　和　善

1　はじめに

ゲノム科学・情報工学・金融工学等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 マイクロアレイのようなゲノムデータは，次元数が数万にものぼり，それに対して標本数は 100 にも満たない事例が多く見られる．1990 年代頃から，多変量統計解析の理論と方法論の枠組みで高次元データを捉えようとする研究が各方面から始まった．しかしながら，理論的制約から次元数は標本数よりも小さいことが前提とされ，標本数よりも次元数の方が圧倒的に大きい高次元小標本データを理論的に扱うことはできなかった．従来の多変量統計解析法では，高次元小標本データという解析対象はデータの次元数に対して標本数が十分ではないために，信頼のおける解を与えることはできない．それゆえ，高次元小標本データに対する統計的推測の理論と方法論の開発が，急務とされている．

高次元小標本データを対象とした研究を展開するためには，まずは，これまでの多変量統計解析の枠組みを外す必要がある．なぜならば，従来の多変量統計解析の枠組みに囚われた研究では，高次元データ空間の特徴を捉えることができず，高次元データが本来もつ豊富な情報を生かせないからである．

2000 年以降，確率論と理論物理の方面から，ランダム行列の理論に基づき，高次元統計解析の基礎理論に重要な結果がもたらされた．Baik *et al.* [14], Baik and Silverstein [15], Johnstone [35], Paul [42] 等による，標本固有値の漸近的性質の研究である．高次元におけるランダム行列の理論は，Bai and Silverstein [13] が詳しい．しかしながら，これらの研究は，データの次元数 p と標本数 n が $n/p \to c > 0$ 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もしくは類する条件を仮定していた．次元数が数万にものぼり標本数は高々 100 程度といった高次元小標本においては，標本数を次元数と同程度には仮定できない．それゆえ，n が p に依存しないような設定で，もしくは，$n = n(p)$ であっても $n/p \to 0$ となる設定で，高次元漸近理論を展開する必要がある．

Yata and Aoshima [54] は，高次元小標本のもとで ‘クロスデータ行列法’ とよばれるノンパラメトリックな方法論を考案した．クロスデータ行列法は，データセットを 2 分割して掛け合わせ，クロスデータ行列という非正則な行列を定義し，これに基づいて高速かつ高精度な汎用性の高い推測を可能にする．Yata and Aoshima [54] は，クロスデータ行列の特異値分解に基づいて固有値の推定と漸近分布を求め，さらに固有ベクトルや主成分スコアの推定も与えて，それらが高次元小標本のもとで一致性をもつことを証明した．一方，高次元小標本データ空間を幾何学的に捉えるための研究もある．

Ahn *et al.* [1], Hall *et al.* [31], Yata and Aoshima [56] は，標本数 n を高々 100 程度に固定して次元数 p を $p \to \infty$ としたときの高次元小標本データの漸近的振る舞いを考察し，高次元データ空間の幾何学的表現を見つけている．Ahn *et al.* [1] や Hall *et al.* [31] は，母集団分布が正規分布もしくは類する条件を満たすものであることを仮定した．それに対し，Yata and Aoshima [56] は分布の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．非正規性の尺度を境にした，固有空間の球面集中現象と座標軸集中現象である．

Aoshima and Yata [8], [9] は，高次元小標本における統計的推測に，幾何学的表現に基づいた各種方法論を考案し，高次元漸近正規性・標本数の設計・精度保証に至るまでの一連の基礎理論を築き上げた．Aoshima and Yata [8], [9] の研究は多岐にわたり，高次元球面上の与えられたバンド幅の信頼領域，高次元二標本問題，高次元共分散行列の推定・検定，高次元判別分析，高次元回帰分析，高次元変数選択問題，パスウェイ解析等，高次元小標本データに対する統計的推測の 8 つの重要な問題に，高次元データの豊富な情報を生かすための理論と方法論を与えている．Fujikoshi *et al.* [29] も高次元データにおける統計的推測の問題を部分的に扱っているが，母集団分布に正規分布を仮定して，$n/p \to c > 0$ なる高次元大標本の枠組みであることに注意する．

工学における機械学習の方面から，高次元データ解析の方法論に関する研究が膨大にある．高次元データへのアプローチは，高次元データが有する非線形性の扱いに特徴がある．例えば，回帰分析と判別分析は，機械学習の領域では教師あり学習という立場で広く研究され，代表的な手法に Vapnik [47] が考案したサポートベクトルマシン (SVM) がある．高次元データ解析において疎な解が得られ，汎化性能がよいことも知られているが，理論的に精度を保証するものではなく，また，計算コストに問題を抱える部分がある．機械学習の詳細は，例えば，Bishop [19], Hastie *et al.* [33] 等を参照されたい．

本稿では，数理統計学の立場から，高次元小標本における理論と方法論の最新の研究を紹介する．

2 節では，高次元小標本における統計的推測の鍵となる，データの幾何学的表現について紹介する．従来の多変量統計解析における推測では，大標本漸近理論に見られるように，データが中心に集まっていく法則を用いる．これに対し，高次元小標本における統計的推測では様相が大きく異なり，データが中心から離れていく法則を用いることになる．

3 節では，高次元小標本データの主成分分析 (PCA) を紹介する．高次元小標本に対して従来型の PCA を用いると，不適解を生じることが知られている．解決策として，高次元小標本データの幾何学的表現に基づいて，‘クロスデータ行列法’ と ‘ノイズ掃き出し法’ という 2 つの新しい PCA を紹介する．応用例の一つとして，高次元小標本データのクラスター分析を紹介する．

4 節では，高次元小標本における推定と検定を紹介する．要求されるバンド幅をもつ信頼領域問題と，要求される有意水準と検出力をもつ二標本問題を扱い，高次元小標本ならではの新しい統計的推測を紹介する．高次元小標本データの幾何学的表現に根ざした各種統計量について，高次元漸近正規性を有することを示し，精度を保証するために必要となる標本数を二段階推定法で決定する．

5 節では，高次元小標本における各種パラメータの推定と検定に，計算コストを著しく削減し，かつ，漸近分散が小さい不偏推定量を与えるための ‘拡張クロスデータ行列法’ を紹介する．

6 節では，高次元小標本データの判別分析を紹介する．2 群から抽出される高次元小標本データの

幾何学的差異を利用した判別関数を紹介し，それが高次元漸近正規性を有し，判別方式が誤判別確率に対して要求される精度を保証することを示す．

なお，本稿で扱う理論と方法論は，母集団分布に対する正規性や母集団間の共分散行列に対する共通性の制約を特に必要とするものではない．それらの制約は高次元小標本データの様相を正しく反映しておらず，それらに限定された理論と方法論を扱うことは，高次元データが本来もつ豊富な情報を見落すことになる．本稿では，高次元小標本データが織り成す幾何学的表現に根ざして，理論と方法論を扱うことにする．

2　高次元小標本データの幾何学的表現

平均に p 次の $\mathbf{0}$ ベクトル，共分散行列に p 次の正定値対称行列 $\Sigma_{(p)}(>\boldsymbol{O})$ をもつ母集団を考える．n 個の p 次データベクトル $\boldsymbol{x}_{1(p)},\ldots,\boldsymbol{x}_{n(p)}$ を無作為に抽出して，データ行列 $\boldsymbol{X}_{(p)}: p\times n=[\boldsymbol{x}_{1(p)},\ldots,\boldsymbol{x}_{n(p)}]$ を定義する．ただし，$p>n$ である．$\Sigma_{(p)}$ の固有値を $\lambda_{1(p)}\geq\cdots\geq\lambda_{p(p)}(>0)$ とし，適当な直交行列 $\boldsymbol{H}_{(p)}=[\boldsymbol{h}_{1(p)},\ldots,\boldsymbol{h}_{p(p)}]$ で $\Sigma_{(p)}$ を $\Sigma_{(p)}=\boldsymbol{H}_{(p)}\boldsymbol{\Lambda}_{(p)}\boldsymbol{H}_{(p)}^T$，$\boldsymbol{\Lambda}_{(p)}=\mathrm{diag}(\lambda_{1(p)},\ldots,\lambda_{p(p)})$ と分解する．そのとき $\boldsymbol{Z}_{(p)}=\boldsymbol{\Lambda}_{(p)}^{-1/2}\boldsymbol{H}_{(p)}^T\boldsymbol{X}_{(p)}$ とおき，$\boldsymbol{Z}_{(p)}=[\boldsymbol{z}_{1(p)},\ldots,\boldsymbol{z}_{p(p)}]^T$，$\boldsymbol{z}_{i(p)}=(z_{i1(p)},\ldots,z_{in(p)})^T$ と表記する．ただし，$\boldsymbol{Z}_{(p)}$ の成分は，4 次モーメントが一様有界であることを仮定する．今後，次元数を意識して付した添え字 (p) は省いて表記する．

高次元小標本データを解析する上で鍵となるのは，データがもつ特有の幾何学的表現である．Ahn *et al.* [1], Hall *et al.* [31] は，母集団に正規分布もしくは ρ-mixing を仮定して，高次元小標本データの幾何学的表現を導いた．また，Jung and Marron [36], Jung *et al.* [37] は，母集団に正規分布もしくは ρ-mixing を仮定して，PCA に関する幾何学的表現を導いた．ρ-mixing の詳細については，Bradley [20] を参照のこと．これらに対して，Yata and Aoshima [56] は，母集団に正規分布や ρ-mixing を仮定せずに，高次元漸近理論を発展させることで，高次元小標本データが有する幾何学的表現を発見することに成功した．

いま，標本共分散行列を $\boldsymbol{S}=n^{-1}\boldsymbol{X}\boldsymbol{X}^T$ とする．そのとき，$\boldsymbol{S}_D=n^{-1}\boldsymbol{X}^T\boldsymbol{X}$ は $\boldsymbol{S}$ と正の固有値を共有する．$\boldsymbol{S}_D$ の固有値を $\hat{\lambda}_1\geq\cdots\geq\hat{\lambda}_n$ 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 $\boldsymbol{S}_D=\sum_{j=1}^n\hat{\lambda}_j\hat{\boldsymbol{u}}_j\hat{\boldsymbol{u}}_j^T$ とおく．次のような球形条件を考える．

$$\frac{\sum_{i=1}^p\lambda_i^2}{(\sum_{i=1}^p\lambda_i)^2}\to 0,\quad p\to\infty. \tag{1}$$

そのとき，Ahn *et al.* [1], Jung and Marron [36] は，$\boldsymbol{X}$ に正規分布もしくは $\boldsymbol{Z}$ に ρ-mixing を仮定して

$$\frac{n}{\sum_{i=1}^p\lambda_i}\boldsymbol{S}_D\xrightarrow{P}\boldsymbol{I}_n,\quad p\to\infty \tag{2}$$

を示した．ただし，$\boldsymbol{I}_n$ は n 次単位行列である．Yata and Aoshima [56] は，$\boldsymbol{w}_j=(n/\sum_{i=1}^p\lambda_i)\boldsymbol{S}_D\hat{\boldsymbol{u}}_j$，すなわち，$\boldsymbol{w}_j=(n/\sum_{i=1}^p\lambda_i)\hat{\lambda}_j\hat{\boldsymbol{u}}_j$ について，$\boldsymbol{X}$ が正規分布もしくは $\boldsymbol{Z}$ が ρ-mixing の場合に，$p\to\infty$ のとき次の結果を示した．

$$\boldsymbol{w}_j\in\boldsymbol{R}_n,\quad j=1,\ldots,n. \tag{3}$$

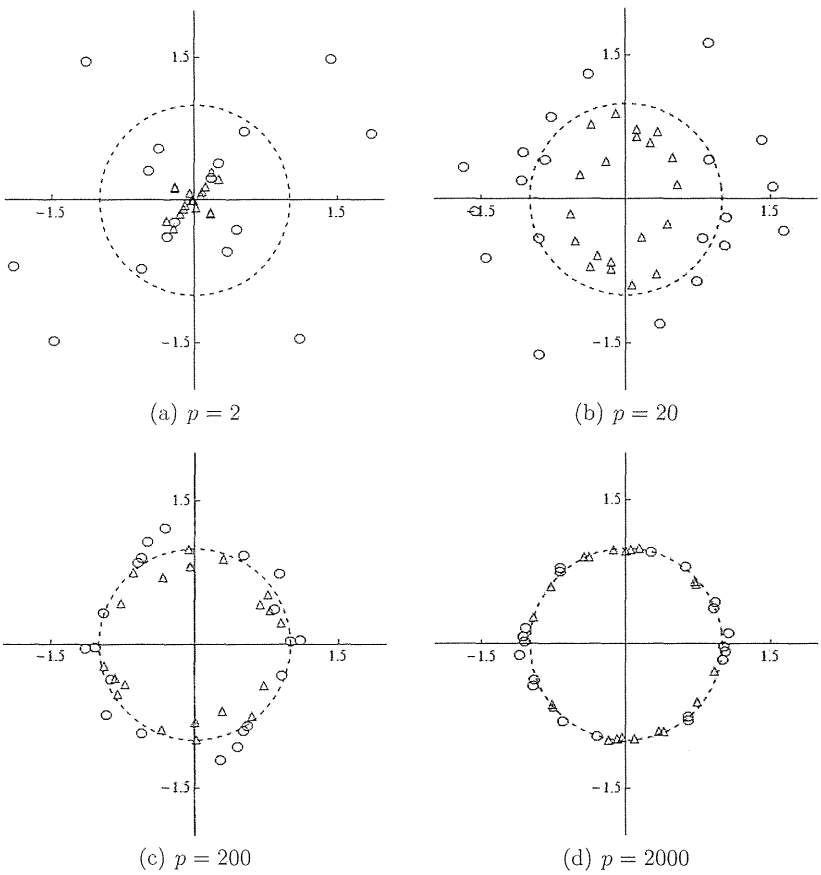
ここで，$\boldsymbol{R}_n = \{\boldsymbol{e}_n \in \boldsymbol{R}^n | \ ||\boldsymbol{e}_n|| = 1\}$ である．(2)，(3) は高次元小標本データがもつ一つの幾何学的表現であり，$p \to \infty$ のとき $\boldsymbol{S}_D$ の固有ベクトルは方向が定まらず，固有値は定まるものの互いの差異がなくなり，$\boldsymbol{S}_D$ を用いて固有値・固有ベクトルを推定することは困難になることが容易に想像できる．また，Yata and Aoshima [55], [56] は，非正規分布かつ非 ρ-mixing の場合に，次の結果を示した．

$$\frac{n}{\sum_{i=1}^{p} \lambda_i} \boldsymbol{S}_D \xrightarrow{P} \boldsymbol{D}_n, \quad p \to \infty. \tag{4}$$

ここで，$\boldsymbol{D}_n$ は対角成分が $O_P(1)$ となる対角行列である．Yata and Aoshima [56] は，非正規分布かつ非 ρ-mixing の場合に，$p \to \infty$ のとき次の結果も示している．

$$\hat{\boldsymbol{u}}_j \in \boldsymbol{R}_{n*}, \quad j = 1, \ldots, n. \tag{5}$$

ここで，$\boldsymbol{R}_{n*} = \{(1, 0, \ldots, 0)^T, (0, 1, \ldots, 0)^T, \ldots, (0, 0, \ldots, 1)^T\}$ である．(4)，(5) は (2)，(3) の対極にある高次元小標本データの幾何学的表現といえ，$\boldsymbol{S}_D$ の固有ベクトルが座標軸と重なり方向は定まるものの，固有値は定まらないことを意味している．以上の観察を精密に述べれば次のようになる．



(a) $p = 2$ (b) $p = 20$

(c) $p = 200$ (d) $p = 2000$

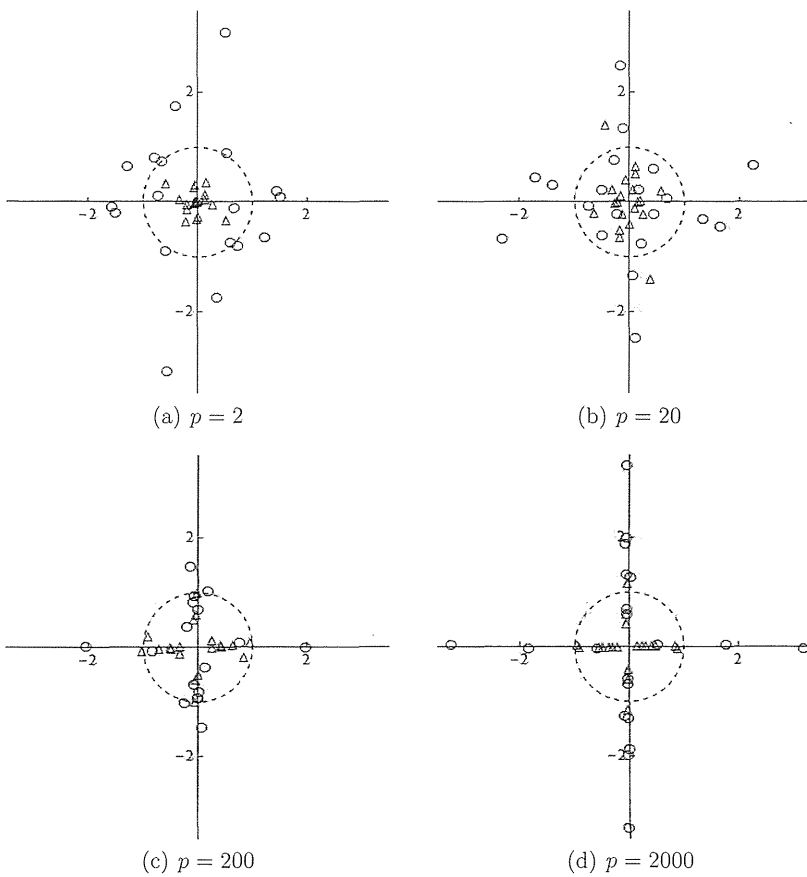
図 1：p 次元正規分布 $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ における 10 組の $\pm \boldsymbol{w}_j (j = 1, 2)$ の幾何学的表現．
○ は $\boldsymbol{w}_1$ を，△ は $\boldsymbol{w}_2$ を表す．

定理 1 ([56]) Z の成分の 2 次モーメントについて $z_{k*}=(z_{1k}^2-1,\dots,z_{pk}^2-1)^T$, $k=1,\dots,n$ とおき，その共分散行列を $\mathrm{Cov}(z_{k*})=(\phi_{ij})$ とする．もしも，Z が

$$\frac{\sum_{i,j}^{p}\lambda_i\lambda_j\phi_{ij}}{(\sum_{j=1}^{p}\lambda_j)^2}\to 0,\quad p\to\infty \tag{6}$$

なる条件を満たすならば，仮定 (1) のもと $p\to\infty$ (n は固定したままでよい) で (2) と (3) が成り立つ．もしも，Z が条件 (6) を満たさないならば，仮定 (1) のもと $p\to\infty$ (n は固定したままでよい) で (4) と (ある正則条件のもと) (5) が成り立つ．

条件 (6) を満たす例と満たさない例として，p 次元正規分布 $N_p(\mathbf{0},\boldsymbol{I}_p)$ と，平均が $\mathbf{0}$ で共分散行列が $\boldsymbol{I}_p$ (尺度行列が $(3/5)\boldsymbol{I}_p$) で自由度が 5 の p 次元 t 分布を考える．ここで，当該の p 次元 t 分布の密度関数は，C を正規化定数とすると，$C\{1+(\boldsymbol{x}^T\boldsymbol{x})/3\}^{-(p+5)/2}$, $\boldsymbol{x}\in\boldsymbol{R}^p$ で与えられる．標本数を $n=2$ として，次元数が $p=2$, 20, 200, 2000 の各場合で，$\pm\boldsymbol{w}_j$ $(j=1,2)$ の対を独立に 10 組発生させた．出力結果を $\boldsymbol{w}_1$ は ○ で，$\boldsymbol{w}_2$ は △ で表示する．図 1 (a)–(d) は正規分布について，図 2 の (a)–(d) は t 分布について，結果を纏めたものである．



(a) $p=2$　(b) $p=20$

(c) $p=200$　(d) $p=2000$

図 2：平均 $\mathbf{0}$，共分散行列 $\boldsymbol{I}_p$ (尺度行列が $(3/5)\boldsymbol{I}_p$)，自由度 5 の p 次元 t 分布における 10 組の $\pm\boldsymbol{w}_j(j=1,2)$ の幾何学的表現．○ は $\boldsymbol{w}_1$ を，△ は $\boldsymbol{w}_2$ を表す．

次元数が増えるにつれ，条件 (6) を満たす場合とそうでない場合の，それぞれのデータの特徴が明瞭になっていく様子が見てとれる．高次元小標本データに対して理論と方法論を構築するためには，データに関するこれらの幾何学的表現を知っておくことが重要である．

3 高次元小標本データの主成分分析

2節と同様に，平均が p 次の $\mathbf{0}$ ベクトル，共分散行列が p 次の正定値対称行列 Σ $(>\boldsymbol{O})$ で，$\boldsymbol{Z}$ の成分の4次モーメントが一様有界であることを仮定する．Σ の固有値には，次のモデルを仮定する．

$$\lambda_i = a_i p^{\alpha_i} \ (i=1,\dots,m), \quad \lambda_i = c_i \ (i=m+1,\dots,p). \tag{7}$$

ここで，a_i (>0)，c_i (>0)，α_i $(\alpha_1 \geq \cdots \geq \alpha_m > 0)$ はともに未知の (p に依存しない固定された) 実数で，m は未知の (p に依存しない固定された) 自然数とし，これらのパラメータは $\lambda_1 \geq \cdots \geq \lambda_p > 0$ を満たすものとする．(7) の後半 (λ_i の $m+1$ 番目以降) はノイズを表し，$\sum_{i=m+1}^{p} \lambda_i^2 / (\sum_{i=m+1}^{p} \lambda_i)^2 \to 0$，$p \to \infty$ となるので，(1) の球形条件を満たすものになっている．Jung and Marron [36] 等も同様のモデルを考えているが，彼らのモデルには $\alpha_i > 1$ という制約がある．標本共分散行列を $\boldsymbol{S} = n^{-1}\boldsymbol{X}\boldsymbol{X}^T$ とする．本節では，$\boldsymbol{Z}$ の成分について，条件

(*) z_{ij}, $i=1,\dots,p$ $(j=1,\dots,n)$ が互いに独立

を満たすときと満たさないときに場合分けして，高次元小標本データに対する主成分分析 (PCA) の性質を纏める．(*) を満たす一つの例は，$\boldsymbol{X}$ が正規分布に従う場合である．本節では，$n(p)$ を p に依存して定まる標本数とする．

3.1 従来型 PCA の限界

Jung and Marron [36]，Yata and Aoshima [53] は，高次元小標本データに対する従来型 PCA の性質を研究した．Lee *et al.* [38] も従来型 PCA の性質をランダム行列理論に基づいて研究しているが，$n/p \to c > 0$ のもとであることに注意する．Jung and Marron [36] は母集団に正規分布もしくは ρ-mixing を仮定して，従来型 PCA に不適解が生じることを示した．一方で，Yata and Aoshima [53] は，母集団分布に制約を課すことなしに，従来型 PCA による推定が一致性をもつための条件を $n(p) = p^{\gamma}$ $(\gamma > 0)$ という形式で導き，高次元小標本における従来型 PCA の限界を示した．例えば，固有値については，次の2つの定理を与えた．

定理 2 ([53]) $\boldsymbol{Z}$ が (*) を満たすとき，条件

(i) $\alpha_i > 1$ のとき $p \to \infty$, $n \to \infty$,

(ii) $\alpha_i \in (0,1]$ のとき $p \to \infty$, $p^{1-\alpha_i}/n(p) \to 0$

のもとで，$\hat{\lambda}_i$ $(i=1,\dots,m)$ について次が成り立つ．

$$\frac{\hat{\lambda}_i}{\lambda_i} = 1 + o_P(1). \tag{8}$$

定理 3 ([53]) $\boldsymbol{Z}$ が (*) を満たさないとき，(8) は次の条件のもとで成り立つ．

(i) $\alpha_i > 1$ のとき $p \to \infty$, $n \to \infty$,

(ii) $\alpha_i \in (0,1]$ のとき $p \to \infty$, $p^{2-2\alpha_i}/n(p) \to 0$.

これらの定理は，従来型 PCA による固有値の推定が一致性をもつための条件を与える．定理2と3から，$\alpha_i \in (0,1]$ のとき従来型 PCA が一致性をもつためには，標本数 n を次元数 p に依存して決め

るべきだと分かる．また，$\boldsymbol{Z}$ が (*) を満たさない場合，(ii) の条件が厳しくなっているため，高次元小標本のもとで推定が一致性をもつことは極めて難しいことが分かる．(次節の注意 1 を参照のこと．) この困難を克服するために，Yata and Aoshima [54], [56] は 'クロスデータ行列法' と 'ノイズ掃き出し法' という 2 つの異なるアプローチを提案し，これらの方法論に基づく新しい PCA を開発した．

3.2 ノイズ掃き出し法による新しい PCA

Yata and Aoshima [56] は，条件 (*) が仮定される高次元データに対して，図 1 で見た高次元小標本の幾何学的表現に着目して，ノイズ掃き出し法という方法論を提案した．それは，次のような固有値の推定に基づくものである．

$$\acute{\lambda}_i = \hat{\lambda}_i - \frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{j=1}^{i} \hat{\lambda}_j}{n-i} \quad (i = 1, \ldots, n-1). \tag{9}$$

(9) の第 2 項がノイズの掃き出しを意味している．$\boldsymbol{S}_D$ は

$$n\boldsymbol{S}_D = \sum_{j=1}^{m} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^T + \sum_{j=m+1}^{p} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^T$$

と分解でき，ノイズを表す分解の第 2 項は条件 (*) のもとで

$$\frac{\sum_{i,j=m+1}^{p} \lambda_i \lambda_j \phi_{ij}}{(\sum_{j=m+1}^{p} \lambda_j)^2} = O\Big(\frac{\sum_{j=m+1}^{p} \lambda_j^2}{(\sum_{j=m+1}^{p} \lambda_j)^2}\Big) \to 0, \quad p \to \infty$$

を満たす．よって，定理 1 により

$$\frac{\sum_{j=m+1}^{p} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^T}{\sum_{j=m+1}^{p} \lambda_j} \xrightarrow{P} \boldsymbol{I}_n, \quad p \to \infty$$

なる幾何学的表現をもつ．従って，ノイズの大きさが確率的に定まり，それを (9) の第 2 項のように除去すれば固有値の望ましい推定ができる．ノイズの漸近的挙動を精密に評価すると，次の定理を得る．

定理 4 ([56]) $\boldsymbol{Z}$ に (*) を仮定する．そのとき，条件

(i) $\alpha_i > 1/2$ のとき $p \to \infty$, $n \to \infty$,

(ii) $\alpha_i \in (0, 1/2]$ のとき $p \to \infty$, $p^{1-2\alpha_i}/n(p) \to 0$

のもとで，$\acute{\lambda}_i$ $(i = 1, \ldots, m)$ について次が成り立つ．

$$\frac{\acute{\lambda}_i}{\lambda_i} = 1 + o_P(1). \tag{10}$$

定理 4 の条件を定理 2 と見比べると，推定量 (9) は従来型 PCA よりも緩い条件のもとで一致性をもつことが分かる．

注意 1 条件 (*) のもとで，$\boldsymbol{S}_D$ のノイズの大きさは漸近的に $\sum_{j=m+1}^{p} \lambda_j/n$ である．定理 2 の (ii) は，ノイズが λ_i $(i \le m)$ よりも小さくなるように標本数 n をとることができれば $\hat{\lambda}_i$ は一致性をもつ，というものである．一方，定理 4 の (ii) では，ノイズの主要項を掃き出してから $\acute{\lambda}_i$ を定義しているので，一致性をもつための標本数 n にかかる条件は軽減されている．なお，条件 (*) を満たさない場合，図 2 のような散らばったノイズも λ_i $(i \le m)$ より小さくなるように標本数 n をとらなけ

ればならないので，一致性をもつための条件は必然的に厳しくなる．

定理 5 ([56]) $\boldsymbol{Z}$ に (∗) を仮定する．$\mathrm{Var}(z_{ij}^2) = M_i\ (>0)$, $i = 1, \ldots, p$ $(j = 1, \ldots, n)$ とおく．そのとき，条件

(i) $\alpha_i > 1/2$ のとき $p \to \infty$, $n \to \infty$,

(ii) $\alpha_i \in (0, 1/2]$ のとき $p \to \infty$, $p^{2-4\alpha_i}/n(p) \to 0$

のもとで，単根の固有値 λ_i $(i \leq m)$ に対して次が成り立つ．

$$\sqrt{\frac{n}{M_i}}\left(\frac{\acute{\lambda}_i}{\lambda_i} - 1\right) \Rightarrow N(0,1). \tag{11}$$

ここで，$\Rightarrow$ は分布収束を表す．

注意 2 $\boldsymbol{Z}$ が正規分布のとき，$M_i = 2$ $(i = 1, \ldots, p)$ である．なお，重根の固有値に対しては，対応する Σ の固有ベクトルが一意に定まらないので，(11) のような漸近正規性は成り立たない．

次に，Σ の固有ベクトルについて，ノイズ掃き出し法による推定を考える．$\boldsymbol{S}_D$ のスペクトル分解を $\boldsymbol{S}_D = \sum_{i=1}^n \hat{\lambda}_i \hat{\boldsymbol{u}}_i \hat{\boldsymbol{u}}_i^T$ とする．推定量 (9) に基づいて，Σ の固有ベクトル $\boldsymbol{h}_i$ を $\acute{\boldsymbol{h}}_i = (n\acute{\lambda}_i)^{-1/2}\boldsymbol{X}\hat{\boldsymbol{u}}_i$ で推定する．ここで，$\boldsymbol{h}_i$ には符号の自由度があるため，各 i で $\acute{\boldsymbol{h}}_i^T \boldsymbol{h}_i \geq 0$ を仮定する．そのとき，次の定理を得る．

定理 6 ([56]) $\boldsymbol{Z}$ に (∗) を仮定する．そのとき，定理 4 の条件 (i)–(ii) のもとで，単根の固有値 λ_i $(i \leq m)$ に対して次が成り立つ．

$$\acute{\boldsymbol{h}}_i^T \boldsymbol{h}_i = 1 + o_P(1). \tag{12}$$

次に，主成分スコアについて，ノイズ掃き出し法による推定を考える．データ $\boldsymbol{x}_j$ の第 i 主成分スコアは，$\boldsymbol{h}_i^T \boldsymbol{x}_j = \sqrt{\lambda_i} z_{ij}$ $(= s_{ij}$ とおく$)$ である．$\boldsymbol{S}_D$ の固有ベクトルの成分を $\hat{\boldsymbol{u}}_i = (\hat{u}_{i1}, \ldots, \hat{u}_{in})^T$ とする．推定量 (9) に基づいて，第 i 主成分スコアを $\hat{u}_{ij}\sqrt{n\acute{\lambda}_i}$ $(= \acute{s}_{ij}$ とおく$)$ で推定する．そのとき，次の定理を得る．

定理 7 ([56]) $\boldsymbol{Z}$ に (∗) を仮定する．$\mathrm{MSE}(\acute{s}_i) = n^{-1}\sum_{j=1}^n (\acute{s}_{ij} - s_{ij})^2$ とおく．そのとき，定理 4 の条件 (i)–(ii) のもとで，単根の固有値 $\lambda_i\,(i \leq m)$ に対して次が成り立つ．

$$\frac{\mathrm{MSE}(\acute{s}_i)}{\lambda_i} = o_P(1). \tag{13}$$

系 1 ([56]) 平均が 0 でないとき，$\boldsymbol{S}_{oD} = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^T(\boldsymbol{X} - \overline{\boldsymbol{X}})$ とおく．ここで，$\overline{\boldsymbol{X}} = [\bar{\boldsymbol{x}}, \ldots, \bar{\boldsymbol{x}}]$, $\bar{\boldsymbol{x}} = \sum_{j=1}^n \boldsymbol{x}_j/n$ である．そのとき，$(\boldsymbol{S}_D,\ n)$ の代わりに $(\boldsymbol{S}_{oD},\ n-1)$ を用いれば，定理 4 から定理 7 が成り立つ．

条件 (∗) が仮定されるとき，ノイズ掃き出し法による固有値，固有ベクトルと主成分スコアの推定は，従来型 PCA を著しく改良することが報告されている．ノイズ掃き出し法の応用として，Yata and Aoshima [56] は高次元小標本における Σ の逆行列の推定を与えている．高次元における Σ の推測については，Bickel and Levina [17], [18] 等を参照のこと．

3.3 クロスデータ行列法による新しい PCA

本節では，$\boldsymbol{Z}$ に (∗) が仮定できない場合を考える．このとき，図 2 のように固有空間が座標軸上に集まる場合も考慮しなければならない．もはや，ノイズの大きさが定まらず，(9) のようなノイズ除去

をすることはできない．このような状況にあっても有効な方法論として，Yata and Aoshima [54] はクロスデータ行列法を考案した．クロスデータ行列法はノンパラメトリックな方法論であり，次のようなアイデアに基づくものである：いま，$n_{(1)} = \langle n/2 \rangle + 1$，$n_{(2)} = n - n_{(1)}$ とおく．ここで，$\langle x \rangle$ は x を越えない最大の整数を表す．データ行列 $\boldsymbol{X}$ を（無作為に）2つに分割して，$\boldsymbol{X}_l : p \times n_{(l)} = [\boldsymbol{x}_{l1}, \ldots, \boldsymbol{x}_{ln_{(l)}}]$ $(l = 1, 2)$ を定義する．そのとき，$\boldsymbol{S}_{D(1)} = (n_{(1)}n_{(2)})^{-1/2}\boldsymbol{X}_1^T\boldsymbol{X}_2$ をクロスデータ行列とよぶことにする．いま，$n_{(1)} \geq n_{(2)}$ に注意し，$\boldsymbol{S}_{D(1)}$ の特異値分解を $\boldsymbol{S}_{D(1)} = \sum_{i=1}^{n_{(2)}} \tilde{\lambda}_i \tilde{\boldsymbol{u}}_{i(1)} \tilde{\boldsymbol{u}}_{i(2)}^T$ とする．ここで，$\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_{n_{(2)}}$ (≥ 0) は $\boldsymbol{S}_{D(1)}$ の特異値，$\tilde{\boldsymbol{u}}_{i(1)}$（もしくは $\tilde{\boldsymbol{u}}_{i(2)}$）は左特異ベクトル（もしくは右特異ベクトル）である．クロスデータ行列 $\boldsymbol{S}_{D(1)}$ の特異値分解に着目する理由は，高次元小標本データの幾何学的表現にある．いま，$\boldsymbol{X}$ の分割に対応して $\boldsymbol{Z}$ を2つに分割し，$\boldsymbol{z}_{1j} = (z_{1j1}, \ldots, z_{1jn_{(1)}})^T$，$\boldsymbol{z}_{2j} = (z_{2j1}, \ldots, z_{2jn_{(2)}})^T$，$j = 1, \ldots, p$ を定義する．そのとき，

$$(n_{(1)}n_{(2)})^{1/2}\boldsymbol{S}_{D(1)} = \sum_{j=1}^{m} \lambda_j \boldsymbol{z}_{1j}\boldsymbol{z}_{2j}^T + \sum_{j=m+1}^{p} \lambda_j \boldsymbol{z}_{1j}\boldsymbol{z}_{2j}^T$$

と書ける．第2項はノイズを表し，無条件で

$$\frac{\sum_{j=m+1}^{p} \lambda_j \boldsymbol{z}_{1j}\boldsymbol{z}_{2j}^T}{\sum_{j=m+1}^{p} \lambda_j} \xrightarrow{P} \boldsymbol{O}, \quad p \to \infty$$

なる幾何学的表現を有する．これは，クロスデータ行列を用いることで，ノイズを自動的に除去することができることを意味している．ノイズの漸近的挙動を精密に評価すると，次の定理を得る．

定理 8 ([54]) 条件

(i) $\alpha_i > 1/2$ のとき $p \to \infty$, $n \to \infty$,

(ii) $\alpha_i \in (0, 1/2]$ のとき $p \to \infty$, $p^{2-2\alpha_i}/n(p) \to 0$

のもとで，$\tilde{\lambda}_i$ $(i = 1, \ldots, m)$ について次が成り立つ．

$$\frac{\tilde{\lambda}_i}{\lambda_i} = 1 + o_P(1). \tag{14}$$

定理8の条件を定理3と見比べると，特異値 $\tilde{\lambda}_i$ は従来型PCAよりも緩い条件のもとで一致性をもつことが分かる．

注意 3　この性質を利用して，Yata and Aoshima [55] は，高次元データ空間に内在する潜在空間の次元推定を考えている．

定理 9 ([54]) $\mathrm{Var}(z_{ij}^2) = M_i$ (> 0), $i = 1, \ldots, p$ $(j = 1, \ldots, n)$ とおく．定理8の条件 (i)–(ii) のもとで，単根の固有値 λ_i $(i \leq m)$ に対して次が成り立つ．

$$\sqrt{\frac{n}{M_i}} \left(\frac{\tilde{\lambda}_i}{\lambda_i} - 1 \right) \Rightarrow N(0, 1). \tag{15}$$

次に，$\boldsymbol{\Sigma}$ の固有ベクトルについて，クロスデータ行列法による推定を考える．$\boldsymbol{S}_{D(1)}$ の特異値 $\tilde{\lambda}_i$ と特異ベクトル $\tilde{\boldsymbol{u}}_{i(j)}$, $j = 1, 2$ に基づいて，固有ベクトルを $\tilde{\boldsymbol{h}}_i = \tilde{\lambda}_i^{-1/2}(n_{(1)}^{-1/2}\boldsymbol{X}_1\tilde{\boldsymbol{u}}_{i(1)} + n_{(2)}^{-1/2}\boldsymbol{X}_2\tilde{\boldsymbol{u}}_{i(2)})/2$ で推定する．各 i で $\tilde{\boldsymbol{h}}_i^T\boldsymbol{h}_i \geq 0$ を仮定する．そのとき，次の定理を得る．

定理 10 ([54]) 定理8の条件 (i)–(ii) のもとで，単根の固有値 λ_i $(i \leq m)$ に対して次が成り立つ．

$$\tilde{\boldsymbol{h}}_i^T \boldsymbol{h}_i = 1 + o_P(1). \tag{16}$$

次に，主成分スコアについて，クロスデータ行列法による推定を考える．$\boldsymbol{S}_{D(1)}$ の特異ベクトルを $\tilde{\boldsymbol{u}}_{i(l)} = (\tilde{u}_{i1(l)}, \ldots, \tilde{u}_{in_{(l)}(l)})^T$ $(l = 1, 2)$ と成分表示する．$\boldsymbol{S}_{D(1)}$ の特異値と特異ベクトルに基づいて，$\boldsymbol{x}_{lj}$ $(l = 1, 2)$ の第 i 主成分スコアを $\tilde{u}_{ij(l)}\sqrt{n_{(l)}\tilde{\lambda}_i}$ $(= \tilde{s}_{ij(l)}$ とおく$)$ で推定する．各 i で，$\tilde{s}_{ij(l)}$, $j = 1, \ldots, n_{(l)}$, $l = 1, 2$ に $\tilde{s}_{ij}$, $j = 1, \ldots, n$ という通し番号を付ける．そのとき，次の定理を得る．

定理 11 ([54]) $\mathrm{MSE}(\tilde{s}_i) = n^{-1}\sum_{j=1}^{n}(\tilde{s}_{ij} - s_{ij})^2$ とおく．定理 8 の条件 (i)–(ii) のもとで，単根の固有値 λ_i $(i \leq m)$ に対して次が成り立つ．

$$\frac{\mathrm{MSE}(\tilde{s}_i)}{\lambda_i} = o_P(1). \tag{17}$$

系 2 ([54]) $\boldsymbol{Z}$ に $(*)$ を仮定する．そのとき，定理 4 の条件 (i)–(ii) のもとで，(14) が成り立ち，単根の固有値 λ_i $(i \leq m)$ に対して (16)–(17) が成り立つ．定理 5 の条件 (i)–(ii) のもとで (15) が成り立つ．

系 3 ([54]) 平均が 0 でないとき，$\boldsymbol{S}_{oD(1)} = (n_{(1)}n_{(2)})^{-1/2}(\boldsymbol{X}_1 - \overline{\boldsymbol{X}}_1)^T(\boldsymbol{X}_2 - \overline{\boldsymbol{X}}_2)$ とおく．ここで，$\overline{\boldsymbol{X}}_l = [\bar{\boldsymbol{x}}_l, \ldots, \bar{\boldsymbol{x}}_l]$, $\bar{\boldsymbol{x}}_l = n_{(l)}^{-1}\sum_{j=1}^{n_{(l)}}\boldsymbol{x}_{lj}$ $(l = 1, 2)$ である．そのとき，$\boldsymbol{S}_{D(1)}$ の代わりに $\boldsymbol{S}_{oD(1)}$ を用いれば，定理 8 から定理 11 および系 2 が成り立つ．

クロスデータ行列法による固有値，固有ベクトルと主成分スコアの推定は，条件 $(*)$ が仮定できないノンパラメトリックな状況において，従来型の PCA を著しく改良することが報告されている．また，ノイズ掃き出し法とクロスデータ行列法の比較が，理論的かつ数値的に Yata [51] と Yata and Aoshima [59] で研究されている．

3.4 高次元小標本データのクラスター分析

高次元小標本における PCA の一つの応用として，クラスター分析を考える．2 つのクラスターがあるとする．それらを π_1, π_2 と名付け，平均 $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ と，共分散行列 Σ_1, Σ_2 をもつと仮定する．データは，p.d.f. (もしくは，p.f.)

$$f(\boldsymbol{x}) = \varepsilon f_1(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1) + (1 - \varepsilon) f_2(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2),\ \varepsilon \in (0, 1)$$

をもつ混合分布からの標本とみなす．ただし，$f_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i)$ 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 $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$ を定義する．そのとき，$E(\boldsymbol{x}_i) = \varepsilon\boldsymbol{\mu}_1 + (1 - \varepsilon)\boldsymbol{\mu}_2$, $\mathrm{Var}(\boldsymbol{x}_i) = \varepsilon\Sigma_1 + (1 - \varepsilon)\Sigma_2 + \varepsilon(1 - \varepsilon)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$ $(= \Sigma$ とおく$)$ である．Σ_i の最大固有値を λ_{i1} とし，$\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ とする．ここで，$\Delta = cp^{\beta}$, $c > 0$, $\beta > 0$, かつ，$\lambda_{i1}/\Delta \to 0$, $i = 1, 2$ を仮定し，$\boldsymbol{h}_1^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0$ とする．Yata and Aoshima [54] は，第 1 主成分スコア s_{1j} と Σ の最大固有値 λ_1 について $p \to \infty$ のとき

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

となることを示した．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小標本データを高い精度で分類できる．混合分布は非正規分布なので，3.3 節で紹介したクロスデータ行列

法を用いて主成分スコアを推定する．実際，Aoshima and Yata [10]，Yata and Aoshima [54] は，推定量 $\tilde{s}_{ij}$ を用いてマイクロアレイデータのクラスター分析を行い，性能の高さを確認している．

4 高次元小標本の平均ベクトルに関する推定と検定

本節では，母集団が k 個あると想定し，各母集団 (π_i) の分布には平均ベクトル $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ip})^T$ と共分散行列 Σ_i $(> \boldsymbol{O})$ の存在を仮定する．高次元データに対して $\Sigma_i = \Sigma_j$ $(i \neq j)$ を想定することは現実的ではないので，共分散行列の共通性は仮定しない．むしろ，高次元小標本データの幾何学的表現を通して，母集団間の共分散行列の差異を推測に生かすことを考える．この点については，6 節でも触れることにする．Σ_i の固有値を $\lambda_{i1} \geq \cdots \geq \lambda_{ip} (> 0)$ とし，$\liminf_{p\to\infty} \lambda_{ip} > 0$ $(i = 1, \ldots, k)$ を仮定する．Σ_i を $\Sigma_i = \boldsymbol{H}_i \Lambda_i \boldsymbol{H}_i^T$，$\Lambda_i = \mathrm{diag}(\lambda_{i1}, \ldots, \lambda_{ip})$ と対角化する適当な直交行列を $\boldsymbol{H}_i = [\boldsymbol{h}_{i1}, \ldots, \boldsymbol{h}_{ip}]$ とする．いま，各母集団 π_i から $n_i (\geq 4)$ 個の p 次元データ $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$ を無作為に抽出する．$\boldsymbol{z}_{ij} = (z_{i1j}, \ldots, z_{ipj})^T = \Lambda_i^{-1/2} \boldsymbol{H}_i^T (\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i)$ とおく．データには次のモデルを仮定する．

$$\boldsymbol{x}_{ij} = \Gamma_i \boldsymbol{w}_{ij} + \boldsymbol{\mu}_i, \quad i = 1, \ldots, k;\ j = 1, \ldots, n_i. \tag{18}$$

ここで，Γ_i は $\Gamma_i \Gamma_i^T = \Sigma_i$ なる $p \times r_i$ 行列とし，$\boldsymbol{w}_{ij}$ は $E(\boldsymbol{w}_{ij}) = \boldsymbol{0}$，$\mathrm{Var}(\boldsymbol{w}_{ij}) = \boldsymbol{I}_{r_i}$ とする．ただし，$r_i \geq p$ である．このモデルは Bai and Saranadasa [12], Chen and Qin [22], Yata and Aoshima [58] 等で扱われた．(18) は，$\Gamma_i = \boldsymbol{H}_i \Lambda_i^{1/2}$ の場合を含む一般的な多変量モデルといえる．ここで，$\boldsymbol{w}_{ij} = (w_{i1j}, \ldots, w_{ir_ij})^T$ とおき，各成分は 4 次モーメントが一様有界であることを仮定する．母集団 π_i，$i = 1, \ldots, k$ の分布について，次の 3 つのどれか一つを仮定する：

(A-i) $N_p(\boldsymbol{\mu}_i, \Sigma_i)$,

(A-ii) z_{ijl}，$j = 1, \ldots, p$ は互いに独立である，

(A-iii) $E(w_{ijl} w_{isl} w_{itl}) = 0$, $E(w_{ijl}^2 w_{isl}^2) = 1$, $E(w_{ijl} w_{isl} w_{itl} w_{iul}) = 0$, $j \neq s, t, u$.

(A-iii) は (A-ii) を緩めた条件であり，(A-ii) は (A-i) を緩めた条件である．共分散行列には以下を仮定する：

(A-iv) $\mathrm{tr}(\Sigma_i^t)/p < \infty$ $(t = 1, 2)$, $\mathrm{tr}(\Sigma_i^4)/p^2 \to 0$, $p \to \infty$; $i = 1, \ldots, k$.

4.1 要求されるバンド幅をもつ信頼領域

平均ベクトルの 1 次結合 $\boldsymbol{\mu} = \sum_{i=1}^k b_i \boldsymbol{\mu}_i$ の推定を考える．各母集団 π_i から抽出された大きさ n_i の標本に基づいて $\boldsymbol{T}_{\boldsymbol{n}} = \sum_{i=1}^k b_i \overline{\boldsymbol{x}}_{in_i}$ を定義する．ここで，$\boldsymbol{n} = (n_1, \ldots, n_k)$，$\overline{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ である．任意の $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \Sigma_1, \ldots, \Sigma_k)$ に対して要求される精度を満たすための，二段階推定法に基づいた $\boldsymbol{\mu}$ に関する信頼領域は，Aoshima *et al.* [6] によって一致解が与えられ，Aoshima and Takada [5] によって解の 2 次漸近有効性が示された．さらに，Aoshima and Yata [7] は二段階推定法の 2 次漸近一致性の理論を構築し，Yata [49]，Yata and Aoshima [52] によって各種推測問題に拡張された．また，Yata [50] は高次元データに対する $p \to \infty$ 漸近有効な解を考えた．種々の統計的推測における二段階推定法については，例えば，Aoshima [2]，Aoshima *et al.* [3]，Aoshima and Mukhopadhyay [4]，Ghosh *et al.* [30] 等を参照のこと．一連の先行研究は，$n/p \to \infty$ もしくは $n/p \to c$ (> 0) の場合を扱っている．高次元小標本の枠組みでは $n_i/p \to 0$ で信頼領域を構築することになり，その場合，有界な半径をもつ信頼領域は存在しない．Aoshima and Yata [8] は，損

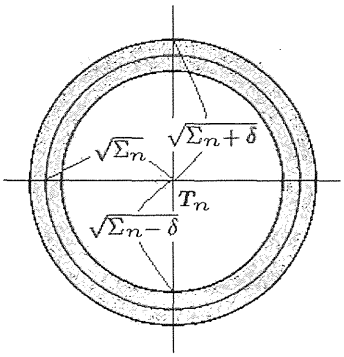


図3：μ の信頼領域 R_{Σ_n}（灰色の領域）

失関数 $||T_n-\mu||^2$ について，与えられる $\delta=o(p^{1/2})>0$ でバンド幅を調節する

$$R_{\Sigma_n}=\{\mu\in R^p:\max\{-\delta+\Sigma_n,\ 0\}\le||T_n-\mu||^2\le\delta+\Sigma_n\} \tag{19}$$

なる信頼領域を考えた．ここで，$\Sigma_n=\sum_{i=1}^k b_i^2\mathrm{tr}(\Sigma_i)/n_i$ である．これは，高次元小標本データの幾何学的表現に基づいて導かれた信頼領域である．データの球面集中現象に着目して μ の存在領域を有効に特定しようというものである．$\Sigma_n>\delta$ のとき，R_{Σ_n} は中心が T_n で半径がそれぞれ $\sqrt{\Sigma_n-\delta}$ と $\sqrt{\Sigma_n+\delta}$ の2つの p 次元球に挟まれる領域になる．図3の灰色の領域が R_{Σ_n} である．

注意4 Aoshima and Yata [8] は，高次元小標本データの判別分析において，この信頼領域の幾何学的表現に基づいて判別関数を考えた．(6節を参照のこと．) また，Yata and Aoshima [57] は，高次元の変数選択にこの信頼領域を応用している．

各母集団 π_i の標本共分散行列 $S_{in_i}=(n_i-1)^{-1}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_{in_i})(x_{ij}-\overline{x}_{in_i})^T$ に基づいて，$\widehat{\Sigma}_n=\sum_{i=1}^k b_i^2\mathrm{tr}(S_{in_i})/n_i$ とおく．このとき，$||T_n-\mu||^2$ について，次の定理を得る．

定理12 ([8]) 母集団 π_i，$i=1,\dots,k$ の分布に (A-ii) を仮定する．(A-iv) のもと，$p\to\infty$，$n_i\to\infty$, $i=1,\dots,k$ のとき次が成り立つ．

$$\frac{||T_n-\mu||^2-\widehat{\Sigma}_n}{\sqrt{2\sum_{i,j}^k b_i^2b_j^2\mathrm{tr}(\Sigma_i\Sigma_j)/(n_in_j)}}\Rightarrow N(0,1).$$

いま，与えられる $\alpha\in(0,1)$ に対して，母集団 π_i $(i=1,\dots,k)$ の標本数を

$$n_i\ge\frac{z_{\alpha/2}\sqrt{2}}{\delta}|b_i|\mathrm{tr}(\Sigma_i^2)^{1/4}\sum_{j=1}^k|b_j|\mathrm{tr}(\Sigma_j^2)^{1/4}\quad(=C_i\text{とおく}) \tag{20}$$

を満たす最小の整数とする．ここで，$z_{\alpha/2}$ は $N(0,1)$ の上側 $\alpha/2$ 点である．(20) は適当な条件のもとで $\sum_{i=1}^k n_i$ が最小になるように導かれている．(A-iv) と $\delta=o(p^{1/2})>0$ から，$C_i/p\to0$, $p\to\infty$ が成り立ち，高次元小標本の枠組みで標本数が定まっていることに注意する．このとき，次の定理を得る．

定理13 ([8]) 母集団 π_i，$i=1,\dots,k$ の分布に (A-ii) を仮定する．n は (20) を満たすものとする．(A-iv) のもと $p\to\infty$ のとき，次が成り立つ．

$$\liminf P_{\boldsymbol{\theta}}(\boldsymbol{\mu} \in \boldsymbol{R}_{\widehat{\Sigma}_n}) \geq 1-\alpha.$$

(20) における $\mathrm{tr}(\Sigma_i^2)$ は未知なので，実際には二段階推定法で標本数を決定する．二段階推定法の初期標本数の決め方については，Mukhopadhyay and Duggan [39], [40] を参照のこと．$\sqrt{\mathrm{tr}(\Sigma_i^2)}$ について，各母集団 π_i から得られる事前情報をもとに既知の下限 $\sigma_{i\star}$ ($\sqrt{\mathrm{tr}(\Sigma_i^2)} > \sigma_{i\star} > 0$) を設定し，$\sigma_{i\star}/\sqrt{\mathrm{tr}(\Sigma_i^2)} \in (0,1)$, $p \to \infty$ を仮定する．いま，$\tau_\star = \min_{1\leq i\leq k} |b_i|\sqrt{\sigma_{i\star}} \sum_{j=1}^k |b_j|\sqrt{\sigma_{j\star}}$ とおいて

$$m = \max\left\{4,\ \left\langle \frac{z_{\alpha/2}\sqrt{2}}{\delta}\tau_\star \right\rangle + 1\right\}$$

を定義する．ここで，$\langle x \rangle$ は x を越えない最大の整数を表す．各母集団 π_i から m 個の初期標本を抽出し，$\mathrm{tr}(\Sigma_i^2)$ の不偏推定量を 3.3 節のクロスデータ行列法，もしくは，5.1 節で紹介する拡張クロスデータ行列法で構築する．例えば，Yata [50] はクロスデータ行列法を用いて $\mathrm{tr}(\Sigma_i^2)$ の不偏推定量 $\mathrm{tr}(\boldsymbol{S}_{im(1)}\boldsymbol{S}_{im(2)})$ を得た．(詳細は 5.1 節を参照のこと．) いま，各母集団の標本数を

$$N_i = \max\left\{m,\ \left\langle \frac{z_{\alpha/2}\sqrt{2}}{\delta}|b_i|\mathrm{tr}(\boldsymbol{S}_{im(1)}\boldsymbol{S}_{im(2)})^{1/4}\sum_{j=1}^k |b_j|\mathrm{tr}(\boldsymbol{S}_{jm(1)}\boldsymbol{S}_{jm(2)})^{1/4}\right\rangle + 1\right\} \quad (21)$$

で定義する．各母集団から $N_i - m$ 個の追加標本を抽出し，初期標本と追加標本を合併して $\boldsymbol{T}_{\boldsymbol{N}} = \sum_{i=1}^k b_i \overline{\boldsymbol{x}}_{iN_i}$ と $\widehat{\Sigma}_{\boldsymbol{N}} = \sum_{i=1}^k b_i^2 \mathrm{tr}(\boldsymbol{S}_{iN_i})/N_i$ を定義する．ここで，$\boldsymbol{N} = (N_1, \ldots, N_k)$ である．このとき，$\boldsymbol{T}_{\boldsymbol{N}}$ と $\widehat{\Sigma}_{\boldsymbol{N}}$ に基づいて計算される (19) の信頼領域について，次の定理を得る．

定理 14 ([8]) 母集団 π_i, $i = 1, \ldots, k$ の分布に (A-ii) を仮定する．(A-iv) のもと $p \to \infty$ のとき，次が成り立つ．

$$\liminf P_{\boldsymbol{\theta}}(\boldsymbol{\mu} \in \boldsymbol{R}_{\widehat{\Sigma}_{\boldsymbol{N}}}) \geq 1-\alpha.$$

定理 15 ([8]) 母集団 π_i, $i = 1, \ldots, k$ の分布に (A-i) を仮定する．(A-iv) のもと $p \to \infty$ のとき，各 π_i で次が成り立つ．

$$\limsup |E_{\boldsymbol{\theta}}(N_i - C_i)| \leq 1, \quad \mathrm{Var}_{\boldsymbol{\theta}}(N_i) = o(p^{1/2}/\delta).$$

注意 5 (21) の N_i について，$\mathrm{tr}(\boldsymbol{S}_{im(1)}\boldsymbol{S}_{im(2)})$ の代わりに 5.1 節で紹介する拡張クロスデータ行列法から導かれる (26) の不偏推定量 W_{im} を用いても，上記の 2 つの定理は成り立つ．

二段階推定法によって構築した信頼領域の精度を，シミュレーション実験で検証する．単純な設定として，$p = 1600$; $k = 2$, $b_1 = b_2 = 1$; $\delta = 5$, $\alpha = 0.05$; $m = 20$ とおく．説明を簡単にするため，π_i $(i = 1, 2)$ の母集団分布には $N_p(\boldsymbol{0}, \Sigma_i)$ を考える．ここで，$\Sigma_i = c_i \boldsymbol{B}(\rho_i^{|i-j|^{1/3}})\boldsymbol{B}$ とおき，$c_i > 0$, $\boldsymbol{B} = \mathrm{diag}(\sqrt{0.5 + 1/(p+1)}, \sqrt{0.5 + 2/(p+1)}, \ldots, \sqrt{0.5 + p/(p+1)})$, $\rho_i \in (0,1)$ とする．次の 3 つの場合を考える：(i) $(c_1, c_2) = (1, 1)$, $(\rho_1, \rho_2) = (0.3, 0.3)$; (ii) $(c_1, c_2) = (1, 1)$, $(\rho_1, \rho_2) = (0.3, 0.4)$; (iii) $(c_1, c_2) = (1, 1.5)$, $(\rho_1, \rho_2) = (0.3, 0.3)$. 2000 回のシミュレーションによる標本数の平均，分散，被覆確率，その標準偏差を纏めたものが表 1 である．ここでは割愛するが，設定を変えて実験をしたときにも，二段階推定法による信頼領域は数千の次元数で要求精度を満たすことが確認されている．

注意 6 Yata and Aoshima [57] は，$||\boldsymbol{\mu}||^2$ の推定量 $\widehat{T}_{\boldsymbol{n}} = ||\boldsymbol{T}_{\boldsymbol{n}}||^2 - \widehat{\Sigma}_{\boldsymbol{n}}$ を考え，p に依存する

表 1：二段階推定法で構築した信頼領域の精度 ($p = 1600$, $k = 2$, $\delta = 5$, $\alpha = 0.05$, $m = 20$)

		$\overline{N}$	$\overline{N} - C$	$\mathrm{Var}(N)$	$\overline{P}$	$s(\overline{P})$
		$(c_1, c_2) = (1, 1)$, $(\rho_1, \rho_2) = (0.3, 0.3)$				
C	116.29	117.00	0.72	47.81	0.943	0.00518
C_1	58.14	58.50	0.36	15.13		
C_2	58.14	58.50	0.36	14.83		
		$(c_1, c_2) = (1, 1)$, $(\rho_1, \rho_2) = (0.3, 0.4)$				
C	131.66	132.24	0.58	69.54	0.950	0.00487
C_1	61.87	62.17	0.30	16.60		
C_2	69.79	70.07	0.28	27.08		
		$(c_1, c_2) = (1, 1.5)$, $(\rho_1, \rho_2) = (0.3, 0.3)$				
C	143.89	144.21	0.32	74.89	0.946	0.00505
C_1	64.68	64.88	0.20	17.53		
C_2	79.21	79.33	0.12	29.48		

δ で区間幅を調節するような

$$\boldsymbol{R}_{\boldsymbol{n},\delta} = \{\boldsymbol{\mu} \in R^p : \max\{\widehat{T}_{\boldsymbol{n}} - \delta,\ 0\} \leq ||\boldsymbol{\mu}||^2 \leq \max\{\widehat{T}_{\boldsymbol{n}} + \delta,\ 0\}\}$$

なる信頼領域を構築し，マイクロアレイデータの解析に利用している．

4.2 要求される有意水準と検出力をもつ二標本問題

2 つの母集団の平均ベクトル $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$ について，次の検定を考える．

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

高次元二標本問題について，Dempster [24], [25] は，母集団分布に正規性と $\Sigma_1 = \Sigma_2$ を仮定して検定統計量を与えた．Bai and Saranadasa [12] は，正規性の条件を緩めた仮定のもとで，下記の (22) で与えられる検定統計量を提案した．しかしながら，依然として $\Sigma_1 = \Sigma_2$ なる厳しい条件を仮定していた．最近，Chen and Qin [22] は，非正規かつ $\Sigma_1 \neq \Sigma_2$ のもとで，(22) で与えられる検定統計量の漸近分布を導出した．一方，Aoshima and Yata [8] は，非正規かつ $\Sigma_1 \neq \Sigma_2$ のもとで，$n_i/p \to 0$ において $\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ について与えられる要求 α，$\beta \in (0, 1/2)$，$\Delta_L = o(p^{1/2})$ (> 0) に対して，有意水準 (size) $\leq \alpha$ となり，かつ，$\Delta \geq \Delta_L$ のときに検出力 (power) $\geq 1 - \beta$ となるような検定方式を与えた．

各母集団から抽出される大きさ n_i の標本に基づいて，$\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ の推定量

$$\widetilde{T}_{\boldsymbol{n}} = \sum_{i=1}^{2} \frac{\sum_{j \neq j'}^{n_i} \boldsymbol{x}_{ij}^T \boldsymbol{x}_{ij'}}{n_i(n_i - 1)} - 2\frac{\sum_{j=1}^{n_1}\sum_{j'=1}^{n_2} \boldsymbol{x}_{1j}^T \boldsymbol{x}_{2j'}}{n_1 n_2} \tag{22}$$

を考える．そのとき，$E_{\boldsymbol{\theta}}(\widetilde{T}_{\boldsymbol{n}}) = \Delta$ となり

$$\mathrm{Var}_{\boldsymbol{\theta}}(\widetilde{T}_{\boldsymbol{n}}) = \sum_{i=1}^{2} \frac{2}{n_i(n_i - 1)}\mathrm{tr}(\Sigma_i^2) + \frac{4}{n_1 n_2}\mathrm{tr}(\Sigma_1\Sigma_2) + \sum_{i=1}^{2} \frac{4}{n_i}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

となる．そのとき，次の定理を得る．

定理 16 ([8], [22]) 母集団分布に (A-iii) を仮定する．正則条件として $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o\{\mathrm{tr}(\Sigma_i^2)/n_i\}$, $i = 1, 2$ を仮定する．そのとき，(A-iv) のもと $p \to \infty$, $n_i \to \infty$, $i = 1, 2$ の

とき次が成り立つ.

$$\frac{\widetilde{T}_{\boldsymbol{n}}-\Delta}{\sqrt{\mathrm{Var}_{\boldsymbol{\theta}}(\widetilde{T}_{\boldsymbol{n}})}} \Rightarrow N(0,1).$$

Aoshima and Yata [8] は，各母集団の標本数を

$$n_i \geq \frac{(z_\alpha+z_\beta)\sqrt{2}}{\Delta_L}\mathrm{tr}(\Sigma_i^2)^{1/4}\sum_{j=1}^{2}\mathrm{tr}(\Sigma_j^2)^{1/4} \quad (=C_i\text{とおく}) \tag{23}$$

を満たす最小の整数とし，(22) の検定統計量に基づく検定方式を

$$H_0\text{を棄却} \iff \widetilde{T}_{\boldsymbol{n}} > \frac{\Delta_L z_\alpha}{z_\alpha+z_\beta} \tag{24}$$

で定義した．そのとき，次の定理を得る.

定理 17 ([8]) 母集団分布に (A-ii) を仮定する．標本数 n_1, n_2 は (23) を満たすものとする．(A-iv) のもと $p\to\infty$ のとき，検定方式 (24) は次が成り立つ.

$$\limsup \mathrm{size} \leq \alpha, \qquad \liminf \mathrm{power}(\Delta_L) \geq 1-\beta. \tag{25}$$

ただし，$\mathrm{power}(\Delta_L)$ は $\Delta=\Delta_L$ のときの検出力である.

(23) における $\mathrm{tr}(\Sigma_i^2)$ は未知である．Aoshima and Yata [8], [9] は，3.3 節で紹介したクロスデータ行列法を用いて 2 つの異なる C_i の推定法を提示し，(25) を満たすような検定方式を二段階推定法で与えている．そのときの標本数 N_i は，$N_i/p=o_P(1)$, $p\to\infty$ が成り立つ.

注意 7 Aoshima and Yata [8] は，高次元小標本における様々な検定問題を扱っている．例えば，$H:\mathrm{tr}(\Sigma_1)=\mathrm{tr}(\Sigma_2)$ の検定や，$\boldsymbol{\mu}_1-\boldsymbol{\mu}_2$ の成分に関する多重検定および変数選択の問題にも，有意水準と検出力に関して精度を保証する検定方式を与えている．一方で，Srivastava [45] も高次元における検定問題を扱っているが，母集団分布に正規性を仮定し，導かれる結果も正規分布に限定される結果であることに注意する.

5 高次元小標本における各種パラメータの推定と検定

Yata and Aoshima [58] は，3.3 節で紹介したクロスデータ行列法における 2 分割の組合せを考慮して，'拡張クロスデータ行列法 (ECDM)' を開発した．本節では，ECDM を用いて各種パラメータの推定量や検定統計量の構築を解説する．母集団は p 次の平均ベクトル $\boldsymbol{\mu}$，共分散行列 Σ $(>O)$ をもち，母集団から n (≥ 4) 個の p 次元データ $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ が無作為に抽出されたとする．4 節と同様に，データには $\boldsymbol{x}_j=\Gamma\boldsymbol{w}_j+\boldsymbol{\mu}$, $j=1,\ldots,n$ なるモデルを考える．ここで，Γ は $\Gamma\Gamma^T=\Sigma$ となる $p\times r$ 行列とし，$\boldsymbol{w}_j$ は $E(\boldsymbol{w}_j)=\boldsymbol{0}$, $\mathrm{Var}(\boldsymbol{w}_j)=\boldsymbol{I}_r$ とする．ただし，$r\geq p$ である．いま，$\boldsymbol{w}_j=(w_{1j},\ldots,w_{rj})^T$ とおき，各成分は 4 次モーメントが一様有界と仮定する．母集団分布には，次の 2 つのどちらか一つを仮定する：

(A-v) w_{ij}, $i=1,\ldots,r$ は互いに独立である,

(A-vi) $E(w_{ij}w_{sj}w_{tj})=0$, $E(w_{ij}^2w_{sj}^2)=1$, $E(w_{ij}w_{sj}w_{tj}w_{uj})=0$, $i\neq s,t,u$.

(A-vi) は (A-v) を緩めた条件であり，(A-v) は正規性を緩めた条件である.

5.1 $\mathrm{tr}(\Sigma^2)$ の推定量

高次元小標本の推測に精度保証をする上で，しばしば重要になるのが $\mathrm{tr}(\Sigma^2)$ の推定である．高次元データに対しては，単純な推定量 $\mathrm{tr}(S^2)$ は非常に大きなバイアスをもつので役に立たない．Yata [50] は，Yata and Aoshima [54] で開発されたクロスデータ行列法を用いて $\mathrm{tr}(\Sigma^2)$ の推定を考えた．標本を 2 分割し，$n_{(1)} = \langle n/2 \rangle + 1$，$n_{(2)} = n - n_{(1)}$ とおく．それぞれの分割から標本共分散行列

$$S_{n(1)} = (n_{(1)} - 1)^{-1} \sum_{j=1}^{n_{(1)}} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_{n_{(1)}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_{n_{(1)}})^T,$$

$$S_{n(2)} = (n_{(2)} - 1)^{-1} \sum_{j=n_{(1)}+1}^{n} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_{n_{(2)}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_{n_{(2)}})^T$$

を計算する．ここで，$\overline{\boldsymbol{x}}_{n_{(1)}} = \sum_{j=1}^{n_{(1)}} \boldsymbol{x}_j / n_{(1)}$，$\overline{\boldsymbol{x}}_{n_{(2)}} = \sum_{j=n_{(1)}+1}^{n} \boldsymbol{x}_j / n_{(2)}$ である．そのとき，不偏性 $E\{\mathrm{tr}(S_{n(1)} S_{n(2)})\} = \mathrm{tr}(\Sigma^2)$ が示され，$\mathrm{tr}(S_{n(1)} S_{n(2)})$ は $\mathrm{tr}(\Sigma^2)$ の不偏推定量になる．母集団分布に (A-vi) を仮定すると，$p \to \infty$, $n \to \infty$ のとき $\mathrm{Var}\{\mathrm{tr}(S_{n(1)} S_{n(2)})/\mathrm{tr}(\Sigma^2)\} = 8\{1 + o(1)\}/n^2 + O[\mathrm{tr}(\Sigma^4)/\{n\mathrm{tr}(\Sigma^2)^2\}] \to 0$ となり，$n/p \to 0$ なる高次元小標本の枠組みで一致性が主張できる．

Bai and Saranadasa [12], Srivastava [44] は，推定量 $\mathrm{tr}(\widehat{\Sigma^2}) = c_n^{-1}\{\mathrm{tr}(S_n^2) - \mathrm{tr}(S_n)^2/(n-1)\}$ を与えた．ここで，S_n は不偏標本共分散行列であり，$c_n = (n-2)(n+1)/(n-1)^2$ である．そのとき，母集団分布に正規性を仮定すれば，$E\{\mathrm{tr}(\widehat{\Sigma^2})\} = \mathrm{tr}(\Sigma^2)$ なる不偏性をもち，$p \to \infty$, $n \to \infty$ のとき $\mathrm{Var}\{\mathrm{tr}(\widehat{\Sigma^2})/\mathrm{tr}(\Sigma^2)\} = 4\{1 + o(1)\}/n^2 + 8\mathrm{tr}(\Sigma^4)\{1 + o(1)\}/\{n\mathrm{tr}(\Sigma^2)^2\} \to 0$ となる．しかしながら，母集団分布が非正規の場合には，$\mathrm{tr}(\widehat{\Sigma^2})$ の不偏性は主張できず，高次元において非常に大きなバイアスを生じる．これは，$\mathrm{tr}(\widehat{\Sigma^2})$ が正規分布の 4 次モーメントの情報まで必要とすることに起因する．実際，(A-vi) のもとでは，$E(\mathrm{tr}(\widehat{\Sigma^2})) = \mathrm{tr}(\Sigma^2) + O\{\mathrm{tr}(\Sigma^2)/n\}$ となる．さらに，$\boldsymbol{w}_j$ の成分について 8 次モーメントの一様有界性まで仮定できないと，$\mathrm{Var}\{\mathrm{tr}(\widehat{\Sigma^2})/\mathrm{tr}(\Sigma^2)\} < \infty$ さえ保証できない．上記以外に，Chen *et al.* [23] も $\mathrm{tr}(\Sigma^2)$ の U-統計量に基づく不偏推定量を提案しているが，計算コストが $O(pn^4)$ と非常に大きく，実用に向かない．

Yata and Aoshima [58] は，拡張クロスデータ行列法 (ECDM) を次のように開発した：2 つの集合 $\boldsymbol{V}_{n(1)(k)}$, $\boldsymbol{V}_{n(2)(k)}$ $(k = 3, \ldots, 2n-1)$ を次のように定義する．

$$\boldsymbol{V}_{n(1)(k)} = \begin{cases} \{\lfloor k/2 \rfloor - n_{(1)} + 1, \ldots, \lfloor k/2 \rfloor\}, & \lfloor k/2 \rfloor \geq n_{(1)} \text{のとき}, \\ \{1, \ldots, \lfloor k/2 \rfloor\} \cup \{\lfloor k/2 \rfloor + n_{(2)} + 1, \ldots, n\}, & \text{それ以外}, \end{cases}$$

$$\boldsymbol{V}_{n(2)(k)} = \begin{cases} \{\lfloor k/2 \rfloor + 1, \ldots, \lfloor k/2 \rfloor + n_{(2)}\}, & \lfloor k/2 \rfloor \leq n_{(1)} \text{のとき}, \\ \{1, \ldots, \lfloor k/2 \rfloor - n_{(1)}\} \cup \{\lfloor k/2 \rfloor + 1, \ldots, n\}, & \text{それ以外}. \end{cases}$$

ここで，$\lfloor x \rfloor$ は x 以下の最大の整数を表す．そのとき，$k = 3, \ldots, 2n-1$ について，$|\boldsymbol{V}_{n(l)(k)}| = n_{(l)}$, $l = 1, 2$, $\boldsymbol{V}_{n(1)(k)} \cap \boldsymbol{V}_{n(2)(k)} = \emptyset$, $\boldsymbol{V}_{n(1)(k)} \cup \boldsymbol{V}_{n(2)(k)} = \{1, \ldots, n\}$ となること，及び，$i < j$ $(\leq n)$ について

$$i \in \boldsymbol{V}_{n(1)(i+j)}, \quad j \in \boldsymbol{V}_{n(2)(i+j)}$$

となることに注意する．ここで，$|S|$ は集合 S の要素の個数を表す．$\boldsymbol{V}_{n(1)(i+j)}$ と $\boldsymbol{V}_{n(2)(i+j)}$ に対応してデータ集合を 2 分割し，それらに基づいて不偏推定量を計算して，$i<j$ $(\leq n)$ のすべての組合せについて平均をとる．これが ECDM である．

例えば，$k=3,\ldots,2n-1$ について，標本平均を

$$\overline{\boldsymbol{x}}_{n(1)(k)}=n_{(1)}^{-1}\sum_{j\in \boldsymbol{V}_{n(1)(k)}}\boldsymbol{x}_j,\qquad \overline{\boldsymbol{x}}_{n(2)(k)}=n_{(2)}^{-1}\sum_{j\in \boldsymbol{V}_{n(2)(k)}}\boldsymbol{x}_j$$

とし，$\mathrm{tr}(\Sigma^2)$ の不偏推定量として $u_n\{(\boldsymbol{x}_i-\overline{\boldsymbol{x}}_{n(1)(i+j)})^T(\boldsymbol{x}_j-\overline{\boldsymbol{x}}_{n(2)(i+j)})\}^2$ $(i<j)$ を計算する．ただし，$u_n=n_{(1)}n_{(2)}/\{(n_{(1)}-1)(n_{(2)}-1)\}$ である．すべての組合せの平均をとって

$$W_n=\frac{2u_n}{n(n-1)}\sum_{i<j}^{n}\{(\boldsymbol{x}_i-\overline{\boldsymbol{x}}_{n(1)(i+j)})^T(\boldsymbol{x}_j-\overline{\boldsymbol{x}}_{n(2)(i+j)})\}^2 \tag{26}$$

を定義する．そのとき，W_n は，母集団分布に依らずに不偏性 $E(W_n)=\mathrm{tr}(\Sigma^2)$ が成り立ち，一致性について次の定理を得る．

定理 18 ([11], [58])　母集団分布に (A-vi) を仮定する．$p\to\infty$, $n\to\infty$ のとき，次が成り立つ．

$$\mathrm{Var}\Big(\frac{W_n}{\mathrm{tr}(\Sigma^2)}\Big)=\frac{4}{n^2}\{1+o(1)\}+O\Big\{\frac{\mathrm{tr}(\Sigma^4)}{n\mathrm{tr}(\Sigma^2)^2}\Big\}\to 0.$$

ECDM はクロスデータ行列法よりも漸近分散が小さい不偏推定量を構築できる．また，ECDM の計算コストは $O(pn^2)$ なので，U-統計量に基づくものよりも高速で実用的である．さらに，母集団分布に正規性を仮定すると $\mathrm{Var}\{W_n/\mathrm{tr}(\Sigma^2)\}=4\{1+o(1)\}/n^2+8\mathrm{tr}(\Sigma^4)\{1+o(1)\}/\{n\mathrm{tr}(\Sigma^2)^2\}$ となり，正規分布に限定して提案された前掲の $\mathrm{tr}(\widehat{\Sigma^2})$ と同等の漸近分散をもつ．

注意 8　Aoshima and Yata [10] は，クロスデータ行列法におけるデータ集合の 2 分割を計算機上でランダムに実行して推定量を構築するという '一般化クロスデータ行列法 (GCDM)' とよぶ方法を提案した．

5.2　高次元における相関ベクトルの検定

母集団に $p+1$ 次元の分布を考え，n 個のデータ $\boldsymbol{x}_{1(*)},\ldots,\boldsymbol{x}_{n(*)}$ を無作為に抽出したとする．ただし，$\boldsymbol{x}_{j(*)}=(\boldsymbol{x}_j^T,x_{j(*)})^T$, $j=1,\ldots,n$ とする．ここで，$x_{j(*)}$ の分散を $\mathrm{Var}(x_{j(*)})=\sigma_*^2$, $\boldsymbol{x}_j$ と $x_{j(*)}$ の相関ベクトルを $\mathrm{Corr}(\boldsymbol{x}_j,x_{j(*)})=\boldsymbol{\rho}$, 共分散ベクトルを $\mathrm{Cov}(\boldsymbol{x}_j,x_{j(*)})=\boldsymbol{\sigma}$ とおく．そのとき，次の検定を考える．

$$H_0:\boldsymbol{\rho}=\boldsymbol{0}\quad\text{vs.}\quad H_1:\boldsymbol{\rho}\neq\boldsymbol{0}.$$

高次元における相関の検定は，グラフィカルモデリングやパス解析等に用いられる．詳細は，例えば，Drton and Perlman [26] や Hero and Rajaratnam [34], Wille *et al.* [48] 等を参照のこと．Aoshima and Yata [8] は，クロスデータ行列法を用いて検定統計量を構築し，高次元漸近正規性を証明した．本節では，Aoshima and Yata [8] の結果を ECDM を用いて発展させた Yata and Aoshima [58] の検定統計量を紹介する．

いま，$k=3,\ldots,2n-1$ について

$$\overline{x}_{n(1*)(k)} = n_{(1)}^{-1} \sum_{j \in V_{n(1)(k)}} x_{j(*)}, \qquad \overline{x}_{n(2*)(k)} = n_{(2)}^{-1} \sum_{j \in V_{n(2)(k)}} x_{j(*)}$$

を計算し，検定統計量を

$$\widehat{T}_{n,\boldsymbol{\sigma}} = \frac{2u_n}{n(n-1)} \sum_{i<j}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{n(1)(i+j)})^T (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_{n(2)(i+j)}) \times (x_{i(*)} - \overline{x}_{n(1*)(i+j)})(x_{j(*)} - \overline{x}_{n(2*)(i+j)})$$

で定義する．そのとき，母集団分布に依らずに不偏性 $E(\widehat{T}_{n,\boldsymbol{\sigma}}) = ||\boldsymbol{\sigma}||^2$ が成り立つ．

注意 9 Zhong *et al.* [60] は，回帰モデルにおいて U-統計量に基づく $||\boldsymbol{\sigma}||^2$ の不偏推定量を提案しているが，計算コストが $O(pn^4)$ と非常に大きく，実用に向かない．

次の 2 つを仮定する：

(A-vii) $p \to \infty$ のとき $\mathrm{tr}(\Sigma^4)/\mathrm{tr}(\Sigma^2)^2 \to 0$,

(A-viii) $p \to \infty$, $n \to \infty$ と $||\boldsymbol{\sigma}|| \neq 0$ のとき，$\liminf \sigma_*^2 \sqrt{\mathrm{tr}(\Sigma^2)}/(n||\boldsymbol{\sigma}||^2) > 0$.

そのとき，次の定理を得る．

定理 19 ([58]) 母集団分布に (A-v) を仮定する．そのとき，$x_{j(*)}$ のある正則条件と (A-vii)–(A-viii) のもと $p \to \infty, n \to \infty$ のとき，次が成り立つ．

$$\frac{\widehat{T}_{n,\boldsymbol{\sigma}} - ||\boldsymbol{\sigma}||^2}{S_{n(*)}\sqrt{2W_n}/n} \Rightarrow N(0,1).$$

ここで，$S_{n(*)}$ は σ_*^2 の不偏標本分散である．

この定理に基づいて，Yata and Aoshima [58] は，有意水準と検出力に関して要求される精度を保証する検定方式を与え，マイクロアレイデータのパスウェイ解析に応用した．

6 高次元小標本データの判別分析

高次元データの所属について，2 群の判別問題を考える．高次元小標本においては標本共分散行列 $\boldsymbol{S}_{in_i}$ の逆行列が存在しない．そのため，Fisher の線形判別方式や 2 次判別方式は適用できない．2 群の共分散行列が等しいと仮定する場合，Saranadasa [43] による単位行列を代入した判別方式や，Bickel and Levina [16] による標本共分散行列の対角成分だけを使った判別方式，Srivastava and Kubokawa [46] によるリッジ型逆行列による判別方式がある．これらに対し，Yata and Aoshima [56] は，3.2 節で紹介したノイズ掃き出し法を用いて，推定した固有空間から共分散行列の逆行列を構築して判別方式を与え，その判別精度が先行研究よりも優ることを示した．しかしながら，2 群の共分散行列が等しいと仮定する問題設定の単純化は，高次元データが本来もつ 2 群の差異に関する情報に目を瞑ることになり，望ましいものではない．2 群の共分散行列が等しいことを仮定しない場合，Dudoit *et al.* [27] による標本共分散行列の対角成分だけを使った判別方式や，Chan and Hall [21], Hall *et al.* [31], [32], Aoshima and Yata [11] 等によるユークリッド距離に基づく判別方式，そして，Aoshima and Yata [8] が与えた高次元小標本データの幾何学的表現に基づく判別方式がある．他に，Aoshima and Yata [8], Fan and Fan [28] 等の変数選択に基づく判別方式もある．本節では，高次元小標本における理論的な方法論の構築に特に重要な，Aoshima and Yata [8] の判別方式を紹

介する．2群の共分散行列の差異がもたらす高次元小標本データの幾何学的表現の違いに着目した方法論として，その統計的推測は興味深いであろう．

6.1 幾何学的表現に基づく判別方式

2つの母集団 π_i, $i=1,2$ は，p 次の平均ベクトル $\boldsymbol{\mu}_i$ と共分散行列 Σ_i $(>\boldsymbol{O})$ をもつとする．各母集団 π_i から，n_i (≥ 4) 個のデータ $\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{in_i}$ を抽出する．母集団分布には4節と同様の仮定をする．判別の対象となる個体のデータを $\boldsymbol{x}_0$ とする．Aoshima and Yata [8] は，高次元小標本データの幾何学的表現に基づく次のような判別関数 $\omega(\boldsymbol{x}_0|\gamma)$ を考えた．

$$\omega(\boldsymbol{x}_0|\gamma)=\frac{p||\boldsymbol{x}_0-\overline{\boldsymbol{x}}_{1n_1}||^2}{\mathrm{tr}(\boldsymbol{S}_{1n_1})}-\frac{p||\boldsymbol{x}_0-\overline{\boldsymbol{x}}_{2n_2}||^2}{\mathrm{tr}(\boldsymbol{S}_{2n_2})}-p\log\left\{\frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{\mathrm{tr}(\boldsymbol{S}_{1n_1})}\right\}-\frac{p}{n_1}+\frac{p}{n_2}+\gamma. \qquad (27)$$

判別方式は，$\omega(\boldsymbol{x}_0|\gamma)<0$ のとき $\boldsymbol{x}_0\in\pi_1$，$\omega(\boldsymbol{x}_0|\gamma)\geq 0$ のとき $\boldsymbol{x}_0\in\pi_2$ とする．γ は要求精度に依存して決まる値であり，これについては後ほど触れることにする．

誤判別確率を，π_1 の $\boldsymbol{x}_0$ を π_2 に誤判別する場合に $e(2|1)$ と表記し，$e(1|2)$ も同様の表記とする．$\Delta=||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||^2$, $\Delta_{\Sigma_i}=\{\mathrm{tr}(\Sigma_1)-\mathrm{tr}(\Sigma_2)\}^2/\mathrm{tr}(\Sigma_i)$, $i=1,2$ とおき，$\Delta_i=\Delta+\Delta_{\Sigma_i}/2$, $i=1,2$ とする．そのとき，$\Delta_\star=\min_{i=1,2}\Delta_i$ は2群間の差異を表すパラメータと考えられる．次の2つを仮定する：

(A-ix)　各 i で，$p\to\infty$ のとき $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\Sigma_i(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)/\Delta_\star^2\to 0$ かつ $\mathrm{tr}(\Sigma_i^2)\{\mathrm{tr}(\Sigma_1)-\mathrm{tr}(\Sigma_2)\}^2/(\mathrm{tr}(\Sigma_i)^2\Delta_\star^2)\to 0$,

(A-x)　各 i で，$p\to\infty$ のとき $\max_{j=1,2}\mathrm{tr}(\Sigma_j^2)/(n_i\Delta_\star^2)\to 0$.

そのとき，誤判別確率に関する一致性に，次の定理を得る．

定理 20 ([8], [41])　母集団分布に (A-iii) を仮定する．(A-ix)–(A-x) のもと，$\gamma=0$ とした判別方式 (27) は，$p\to\infty$ のとき次が成り立つ．

$$e(2|1)\to 0,\qquad e(1|2)\to 0.$$

$\Delta_\star$ が大きいほど誤判別確率の収束は速い．判別方式 (27) は，$\Delta_\star$ を通して，$\boldsymbol{\mu}_i$, $i=1,2$ の差異だけでなく Σ_i, $i=1,2$ の差異も考慮することで，判別精度を向上させている．

$\omega(\boldsymbol{x}_0|0)$ の漸近的挙動を簡単なシミュレーション実験で示す．2群は $\pi_1:N_p(0,\boldsymbol{I}_p)$, $\pi_2:N_p(0,2\boldsymbol{I}_p)$ とし，$\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$ の状況を設定した．平均間の距離のみに基づく判別方式では，この状況における2群を識別することはできない．いま，$n_1=n_2=5$ と設定し，$p=4,32,256,2048$ の4つの場合を考える．これらの設定は，仮定 (A-iii), (A-ix), (A-x) を満たす．図4は，A: $\boldsymbol{x}_0\in\pi_1$ のとき，B: $\boldsymbol{x}_0\in\pi_2$ のときについて，それぞれ2000回のシミュレーションによる $\omega(\boldsymbol{x}_0|0)$ のヒストグラムを与えている．定理20で主張する通り，次元数が上がるにつれて，2つのヒストグラムが原点を境に完全に分離していく様子が見てとれる．

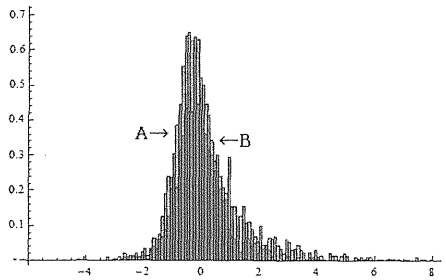
6.2 判別関数の漸近正規性と精度保証

2群の差異を表す $\Delta_\star$ が (A-ix)–(A-x) を満たすほど大きくないときには，標本数 n_i を次元数 p に依存して設計することで，判別方式 (27) の精度は保証される．次の2つを仮定する：

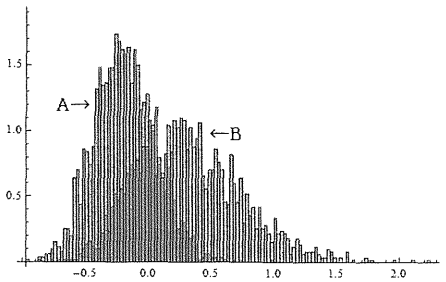
(A-xi)　$p\to\infty$ のとき，$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\Sigma_i(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)/\Delta_\star^2\to 0$ $(i=1,2)$,

(A-xii)　$p\to\infty$ と $n\to\infty$ のとき，$\mathrm{tr}(\Sigma_i^2)/(n_i^2\Delta_\star^2)\to 0$, $\liminf\mathrm{tr}(\Sigma_i^2)/(n_i\Delta_\star^2)>0$ $(i=1,2)$.

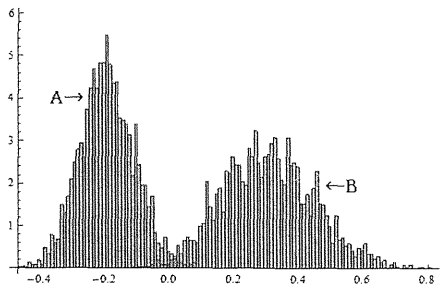
そのとき，次の定理を得る．



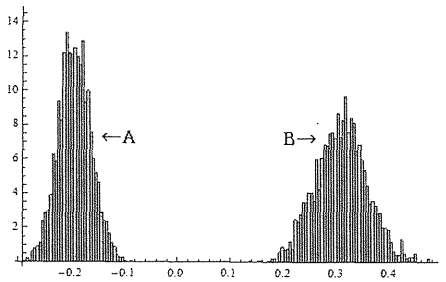
(a) $p = 4$



(b) $p = 32$



(c) $p = 256$



(d) $p = 2048$

図 4：$\omega(\boldsymbol{x}_0|0)$ のヒストグラム．A：$\boldsymbol{x}_0 \in \pi_1(N_p(\boldsymbol{0}, \boldsymbol{I}_p))$ のとき，B：$\boldsymbol{x}_0 \in \pi_2(N_p(\boldsymbol{0}, 2\boldsymbol{I}_p))$ のとき．

定理 21 ([8]) 母集団分布に (A-ii) を仮定する．さらに，$\mathrm{tr}(\Sigma_1)/\mathrm{tr}(\Sigma_2) \to 1$, $p \to \infty$ も仮定する．そのとき，(A-iv)，(A-xi)–(A-xii) のもと，$p \to \infty$ のとき次が成り立つ．

$$\frac{\omega(\boldsymbol{x}_0|0) + \Delta_2\{\mathrm{tr}(\Sigma_2)/p\}^{-1}}{2\sqrt{\{\mathrm{tr}(\Sigma_1)/p\}^{-2}\mathrm{tr}(\Sigma_1^2)/n_1 + \{\mathrm{tr}(\Sigma_2)/p\}^{-2}\mathrm{tr}(\Sigma_1\Sigma_2)/n_2}} \Rightarrow N(0,1), \quad \boldsymbol{x}_0 \in \pi_1 \text{ のとき},$$

$$\frac{\omega(\boldsymbol{x}_0|0) - \Delta_1\{\mathrm{tr}(\Sigma_1)/p\}^{-1}}{2\sqrt{\{\mathrm{tr}(\Sigma_2)/p\}^{-2}\mathrm{tr}(\Sigma_2^2)/n_2 + \{\mathrm{tr}(\Sigma_1)/p\}^{-2}\mathrm{tr}(\Sigma_1\Sigma_2)/n_1}} \Rightarrow N(0,1), \quad \boldsymbol{x}_0 \in \pi_2 \text{ のとき}.$$

定理 21 を用いて，判別方式 (27) の精度保証が考えられる．与えられる α, $\beta \in (0, 1/2)$, $\Delta_L = o(p^{1/2})$ (> 0) に対して，$\Delta_\star \geq \Delta_L$ のときの誤判別確率が $e(2|1) \leq \alpha$, $e(1|2) \leq \beta$ となることを要求精度とする．母集団 π_i $(i = 1, 2)$ の標本数を

$$n_i \geq \frac{(z_\alpha + z_\beta)^2\sigma}{\Delta_L^2}\mathrm{tr}(\Sigma_i^2)^{1/4}\sum_{j=1}^{2}\mathrm{tr}(\Sigma_j^2)^{1/4} \quad (= C_i \text{ とおく}) \tag{28}$$

を満たす最小の整数とする．$\sigma = \max\{\mathrm{tr}(\Sigma_1^2)^{1/2}, \mathrm{tr}(\Sigma_2^2)^{1/2}\}$ である．そのとき，次の定理を得る．

定理 22 ([8]) 母集団分布に (A-ii) を仮定する．n_1, n_2 は (28) を満たすとする．判別方式 (27) において $\gamma = \{\mathrm{tr}(\boldsymbol{S}_{1n_1} + \boldsymbol{S}_{2n_2})/(2p)\}^{-1}\Delta_L(z_\beta - z_\alpha)/(z_\alpha + z_\beta)$ とおく．そのとき，(A-iv) と (A-xi) のもと，$\Delta_\star \geq \Delta_L$ なる $\Delta_\star$ に対して $p \to \infty$ のとき次が成り立つ．

$$\limsup e(2|1) \leq \alpha, \qquad \limsup e(1|2) \leq \beta. \tag{29}$$

(28) における $\mathrm{tr}(\Sigma_i^2)$ は未知なので，Aoshima and Yata [8] はクロスデータ行列法を用いて C_i を推定し，二段階推定法による判別方式が (29) を満たすことを証明した．二段階推定法による標本数 N_i は，$N_i/p = o_P(1)$, $p \to \infty$ が成り立ち，高次元小標本の枠組みで解が得られている．

注意 10 精度保証に関する $(\alpha, \beta, \Delta_L)$ の定め方や諸注意について，Aoshima and Yata [9], [10] を参照のこと．なお，Aoshima and Yata [8] は，高次元小標本における重回帰分析を用いた判別方式や，変数選択を伴う判別方式も扱い，精度を保証する解を与えている．

謝辞 本論説の改訂にあたり，細部にわたり有益なご指摘・ご助言を下さった査読者の方々に，厚く御礼申し上げます．本研究は，科学研究費補助金 基盤研究 (B) 22300094 研究代表者：青嶋 誠 '高次元データの理論と方法論の総合的研究'，および，学術研究助成基金助成金 挑戦的萌芽研究 23650142 研究代表者：青嶋 誠 '高速で頑健かつ高精度な多変量統計手法の新展開'，若手研究 (B) 23740066 研究代表者：矢田和善 '高次元小標本の理論的体系の構築' から研究助成を受けています．

文　献

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

[2] 青嶋誠, 二段階標本抽出による統計的推測, 数学, **54** (2002), 365–382.

[3] M. Aoshima, P. Chen and S. Panchapakesan, Sequential procedures for selecting the most probable multinomial cell when a nuisance cell is present, Comm. Statist. Theory Methods, **32** (2003), 893–906.

[4] M. Aoshima and N. Mukhopadhyay, Fixed-width simultaneous confidence intervals for multinormal means in several intraclass correlation models, J. Multivariate Anal., **66** (1998), 46–63.

[5] M. Aoshima and Y. Takada, Asymptotic second-order efficiency for multivariate two-stage estimation of a linear function of normal mean vectors, Sequential Anal., **23** (2004), 333–353.

[6] M. Aoshima, Y. Takada and M. S. Srivastava, A two-stage procedure for estimating a linear function of k multinormal mean vectors when covariance matrices are unknown, J. Statist. Plann. Inference, **100** (2002), 109–119.

[7] M. Aoshima and K. Yata, Asymptotic second-order consistency for two-stage estimation methodologies and its applications, Ann. Inst. Statist. Math., **62** (2010), 571–600.

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

[9] M. Aoshima and K. Yata, Authors' response, Sequential Anal., **30** (2011), 432–440.

[10] M. Aoshima and K. Yata, Effective methodologies for statistical inference on microarray studies, In: Prostate Cancer—From Bench to Bedside, (ed. P. E. Spiess), InTech, 2011, pp. 13–32.

[11] M. Aoshima and K. Yata, A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data, submitted, 2012.

[12] Z. Bai and H. Saranadasa, Effect of high dimension: by an example of a two sample problem, Statist. Sinica, **6** (1996), 311–329.

[13] Z. Bai and J. W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, Second edition, Springer Ser. Statist., Springer, New York, 2010.

[14] J. Baik, G. Ben Arous and S. Péché, Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices, Ann. Probab., **33** (2005), 1643–1697.

[15] J. Baik and J. W. Silverstein, Eigenvalues of large sample covariance matrices of spiked population models, J. Multivariate Anal., **97** (2006), 1382–1408.

[16] P. J. Bickel and E. Levina, Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations, Bernoulli, **10** (2004), 989–1010.

[17] P. J. Bickel and E. Levina, Covariance regularization by thresholding, Ann. Statist., **36** (2008), 2577–2604.

[18] P. J. Bickel and E. Levina, Regularized estimation of large covariance matrices, Ann. Statist., **36** (2008), 199–227.

[19] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.

[20] R. C. Bradley, Basic properties of strong mixing conditions. A survey and some open questions, Probab. Surv., **2** (2005), 107–144 (electronic).

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

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

[23] S. X. Chen, L.-X. Zhang and P.-S. Zhong, Tests for high-dimensional covariance matrices, J. Amer. Statist. Assoc., **105** (2010), 810–819.

[24] A. P. Dempster, A high dimensional two sample significance test, Ann. Math. Statist., **29** (1958), 995–1010.

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

[26] M. Drton and M. D. Perlman, Multiple testing and error control in Gaussian graphical model

selection, Statist. Sci., **22** (2007), 430–449.
[27] S. Dudoit, J. Fridlyand and T. P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, J. Amer. Statist. Assoc., **97** (2002), 77–87.
[28] J. Fan and Y. Fan, High-dimensional classification using features annealed independence rules, Ann. Statist., **36** (2008), 2605–2637.
[29] Y. Fujikoshi, V. Ulyanov and R. Shimizu, Multivariate Statistics: High-Dimensional and Large-Sample Approximations, Wiley Ser. Probab. Statist., Wiley, Hoboken, NJ, 2010.
[30] M. Ghosh, N. Mukhopadhyay and P. K. Sen, Sequential Estimation, Wiley Ser. Probab. Statist., Wiley, New York, 1997.
[31] P. Hall, J. S. Marron and A. Neeman, Geometric representation of high dimension, low sample size data, J. R. Stat. Soc. Ser. B Stat. Methodol, **67** (2005), 427–444.
[32] P. Hall, Y. Pittelkow and M. Ghosh, Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes, J. R. Stat. Soc. Ser. B Stat. Methodol, **70** (2008), 159–173.
[33] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second edition, Springer, New York, 2009.
[34] A. Hero and B. Rajaratnam, Large-scale correlation screening, J. Amer. Statist. Assoc., **106** (2011), 1540–1552.
[35] I. M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, Ann. Statist., **29** (2001), 295–327.
[36] S. Jung and J. S. Marron, PCA consistency in high dimension, low sample size context, Ann. Statist., **37** (2009), 4104–4130.
[37] S. Jung, A. Sen and J. S. Marron, Boundary behavior in high dimension, low sample size asymptotics of PCA, J. Multivariate Anal., **109** (2012), 190–203.
[38] S. Lee, F. Zou and F. A. Wright, Convergence and prediction of principal component scores in high-dimensional settings, Ann. Statist., **38** (2010), 3605–3629.
[39] N. Mukhopadhyay and W. T. Duggan, Can a two-stage procedure enjoy second-order properties?, Sankhyā Ser. A, **59** (1997), 435–448.
[40] N. Mukhopadhyay and W. Duggan, On a two-stage procedure having second-order properties with applications, Ann. Inst. Statist. Math., **51** (1999), 621–636.
[41] K. Nagahashi, K. Yata and M. Aoshima, Note on classification for high-dimensional data, In: A New Perspective to Statistical Models and Its Related Topics, (eds. M. Akahira and K. Koike), RIMS Kôkyûroku, **1804** (2012), 40–52.
[42] D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, Statist. Sinica, **17** (2007), 1617–1642.
[43] H. Saranadasa, Asymptotic expansion of the misclassification probabilities of D- and A-criteria for discrimination from two high dimensional populations using the theory of large dimensional random matrices, J. Multivariate Anal., **46** (1993), 154–174.
[44] M. S. Srivastava, Some tests concerning the covariance matrix in high dimensional data, J. Japan Statist. Soc., **35** (2005), 251–272.
[45] M. S. Srivastava, Multivariate theory for analyzing high dimensional data, J. Japan Statist. Soc., **37** (2007), 53–86.
[46] M. S. Srivastava and T. Kubokawa, Comparison of discrimination methods for high dimensional data, J. Japan Statist. Soc., **37** (2007), 123–134.
[47] V. N. Vapnic, The Nature of Statistical Learning Theory, Statist. Engineering Information Science, Springer, New York, 1995.
[48] A. Wille, P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelic, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem and P. Bühlmann, Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana, Genome Biol., **5** (2004), R92.
[49] K. Yata, Two-stage equivalence tests that control both size and power, Sequential Anal., **27** (2008), 185–200.
[50] K. Yata, Effective two-stage estimation for a linear function of high-dimensional Gaussian means, Sequential Anal., **29** (2010), 463–482.
[51] K. Yata, Effective methodologies for high-dimension, low sample size data, In: Statistical Experiment and Its Related Topics, (eds. M. Akahira and K. Koike), RIMS Kôkyûroku, **1703** (2010), 180–194.
[52] K. Yata and M. Aoshima, Double shrink methodologies to determine the sample size via covariance structures, J. Statist. Plann. Inference, **139** (2009), 81–99.
[53] K. Yata and M. Aoshima, PCA consistency for non-Gaussian data in high dimension, low sample size context, Comm. Statist. Theory Methods, Special Issue Honoring S. Zacks, (ed. N. Mukhopadhyay), **38** (2009), 2634–2652.
[54] K. Yata and M. Aoshima, Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix, J. Multivariate Anal., **101** (2010), 2060–2077.
[55] K. Yata and M. Aoshima, Intrinsic dimensionality estimation of high-dimension, low sample size data with D-asymptotics, Comm. Statist. Theory Methods, Special Issue Honoring M. Akahira, (ed.

M. Aoshima), **39** (2010), 1511–1521.
[56] K. Yata and M. Aoshima, Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations, J. Multivariate Anal., **105** (2012), 193–215.
[57] K. Yata and M. Aoshima, Inference on high-dimensional mean vectors with fewer observations than the dimension, Methodol. Comput. Appl. Probab., **14** (2012), 459–476.
[58] K. Yata and M. Aoshima, Correlation tests for high-dimensional data using extended cross-data-matrix methodology, J. Multivariate Anal., in press, 2013.
[59] K. Yata and M. Aoshima, PCA consistency for the power spiked model in high-dimensional settings, submitted, 2013.
[60] P.-S. Zhong and S. X. Chen, Tests for high-dimensional regression coefficients with factorial designs, J. Amer. Statist. Assoc., **106** (2011), 260–274.

(2012 年 6 月 25 日提出)
(あおしま まこと・筑波大学数理物質系)
(やた かずよし・筑波大学数理物質系)