数理解析研究所講究録
第1804巻 2012年 40-52

Note on classification for high-dimensional data

筑波大学大学院・数理物質科学研究科　永橋　幸大 (Kota Nagahashi)
Graduate School of Pure and Applied Sciences
University of Tsukuba

筑波大学・数理物質系　矢田　和善 (Kazuyoshi Yata)
Institute of Mathematics
University of Tsukuba

筑波大学・数理物質系　青嶋　誠 (Makoto Aoshima)
Institute of Mathematics
University of Tsukuba

1. はじめに

近代科学におけるデータの一つの特徴は，データがもつ次元数の膨大さにある．例えば，DNAマイクロアレイデータや画像データは，次元数が優に10,000を超える．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，いわゆる頑健性の研究があった．しかしながら，そこでは，次元数は標本数よりも小さいことが仮定され，標本数よりも次元数の方が圧倒的に大きな，本来の高次元データ（高次元小標本データ）を扱うことはできなかった．従来の多変量統計解析の手法にとって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ができないのである．これは，従来の多変量解析法の一致解が得られない現象として理論的に定式化され，一致解と不一致解の境界が Yata and Aoshima (2009) によって与えられた．

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，多変量統計解析の枠組みを外す必要がある．なぜなら，多変量統計解析の枠組みに囚われた研究は，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，理論物理と確率論の立場から，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．Johnstone (2001)，Baik and Silverstein (2006)，Paul (2007) 等による，標本固有値の漸近分布の導出である．しかしながら，これらの論文は，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，母集団にはガウス分布もしくは類する条件を仮定していた．次元数は優に10,000を超えるが標本数は高々100程度といった高次元小標本においては，標本数を次元数と同程度には確保できない．それゆえ，n が p に依存しないような設定で，高次元漸近理論を展開する必要があった．Yata

and Aoshima (2010) 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 (2010) は，クロスデータ行列の特異値分解を使っ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の推定も与えて，それらが高次元小標本の設定で一致解を与えることを証明した．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．Hall et al. (2005)，Ahn et al. (2007)，Yata and Aoshima (2012a) 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．Hall et al. (2005) や Ahn et al. (2007) は母集団をガウス分布もしくは類するものに限定した場合を扱った．それに対し，Yata and Aoshima (2012a) は分布の限定を取り去って非ガウス分布の場合を扱い，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幾何学的表現を発見した．非ガウス性に関するある尺度を境とした，球面集中現象と座標軸への退化現象という，2つの特徴的な幾何学的表現である．これらの幾何学的表現に基づいて，Aoshima and Yata (2011a,b) は，高次元小標本における統計的推測に，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と方法論を築き上げた．Aoshima and Yata (2011a,b) の研究は多岐に亘り，高次元球面上の与えられたバンド幅の信頼領域，高次元の二標本問題，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，変数選択問題，パスウエイ解析など，高次元小標本データの様々な統計的推測について，高次元データが本来もつ豊富な情報を生かすための理論と方法論を与えている．

一方で，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を与えるための研究がある．高次元データへのアプローチは，数理統計学の立場とは異なり，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 (1995) 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，汎化性能が良いことも知られているが，理論的な精度を保証するものではない．

本論文では，2群における高次元データの判別分析と変数選択について，高速な算法でありながら汎化性能に優れ，さらに，高い精度を保証することできる方法論を理論的に考える．その際に，高次元小標本においては標本共分散行列の逆行列が存在しないため，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できる場合，Bickel and Levina (2004) は，標本共分散行列の対角成分だけを使った逆行列で代替する判別方式を与えた．それに対して，Yata and Aoshima (2012a) は，ノイズ掃き出し法とよばれるセミパラメトリックな推定法を提案し，これを用いたリッジ型逆行列による判別方式が Bickel and Levina (2004) 等の判別精度に優ることを示した．しかしながら，2群の共分

散行列が等しいという仮定は，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となり，必ずしも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. (2002) による標本共分散行列の対角成分を使った判別方式や，Hall et al. (2005, 2008)，Chan and Hall (2009)，Yata and Aoshima (2012b) 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 (2011a,b) が与えた高次元小標本の幾何学的表現に基づく判別方式がある．なかでも，Aoshima and Yata (2011a,b) による判別方式は，高次元データが有する2群の共分散行列の差異を有効に利用した方法論であり，その統計的推測は注目に値する．本論文では，まず，Aoshima and Yata (2011a,b) の理論と方法論を解説し，それらの拡張を考えることで，高次元データの判別分析と変数選択に一連の統計的推測を与える．

2. Aoshima and Yata (2011a) の判別方式

2群の判別を考える．2群 π_1, π_2 には，平均ベクトル $\boldsymbol{\mu}_i = (\mu_{i1}, \cdots, \mu_{ip})^T$, $i = 1, 2$ と共分散行列 $\boldsymbol{\Sigma}_i$ $(> \boldsymbol{O})$, $i = 1, 2$ を仮定し，さらに，$||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2/p < \infty$, $p \to \infty$ を仮定する．$\boldsymbol{\Sigma}_i$ の対角成分を $\sigma^2_{(i)j}$, $j = 1, ..., p$ で表し，対角成分が次元数 p に依らないこと，および，$0 < \sigma^2_{(i)j} < \infty$, $j = 1, ..., p$ であることを仮定する．$\boldsymbol{\Sigma}_i$ の固有値を $\lambda_{i1} \geq \cdots \geq \lambda_{ip} > 0$ と表し，適当な直交行列 $\boldsymbol{H}_i = [\boldsymbol{h}_{i1}, \cdots, \boldsymbol{h}_{ip}]$ で $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i \boldsymbol{H}_i^T$, $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1}, \cdots, \lambda_{ip})$ と分解する．さらに，$\boldsymbol{z}_{il} = (z_{i1l}, \cdots, z_{ipl})^T = \boldsymbol{\Lambda}_i^{-1/2} \boldsymbol{H}_i^T (\boldsymbol{x}_{il} - \boldsymbol{\mu}_i)$ を定義し，$E(z^4_{ijl}) < \infty$, $j = 1, ..., p$ $(l = 1, ..., n_i)$ を仮定する．

2群 π_1, π_2 の分布には，次の3つの正則条件のどれか一つを仮定する:

(A-i) $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$;

(A-ii) 各 i で, $E(z^2_{ijl} z^2_{ij'l}) = 1$, $j \neq j' = 1, ..., p$;

(A-iii) 各 i で, $E[\exp\{t|x_{ijl} - \mu_{ij}|/\sigma_{(i)j}\}] < \infty$, $j = 1, ..., p$ なるある t (> 0) が存在する.

ここで，(A-ii), (A-iii) は，(A-i) を緩めた条件であることに注意する．

ラベルなしデータを $\boldsymbol{x}_0$ とする．π_i, $i = 1, 2$ から n_i (≥ 2) 個のトレーニングデータ $\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i} (\in R^p)$ を抽出し，$\overline{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$, $\boldsymbol{S}_{in_i} = \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})^T/(n_i - 1)$ を計算する．Aoshima and Yata (2011a) は，高次元小標本データの幾何学的表現に基づいて2次モーメントを利用し，次のような判別関数を与えた：

$$\omega(\boldsymbol{x}_0) = \frac{p||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}||^2}{\mathrm{tr}(\boldsymbol{S}_{1n_1})} - \frac{p||\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}||^2}{\mathrm{tr}(\boldsymbol{S}_{2n_2})} - p \log \left\{ \frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{\mathrm{tr}(\boldsymbol{S}_{1n_1})} \right\} - \frac{p}{n_1} + \frac{p}{n_2}. \quad (2.1)$$

そのとき，Aoshima and Yata (2011a) の判別方式は，$\omega(\boldsymbol{x}_0) < 0$ のとき $\boldsymbol{x}_0 \in \pi_1$, $\omega(\boldsymbol{x}_0) \geq 0$ のとき $\boldsymbol{x}_0 \in \pi_2$，というものである．

いま，$\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$, $\Delta_{\boldsymbol{\Sigma}_i} = \{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}^2/\mathrm{tr}(\boldsymbol{\Sigma}_i)$, $i = 1, 2$ とおき，$\Delta_i = \Delta + \Delta_{\boldsymbol{\Sigma}_i}/2$, $i = 1, 2$ を定義する．このとき，Aoshima and Yata (2011a) は次の定理を与えた．

定理 2.1 (Aoshima and Yata, 2011a). 適当な正則条件のもと，$p \to \infty$, $n_1 \to \infty$, $n_2 \to \infty$ のとき次が成り立つ．

$$\frac{\omega(\boldsymbol{x}_0) + \Delta_2(\mathrm{tr}(\boldsymbol{\Sigma}_2)/p)^{-1}}{2\sqrt{(\mathrm{tr}(\boldsymbol{\Sigma}_1)/p)^{-2}\mathrm{tr}(\boldsymbol{\Sigma}_1^2)/n_1 + (\mathrm{tr}(\boldsymbol{\Sigma}_2)/p)^{-2}\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_2}} \Rightarrow N(0,1) \quad (\boldsymbol{x}_0 \in \pi_1),$$

$$\frac{\omega(\boldsymbol{x}_0) - \Delta_1(\mathrm{tr}(\boldsymbol{\Sigma}_1)/p)^{-1}}{2\sqrt{(\mathrm{tr}(\boldsymbol{\Sigma}_2)/p)^{-2}\mathrm{tr}(\boldsymbol{\Sigma}_2^2)/n_2 + (\mathrm{tr}(\boldsymbol{\Sigma}_1)/p)^{-2}\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)/n_1}} \Rightarrow N(0,1) \quad (\boldsymbol{x}_0 \in \pi_2).$$

ここで，$\Rightarrow$ は分布収束を意味する．

注意 1. Aoshima and Yata (2011a) は，定理 2.1 を用いて，誤判別確率に関する要求精度を満たすための判別方式を提案している．

3. Aoshima and Yata (2011a) の判別方式の漸近的性質

本節では，2 節で紹介した Aoshima and Yata (2011a) の判別方式について，誤判別確率に関する一致性を研究する．いま，$\Delta_\star = \min_{i=1,2} \Delta_i$ とおいて，次の条件を仮定する:

(A-iv) 各 i で，$p \to \infty$ のとき，

$$\frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_i(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\Delta_\star^2} \to 0 \quad \text{かつ} \quad \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i)^2\Delta_\star^2} \to 0;$$

(A-v) 各 i で，$p \to \infty$ のとき，

n_i は固定もしくは $n_i \to \infty$ のもと，$\dfrac{\max_{j=1,2}\mathrm{tr}(\boldsymbol{\Sigma}_j^2)}{n_i\Delta_\star^2} \to 0.$

そのとき，π_1 の $\boldsymbol{x}_0$ を π_2 に誤判別する確率 $e(2|1)$，π_2 の $\boldsymbol{x}_0$ を π_1 に誤判別する確率 $e(1|2)$ について，次の定理が成り立つ．

定理 3.1. 2 群 π_1, π_2 に，(A-ii) を仮定する．さらに，(A-iv)-(A-v) を仮定する．$p \to \infty$ のとき，次が成り立つ．

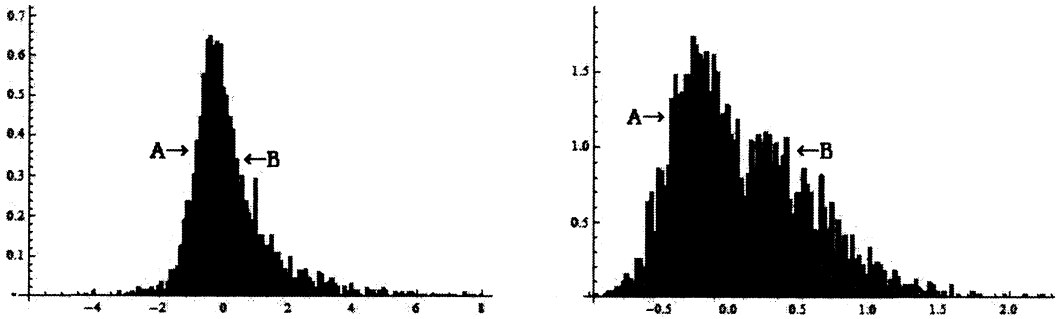
$$e(2|1) \to 0 \quad \text{かつ} \quad e(1|2) \to 0.$$

注意 2. もしも

$$\frac{\max_{j=1,2}\{\mathrm{tr}(\boldsymbol{\Sigma}_j)\}\max_{j=1,2}\{\mathrm{tr}(\boldsymbol{\Sigma}_j^2)^{1/2}\}}{\{\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)\}^2 n_i} = o(1), \ i = 1, 2$$

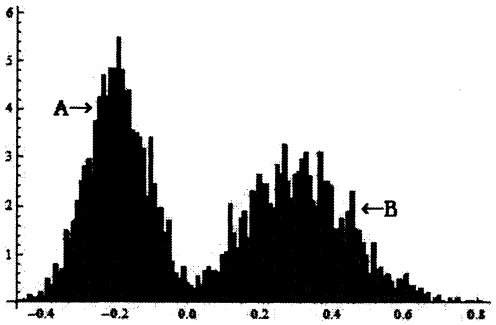
ならば，仮定 (A-v) は必然的に満たされる．例えば，ある定数 c $(\neq 1)$ に対して $\mathrm{tr}(\boldsymbol{\Sigma}_1) = \mathrm{tr}(c\boldsymbol{\Sigma}_2)$ かつ $\max_{j=1,2} \mathrm{tr}(\boldsymbol{\Sigma}_j^2) = o(p^2)$ が成り立つならば，仮に $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ であっても，もしくは，n_1, n_2 が固定であっても，(A-v) は必然的に満たされる．

$\omega(\boldsymbol{x}_0)$ の漸近的性質をシミュレーション実験で検証した．2群は π_i : $N_p(0, \boldsymbol{\Sigma}_i)$, $i = 1, 2$ とし，$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ なるモデルを考えた．平均間の距離のみに基づく判別手法では，このモデルを識別することはできない．いま，$n_1 = n_2 = 5$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_2 = 2\boldsymbol{I}_p$ と設定し，$p = 4$, 32, 256, 2048 としてシミュレーション実験をおこなった．ここで，この設定が仮定 (A-ii), (A-iv), (A-v) を満たすことに注意する．図1は，A : $\boldsymbol{x}_0 \in \pi_1$ と B : $\boldsymbol{x}_0 \in \pi_2$ のそれぞれの場合について，2000回のシミュレーションによる $\omega(\boldsymbol{x}_0)$ のヒストグラムを与えている．定理3.1で主張した通り，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．

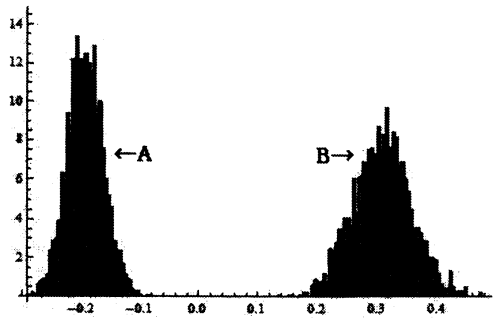


(a) $p = 4$

(b) $p = 32$



(c) $p = 256$



(d) $p = 2048$

図 1. $\pi_1 : N_p(\boldsymbol{0}, \boldsymbol{I}_p)$, $\pi_2 : N_p(\boldsymbol{0}, 2\boldsymbol{I}_p)$ のとき，$A : \boldsymbol{x}_0 \in \pi_1$ と $B : \boldsymbol{x}_0 \in \pi_2$ に対する $\omega(\boldsymbol{x}_0)$ のヒストグラム．

4. 変数選択に基づく判別方式

本節では，Fan and Fan (2008) や Aoshima and Yata (2011a) でも議論された，変数選択に基づく判別方式について考える．次の検定から始める．

$$H_{0j} : \mu_{1j} = \mu_{2j} \quad \text{vs.} \quad H_{1j} : \mu_{1j} \neq \mu_{2j} \quad \text{for } j = 1, ..., p. \tag{4.1}$$

ここで，対立仮説が正しい変数の集合を $\boldsymbol{D} = \{j : \mu_{1j} \neq \mu_{2j}\}$ とおく．ただし，$\boldsymbol{D} \neq \emptyset$ と仮定する．この検定に対して，Aoshima and Yata (2011a) は，FWER (Family-wise error rate) を調整するための標本数の決定方式を与え，2段階推定法に基づく変数選択法を提案した．本節では，$\delta\ (>0)$ を設定し，$p \to \infty$ のとき

$$P(\boldsymbol{D} = \widehat{\boldsymbol{D}}) \to 1 \quad \text{when } \min_{j \in \boldsymbol{D}} |\mu_{1j} - \mu_{2j}| \geq \delta \tag{4.2}$$

となるような変数の集合 $\widehat{\boldsymbol{D}}$ を構築し，$\widehat{\boldsymbol{D}}$ に属する変数に基づいて判別方式を構築することを考える．

4.1. 高次元データの変数選択

いま，$\sigma_i^2 = \max_{j=1,\ldots,p} \sigma_{(i)j}^2$，$i = 1, 2$ とおき，$T_{j(\mathbf{n})} = \overline{x}_{1jn_1} - \overline{x}_{2jn_2}$，$j = 1, \ldots, p$ とおく．ただし，$\boldsymbol{n} = (n_1, n_2)$，$\overline{x}_{ijn_i} = \sum_{l=1}^{n_i} x_{ijl}/n_i$，$i = 1, 2$ である．もしも，適当な $\eta(\geq 0)$ に対して

$$\sum_{i=1}^{2} \frac{\sigma_i^2}{n_i} \leq \frac{\delta^2}{8(\log p)^{1+\eta}} \tag{4.3}$$

を満たすように各母集団の標本数を決定し，各 $j\ (= 1, \ldots, p)$ に対する検定方式

$$H_{0j} \text{ を棄却} \Longleftrightarrow |T_{j(\mathbf{n})}| > \delta/2 \tag{4.4}$$

の結果から $\widehat{\boldsymbol{D}} = \{j : H_{0j} \text{ を棄却}\}$ を定義すれば，そのとき次の定理を得る．証明は，Aoshima and Yata (2011a) の5節と同様の手順を辿る．

定理 4.1. 2群 π_1, π_2 に，(A-i) もしくは (A-iii) を仮定する．(A-i) を仮定するとき，n_1, n_2 は $\eta \geq 0$ で (4.3) を満たすものとする．(A-iii) を仮定するとき，n_1, n_2 は $\eta > 0$ で (4.3) を満たすものとする．そのとき，$p \to \infty$ のもと，検定方式 (4.4) は (4.2) が成り立つ．

注意 3. (4.3) を満たす標本数は，$n_i/p \to 0$，$p \to \infty$ という高次元小標本の枠組みに存在する．そのとき，標本数を

$$n_i \geq \frac{8(\log p)^{1+\eta}}{\delta^2} \sigma_i \sum_{j=1}^{2} \sigma_j \tag{4.5}$$

なるように定めれば，最小の総標本数 $(n_1 + n_2)$ で (4.2) を満足する検定方式 (4.4) を構築することができる．

いま，集合 $\boldsymbol{D}$ の要素の数 $|\boldsymbol{D}|$ について，$|\boldsymbol{D}| \leq s$ かつ $p \to \infty$ のとき $s \to \infty$ となるような，p 以下の正の整数 s が存在すると仮定する．もしも，適当な $\eta(\geq 0)$

に対して

$$\sum_{i=1}^{2}\frac{\sigma_i^2}{n_i}\leq\frac{\delta^2}{2(\sqrt{\log p}+\sqrt{\log s}\,)^2(\log p)^{\eta}} \tag{4.6}$$

を満たすように各母集団の標本数を決定し，各 $j\ (=1,...,p)$ に対する検定方式

$$H_{0j}\text{を棄却}\Longleftrightarrow|T_{j(\mathbf{n})}|>\frac{\delta\sqrt{\log p}}{\sqrt{\log p}+\sqrt{\log s}} \tag{4.7}$$

の結果から $\widehat{\boldsymbol{D}}=\{j:\ H_{0j}\text{を棄却}\}$ を定義すれば，そのとき次の系を得る．

系 4.1. 2群 π_1,π_2 に，(A-i) もしくは (A-iii) を仮定する．(A-i) を仮定するとき，n_1,n_2 は $\eta\geq 0$ で (4.6) を満たすものとする．(A-iii) を仮定するとき，n_1,n_2 は $\eta>0$ で (4.6) を満たすものとする．そのとき，$p\to\infty$ のもと，検定方式 (4.7) は (4.2) が成り立つ．

注意 4. (4.6) を満たす標本数を

$$n_i\geq\frac{2(\sqrt{\log p}+\sqrt{\log s}\,)^2(\log p)^{\eta}}{\delta^2}\sigma_i\sum_{j=1}^{2}\sigma_j$$

なるように定めれば，最小の総標本数 (n_1+n_2) で (4.2) を満足する検定方式 (4.7) を構築することができる．

注意 5. Aoshima and Yata (2011a) では，基準化した平均間の距離 $|\mu_{1j}-\mu_{2j}|/\sqrt{\sigma_{(1)j}^2+\sigma_{(2)j}^2}$ をもとに変数選択法を考え，それに基づく判別方式を提案した．

4.2. 変数選択に基づく判別方式

本節では，4.1 節で変数選択した $\widehat{\boldsymbol{D}}$ の要素を用いて判別方式を考える．いま，$\min_{j\in\boldsymbol{D}}|\mu_{1j}-\mu_{2j}|\geq\delta$ と仮定する．$|\boldsymbol{D}|=S$，$|\widehat{\boldsymbol{D}}|=\widehat{S}$ とおく．そのとき，$\boldsymbol{x}_{1j(S)}$，$\boldsymbol{x}_{2j(S)}$，$\boldsymbol{x}_{0(S)}$ を，$\boldsymbol{x}_{1j}$，$\boldsymbol{x}_{2j}$，$\boldsymbol{x}_0$ の $\boldsymbol{D}$ に含まれる変数だけを並べた S 次元ベクトルとする．同様に，$\boldsymbol{x}_{1j(\widehat{S})}$，$\boldsymbol{x}_{2j(\widehat{S})}$，$\boldsymbol{x}_{0(\widehat{S})}$ を，$\widehat{\boldsymbol{D}}$ に含まれる変数だけを並べた $\widehat{S}$ 次元ベクトルとする．ここで，各 i で $E(\boldsymbol{x}_{ij(S)})=\boldsymbol{\mu}_{i(S)}$，$Var(\boldsymbol{x}_{ij(S)})=\boldsymbol{\Sigma}_{i(S)}$ とおく．そのとき，次の条件を仮定する：

(A-vi) $p\to\infty$ のとき，$S\to\infty$ かつ $\dfrac{(\boldsymbol{\mu}_{1(S)}-\boldsymbol{\mu}_{2(S)})^T\boldsymbol{\Sigma}_{i(S)}(\boldsymbol{\mu}_{1(S)}-\boldsymbol{\mu}_{2(S)})}{||\boldsymbol{\mu}_{1(S)}-\boldsymbol{\mu}_{2(S)}||^4}\to 0,$ $i=1,2.$

いま，$\overline{\boldsymbol{x}}_{in_i(\widehat{S})}=\sum_{j=1}^{n_i}\boldsymbol{x}_{ij(\widehat{S})}/n_i$，$\boldsymbol{S}_{in_i(\widehat{S})}=\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij(\widehat{S})}-\overline{\boldsymbol{x}}_{in_i(\widehat{S})})(\boldsymbol{x}_{ij(\widehat{S})}-\overline{\boldsymbol{x}}_{in_i(\widehat{S})})^T/(n_i-1)$ とし，Aoshima and Yata (2011a) で与えられた次の判別方式を考える．

$$\Big(\boldsymbol{x}_{0(\widehat{S})}-\frac{\overline{\boldsymbol{x}}_{1n_1(\widehat{S})}+\overline{\boldsymbol{x}}_{2n_2(\widehat{S})}}{2}\Big)^T(\overline{\boldsymbol{x}}_{2n_2(\widehat{S})}-\overline{\boldsymbol{x}}_{1n_1(\widehat{S})})-\frac{\mathrm{tr}(\boldsymbol{S}_{1n_1(\widehat{S})})}{2n_1}+\frac{\mathrm{tr}(\boldsymbol{S}_{2n_2(\widehat{S})})}{2n_2}<0 \tag{4.8}$$

のとき $\boldsymbol{x}_0 \in \pi_1$，それ以外のとき $\boldsymbol{x}_0 \in \pi_2$.

上記の判別方式は，Aoshima and Yata (2011a) とは異なる基準で選択された変数に基づいている．そのとき，次の定理を得る．

定理 4.2. $\widehat{\boldsymbol{D}}$ は定理 4.1 もしくは系 4.1 を満たすと仮定する．(A-vi) を仮定する．そのとき，$p \to \infty$ のもと，判別方式 (4.8) は次が成り立つ．

$$e(2|1) \to 0 \quad \text{かつ} \quad e(1|2) \to 0.$$

4.3. シミュレーション

定理4.1と定理4.2をシミュレーションで検証する．2群は $\pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$ とし，$\boldsymbol{\mu}_1 = (0, ..., 0)^T$ とおき，$\boldsymbol{\mu}_2$ は最初と最後の 15 個の成分を 2 とした $\boldsymbol{\mu}_2 = (2, ..., 2, 0, ..., 0, 2, ..., 2)^T$ とおいた．そのとき，$\boldsymbol{D} = \{1, ..., 15, p-14, ..., p\}$, $|\boldsymbol{D}| = S = 30$ となる．また，$\boldsymbol{\Sigma}_1 = \boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$, $\boldsymbol{\Sigma}_2 = 1.2\boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$ とし，$\boldsymbol{B} = \mathrm{diag}\left(\sqrt{0.5 + 1/(p+1)}, \sqrt{0.5 + 2/(p+1)}, ..., \sqrt{0.5 + p/(p+1)}\right)$ とおいた．次元数は $p = 50$, 200, 800 の 3 つの場合を考え，$\delta = 2$ と設定した．標本数は定理 4.1 と (4.5) に基づき，

$$n_i = \left[\frac{8 \log p}{\delta^2} \sigma_i \sum_{j=1}^{2} \sigma_j\right] + 1$$

とした．ただし，$[x]$ は x を超えない最大の整数を表す．そのとき，各 p の場合で，(4.4) の検定をおこない $\widehat{\boldsymbol{D}}$ を求め，それに基づいて判別方式 (4.8) を構築し，$\boldsymbol{x}_0 \in \pi_1$ と $\boldsymbol{x}_0 \in \pi_2$ のそれぞれの状況で正しく判別されるかを確認した．これを独立に 2000 回繰り返し，結果を纏めたものが表 1 である．$P(\widehat{\boldsymbol{D}} = \boldsymbol{D})$ は $\widehat{\boldsymbol{D}} = \boldsymbol{D}$ と正しく変数選択した割合を表し，$s(P(\widehat{\boldsymbol{D}} = \boldsymbol{D}))$ はその標準偏差を表す．また，$\overline{e(2|1)}$, $\overline{e(1|2)}$ は誤判別の割合を表し，$s(\overline{e(2|1)})$, $s(\overline{e(1|2)})$ はその標準偏差を表す．表 1 から，変数選択と判別方式がともに十分機能していることが見てとれる．ここでは割愛するが，設定を変えて実験をしたときにも，同様の結果が得られた．

5. マイクロアレイデータ解析

本節では，4 節で与えた変数選択に基づく判別方式を用いて，Chiaretti et al. (2004) の 12,625(= p) 遺伝子からなるマイクロアレイデータの解析を試みる．このマイクロアレイデータは，π_1: B-cell (95 サンプル) と π_2: T-cell (33 サンプル) の，2 つのタイプの腫瘍のデータからなる．まず，$\delta = 0.8$ と設定した．この設定の妥当性は，矢田 (2011) を参照のこと．次に，$n_1 = 50$, $n_2 = 25$ と設定し，これらが (4.3) を満たすと仮定した．そのとき，検定方式 (4.4) により，

$$\widehat{\boldsymbol{D}} = \{6, 8, 42, ..., 12589, 12593, 12600\}, \quad |\widehat{\boldsymbol{D}}| = \widehat{S} = 1705$$

表 1. $p = 50, 200, 800$ における変数選択法 (4.3)-(4.4) と判別方式 (4.8) の性能.

$\overline{P(\widehat{\boldsymbol{D}} = \boldsymbol{D})}$	$s(\overline{P(\widehat{\boldsymbol{D}} = \boldsymbol{D})})$	$\overline{e(2\|1)}$	$s(\overline{e(2\|1)})$	$\overline{e(1\|2)}$	$s(\overline{e(1\|2)})$
	$p = 50$:	$(n_1, n_2) = (25, 27)$			
0.971	0.00378	0.003	0.00112	0.003	0.00112
	$p = 200$:	$(n_1, n_2) = (34, 37)$			
0.972	0.00369	0.002	0.00087	0.002	0.00087
	$p = 800$:	$(n_1, n_2) = (42, 46)$			
0.966	0.00408	0.001	0.00071	0.002	0.00087

を得た．ここで，$\widehat{\boldsymbol{D}}$ で選択された変数だけを用いて判別方式 (4.8) を定義し，残りの $\pi_1 : 95 - n_1 = 45$ サンプルと $\pi_2 : 33 - n_2 = 8$ サンプルをテストデータとした．その結果，正判別した割合は $\pi_1 : 45/45 = 1$, $\pi_2 : 8/8 = 1$ となり，定理 4.2 の主張の通り，すべてが正しく判別された．なお，2 節で紹介した判別方式を用いた場合の解析例については，Aoshima and Yata (2011a,b,c) を参照のこと．

Appendix

定理 3.1 の証明. まず，$\boldsymbol{x}_0 \in \pi_1$ の場合を考える．いま，$Var\{||\overline{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i||^2 - \mathrm{tr}(\boldsymbol{S}_{in_i})/n_i\} = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/n_i^2\}$, $Var\{(\overline{\boldsymbol{x}}_{in_i} - \boldsymbol{\mu}_i)^T(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)\} = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_1)/n_i\} = O\{\max_{j=1,2} \mathrm{tr}(\boldsymbol{\Sigma}_j^2)/n_i\}$ がいえる．さらに，(A-ii) のもと，$Var\{\mathrm{tr}(\boldsymbol{S}_{in_i})\} = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/n_i\}$ である．ここで，(A-iv)-(A-v) のもとでチェビシェフの不等式を用いると，

$$
\begin{aligned}
\frac{\omega(\boldsymbol{x}_0)}{p} =& \frac{||\boldsymbol{x}_0 - \boldsymbol{\mu}_1 - (\overline{\boldsymbol{x}}_{1n_1} - \boldsymbol{\mu}_1)||^2 - \mathrm{tr}(\boldsymbol{S}_{1n_1})/n_1}{\mathrm{tr}(\boldsymbol{S}_{1n_1})} - \log\left\{\frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{\mathrm{tr}(\boldsymbol{S}_{1n_1})}\right\} \\
& - \frac{||\boldsymbol{x}_0 - \boldsymbol{\mu}_1 - (\overline{\boldsymbol{x}}_{2n_2} - \boldsymbol{\mu}_2) + \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2 - \mathrm{tr}(\boldsymbol{S}_{2n_2})/n_2}{\mathrm{tr}(\boldsymbol{S}_{2n_2})} \\
=& \frac{||\boldsymbol{x}_0 - \boldsymbol{\mu}_1||^2 + o_p(\Delta_\star)}{\mathrm{tr}(\boldsymbol{\Sigma}_1) + o_p(\Delta_\star)} - \frac{||\boldsymbol{x}_0 - \boldsymbol{\mu}_1||^2 + ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2 + o_p(\Delta_\star)}{\mathrm{tr}(\boldsymbol{\Sigma}_2) + o_p(\Delta_\star)} \\
& - \log\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_2) + o_p(\Delta_\star)}{\mathrm{tr}(\boldsymbol{\Sigma}_1) + o_p(\Delta_\star)}\right\} \\
=& \frac{||\boldsymbol{x}_0 - \boldsymbol{\mu}_1||^2\{\mathrm{tr}(\boldsymbol{\Sigma}_2) - \mathrm{tr}(\boldsymbol{\Sigma}_1)\}}{\mathrm{tr}(\boldsymbol{\Sigma}_1)\mathrm{tr}(\boldsymbol{\Sigma}_2)} - \frac{||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2}{\mathrm{tr}(\boldsymbol{\Sigma}_2)} \\
& - \log\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_2)}{\mathrm{tr}(\boldsymbol{\Sigma}_1)}\right\} + o_p\{\Delta_\star/\mathrm{tr}(\boldsymbol{\Sigma}_1)\} + o_p\{\Delta_\star/\mathrm{tr}(\boldsymbol{\Sigma}_2)\} \qquad \text{(A.1)}
\end{aligned}
$$

と評価できる．ここで，$E(||\boldsymbol{x}_0 - \boldsymbol{\mu}_1||^2) = \mathrm{tr}(\boldsymbol{\Sigma}_1)$, $Var(||\boldsymbol{x}_0 - \boldsymbol{\mu}_1||^2) = O\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\}$

に注意する．(A.1) より，$\mathrm{tr}(\Sigma_1)/\mathrm{tr}(\Sigma_2) \to 1$, $p \to \infty$ のもと

$$\frac{\omega(\boldsymbol{x}_0)\mathrm{tr}(\Sigma_2)}{p\Delta_\star} = \frac{\mathrm{tr}(\Sigma_2)-\mathrm{tr}(\Sigma_1)}{\Delta_\star} - \frac{||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||^2}{\Delta_\star} + \frac{\mathrm{tr}(\Sigma_2)}{\Delta_\star}\log\left\{\frac{\mathrm{tr}(\Sigma_1)}{\mathrm{tr}(\Sigma_2)}\right\} + o_p(1) \tag{A.2}$$

$$\begin{aligned}
&= \frac{\mathrm{tr}(\Sigma_2)-\mathrm{tr}(\Sigma_1)}{\Delta_\star} - \frac{||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||^2}{\Delta_\star} + o_p(1) \\
&\quad + \frac{\mathrm{tr}(\Sigma_2)}{\Delta_\star}\left[\frac{\mathrm{tr}(\Sigma_1)-\mathrm{tr}(\Sigma_2)}{\mathrm{tr}(\Sigma_2)} - \frac{\{\mathrm{tr}(\Sigma_1)-\mathrm{tr}(\Sigma_2)\}^2}{2\mathrm{tr}(\Sigma_2)^2}\{1+o(1)\}\right] \\
&= \frac{-\Delta_2}{\Delta_\star} + o_p(1) < 0 \quad \text{w.p.1}
\end{aligned}$$

となる．一方で，$\mathrm{tr}(\Sigma_1)/\mathrm{tr}(\Sigma_2) \neq 1$, $p \to \infty$ のとき，$\Delta_\star/\mathrm{tr}(\Sigma_2) = O(1)$ と

$$1 - \frac{\mathrm{tr}(\Sigma_1)}{\mathrm{tr}(\Sigma_2)} + \log\left\{\frac{\mathrm{tr}(\Sigma_1)}{\mathrm{tr}(\Sigma_2)}\right\} < 0$$

に注意する．そのとき，(A.2) より

$$\frac{\omega(\boldsymbol{x}_0)}{p} = \frac{\mathrm{tr}(\Sigma_2)-\mathrm{tr}(\Sigma_1)}{\mathrm{tr}(\Sigma_2)} - \frac{||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||^2}{\mathrm{tr}(\Sigma_2)} + \log\left\{\frac{\mathrm{tr}(\Sigma_1)}{\mathrm{tr}(\Sigma_2)}\right\} + o_p(1) < 0 \quad \text{w.p.1}$$

が得られる．同様の漸近的評価を，$\boldsymbol{x}_0 \in \pi_2$ の場合にも与えることができる． □

定理 4.1 の証明． まず，(A-iii) を仮定する．そのとき，Aoshima and Yata (2011a) の定理 5.1 の証明にある式 (A.14) より，(4.3) のもと

$$P(|T_{j(\mathbf{n})} - (\mu_{1j}-\mu_{2j})| > \delta/2) = o(p^{-1})$$

が主張できるので，ボンフェロニの不等式から結果が得られる．次に，(A-i) を仮定する．任意の $x > 0$ に対して

$$\int_x^\infty \frac{e^{-t^2/2}}{\sqrt{2\pi}}dt \leq 2e^{-x^2/2}/(1+x)$$

なることに注意する．そのとき，(4.3) のもと

$$\begin{aligned}
&P(|T_{j(\mathbf{n})} - (\mu_{1j}-\mu_{2j})| > \delta/2) \\
&= P\{(\sigma_{(1)j}^2/n_1 + \sigma_{(2)j}^2/n_2)^{-1/2}|T_{j(\mathbf{n})} - (\mu_{1j}-\mu_{2j})| > (\sigma_{(1)j}^2/n_1 + \sigma_{(2)j}^2/n_2)^{-1/2}\delta/2\} \\
&\leq P(|N(0,1)| > \sqrt{2\log p}) \leq 4e^{-\log p}/(1+\sqrt{2\log p}) = o(1/p)
\end{aligned}$$

が主張できるので，ボンフェロニの不等式から結果が得られる． □

系 4.1 の証明． (4.6) より，

$$(\sigma_{(1)j}^2/n_1 + \sigma_{(2)j}^2/n_2)^{-1/2}\delta\frac{\sqrt{\log p}}{\sqrt{\log p}+\sqrt{\log s}} \geq (\log p)^{\eta/2}\sqrt{2\log p}.$$

さらに,

$$(\sigma_{(1)j}^2/n_1+\sigma_{(2)j}^2/n_2)^{-1/2}\Big\{\delta-\delta\frac{\sqrt{\log p}}{\sqrt{\log p}+\sqrt{\log s}}\Big\}\geq(\log p)^{\eta/2}\sqrt{2\log s}$$

も主張できるので，定理 4.1 の証明と同様の手順で結果が得られる． □

定理 4.2 の証明． 仮定から，$P(\boldsymbol{D}=\widehat{\boldsymbol{D}})\to 1,\ p\to\infty$ が成り立つので,

$$\begin{aligned}1-e(2|1)=&P\Big\{\Big(\boldsymbol{x}_{0(\widehat{S})}-\frac{\overline{\boldsymbol{x}}_{1n_1(\widehat{S})}+\overline{\boldsymbol{x}}_{2n_2(\widehat{S})}}{2}\Big)^T(\overline{\boldsymbol{x}}_{2n_2(\widehat{S})}-\overline{\boldsymbol{x}}_{1n_1(\widehat{S})})-\frac{\mathrm{tr}(\boldsymbol{S}_{1n_1(\widehat{S})})}{2n_1}\\&+\frac{\mathrm{tr}(\boldsymbol{S}_{2n_2(\widehat{S})})}{2n_2}<0\Big\}\\=&P\Big\{\Big(\boldsymbol{x}_{0(S)}-\frac{\overline{\boldsymbol{x}}_{1n_1(S)}+\overline{\boldsymbol{x}}_{2n_2(S)}}{2}\Big)^T(\overline{\boldsymbol{x}}_{2n_2(S)}-\overline{\boldsymbol{x}}_{1n_1(S)})-\frac{\mathrm{tr}(\boldsymbol{S}_{1n_1(S)})}{2n_1}\\&+\frac{\mathrm{tr}(\boldsymbol{S}_{2n_2(S)})}{2n_2}<0\Big\}+o(1)\end{aligned}\tag{A.3}$$

が主張できる．ここで，$\overline{\boldsymbol{x}}_{in_i(S)}=\sum_{j=1}^{n_i}\boldsymbol{x}_{ij(S)}/n_i$，$\boldsymbol{S}_{in_i(S)}=\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij(S)}-\overline{\boldsymbol{x}}_{in_i(S)})(\boldsymbol{x}_{ij(S)}-\overline{\boldsymbol{x}}_{in_i(S)})^T/(n_i-1)$ である．いま，仮定 $\min_{j\in\boldsymbol{D}}|\mu_{1j}-\mu_{2j}|\geq\delta$ より，$||\boldsymbol{\mu}_{1(S)}-\boldsymbol{\mu}_{2(S)}||^2/S\geq\delta^2>0$ となる．また，$\max_{j=1,2}\mathrm{tr}(\boldsymbol{\Sigma}_{j(S)}^2)=O(S^2)$，$n_1,n_2\to\infty$，$p\to\infty$ なので,

$$\frac{\max_{j=1,2}\mathrm{tr}(\boldsymbol{\Sigma}_{j(S)}^2)}{||\boldsymbol{\mu}_{1(S)}-\boldsymbol{\mu}_{2(S)}||^4n_i}\to 0,\ p\to\infty\tag{A.4}$$

が主張できる．このとき，Aoshima and Yata (2011a) の定理 7.2 から，仮定 (A-vi) と (A.4) のもと，$\boldsymbol{x}_0\in\pi_1$ の場合に

$$e(2|1)\to 0$$

が主張できる．同様に，$\boldsymbol{x}_0\in\pi_2$ の場合も結果が得られる． □

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

参考文献

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

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

Aoshima, M. and Yata, K. (2011b). Author's response. *Sequential Anal.* **30**, 432-440.

Aoshima, M. and Yata, K. (2011c). Effective methodologies for statistical inference on microarray studies. *In: P.E. Spiess (Ed.), Prostate Cancer - From Bench to Bedside*, InTech, 13-32.

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

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

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

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J. and Foa, R. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood* **103**, 2771-2778.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Statist. Assoc.* **97**, 77-87.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36**, 2605-2637.

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

Hall, P., Pittelkow, Y. and Ghosh, M. (2008). Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *J. R. Statist. Soc.* B **70**, 159-173.

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

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

Vapnic, V.N. (1995). *The Nature of Statistical Learning Theory.* Springer-Verlag, New York.

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

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

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

Yata, K. and Aoshima, M. (2012b). Misclassification rate adjusted classifier for multi-class, high-dimensional data, submitted.

矢田和善 (2011). 高次元小標本における平均ベクトルの推測とその周辺. 数理解析研究所講究録 **1758**, 136-149.