数理解析研究所講究録
第1703巻 2010年 180-194

Effective Methodologies for High-dimension, Low Sample Size Data

筑波大学・数学系　矢田　和善 (Kazuyoshi Yata)
Institute of Mathematics
University of Tsukuba

1. はじめに

マイクロアレイデータやMRIデータにみられるように，データの次元数dが標本数nよりも遥かに大きな高次元小標本(HDLSS)データが，情報化の進展に伴い，ますます増えてきている．一方，高次元データに従来の統計手法を用いると，いわゆる次元の呪いによって解析が上手くいかないことが，先行研究によって数多く報告されている．例えば，低次元空間への次元縮約法の一つに主成分分析(PCA)がある．HDLSSデータにおいて，従来型のPCAが推定に不一致性を起こすことは，Hall et al. (2005), Ahn et al. (2007), Muller et al. (2009), Jung and Marron (2009)によって報告されている．

彼らが扱った高次元固有値モデルは，基本的には，Johnstone (2001), Baik et al. (2005), Paul (2007), Baik and Silverstein (2006)と同様のもので，モデルに正規性，もしくは，それに類する厳しい制約条件が課されている．また，彼らの研究では，推定に一致性を回復させるための方法論については，解決されないままであった．それに対して，Yata and Aoshima (2009)は，正規性の仮定を外したより柔軟なモデル設定のもと，従来型のPCAが一致性を回復するための条件を導いている．

本研究では，Yata and Aoshima (2010ab)に基づいたHDLSSデータに対する新しいPCA手法であるノイズ掃き出し法とクロスデータ行列法について考察する．HDLSSデータにおいて，どちらの手法も従来型のPCAによる推定を改良するものになっている．本研究で，これらの手法に対して漸近的性質を新しく導出し，理論的，かつ，数値的に比較し，各手法の有効な適用条件を考察する．

2. ノイズ掃き出し法によるPCAの改良

平均が$\mathbf{0}$，共分散行列が$\boldsymbol{\Sigma}$のd次元分布をもつ母集団から，n個のデータベクトルを無作為に抽出して，データ行列$\boldsymbol{X}_{(d)} : d \times n = [\boldsymbol{x}_{1(d)}, \cdots, \boldsymbol{x}_{n(d)}]$を定義する．ただし，$d > n$と仮定する．$\boldsymbol{\Sigma}_d$の固有値を$\lambda_{1(d)} \geq \cdots \geq \lambda_{d(d)} \geq 0$とし，適当な直交行列$\boldsymbol{H}_d = [\boldsymbol{h}_{1(d)}, \cdots, \boldsymbol{h}_{d(d)}]$で$\boldsymbol{\Sigma}_d = \boldsymbol{H}_d \boldsymbol{\Lambda}_d \boldsymbol{H}_d^T$，$\boldsymbol{\Lambda}_d = \mathrm{diag}(\lambda_{1(d)}, \cdots, \lambda_{d(d)})$と分解する．そのとき，$\boldsymbol{Z}_{(d)} = \boldsymbol{\Lambda}_d^{-1/2} \boldsymbol{H}_d^T \boldsymbol{X}_{(d)}$を定義し，$\boldsymbol{Z}_{(d)} = [\boldsymbol{z}_{1(d)}, \cdots, \boldsymbol{z}_{d(d)}]^T$，$\boldsymbol{z}_{i(d)} = (z_{i1(d)}, \cdots, z_{in(d)})^T$と表記する．ただし，$\boldsymbol{Z}_{(d)}$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．今後，次元数を意識して付した添え字dは省い

て表記する．さらに，$n(d)$ を d に依存する標本数 n とし，$n(d)=d^{\gamma}$ とおく．ただし，$\gamma>0$ である．

いま，$\boldsymbol{\Sigma}$ の固有値に次のモデルを仮定する．

$$\lambda_i = a_i d^{\alpha_i}\ (i=1,\cdots,m),\quad \lambda_j = c_j\ (j=m+1,\cdots,d). \tag{2.1}$$

ここで，a_i (>0), c_j (>0), α_i $(\alpha_1 \geq \cdots \geq \alpha_m > 0)$ は未知の実数，m は未知の自然数とする．Jung and Marron (2009) 等も同様のモデルを考えているが，彼らのモデルには，$\alpha_i > 1$ なる制約が課されていることに注意する．標本共分散行列を $\boldsymbol{S}=n^{-1}\boldsymbol{X}\boldsymbol{X}^T$ とする．そのとき，$\boldsymbol{S}_D=n^{-1}\boldsymbol{X}^T\boldsymbol{X}$ は正の固有値を共有する Dual な標本共分散行列となる．いま，$\boldsymbol{Z}$ の成分について，次の条件，

(*) z_{jk}, $j=1,...,d$ $(k=1,...,n)$ が互いに独立

を仮定する．ここで，(*) は $\boldsymbol{Z}$ が正規性を有する場合を含む仮定となっていることに注意する．そのとき，$\boldsymbol{S}_D$ の固有値を $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$ とすると，Yata and Aoshima (2009) の Corollary 3.1 から，

$$\frac{\hat{\lambda}_i}{\lambda_i} = 1 + o_p(1)\quad (i=1,\cdots,m) \tag{2.2}$$

が主張できる．ここで収束は，

(YA-i) $\alpha_i > 1$ のとき $d\to\infty$, $n\to\infty$;

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

で保証される．この結果は，高次元データに正規性を緩めた条件 (*) を仮定されるとき，従来型の PCA が一致性をもつための条件を与える．(2.2) から，$\alpha_i \in (0,1]$ のときに従来型の PCA が一致性をもつためには，標本数 n を $n(d)$ とし，次元数 d に依存して決めるべきだと分かる．これが，HDLSS データに対して，従来型の PCA の適用範囲を狭めることになる原因である．

近年，Yata and Aoshima (2010b) は，条件 (*) が仮定される高次元データに対して，HDLSS の幾何学的構造に基づいた，ノイズ掃き出し法を提案した．それは次のような新しい固有値推定を与えた：

$$\acute{\lambda}_j = \hat{\lambda}_j - \frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{i=1}^{j}\hat{\lambda}_j}{n-j}\quad (j=1,\cdots,n-1). \tag{2.3}$$

ただし，$\acute{\lambda}_j \geq 0$, $j=1,...,n-1$，である．このとき，次の定理が成り立つ．

定理 1 (Yata and Aoshima, 2010b). $\boldsymbol{Z}$ に (*) を仮定する．$\acute{\lambda}_i$ $(i=1,\cdots,m)$ について，

$$\frac{\acute{\lambda}_i}{\lambda_i} = 1 + o_p(1).$$

ここで，収束は，

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

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

で保証される.

上の固有値推定は, 一致性をもつための条件が (2.2) よりも緩いことに注意する.

定理 2 (Yata and Aoshima, 2010b). $\lambda_1 > \cdots > \lambda_m$ と仮定する. $\boldsymbol{Z}$ に $(*)$ を仮定する. $V(z_{ik}^2) = M_i$, $i = 1, ..., d$ $(k = 1, ..., n)$ とおく. そのとき, $\acute{\lambda}_i$ $(i = 1, \cdots, m)$ について,

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

ここで分布収束 $\Rightarrow$ は,

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

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

で保証される.

次に固有ベクトルの推定を考える. $\boldsymbol{S}_D$ 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. そのとき, $\boldsymbol{S}$ の固有ベクトルは $\hat{\boldsymbol{h}}_i = (n\hat{\lambda}_i)^{-1/2} \boldsymbol{X} \hat{\boldsymbol{u}}_i$ で表される. Yata and Aoshima (2009) の Corollary 4.1 から, $\boldsymbol{Z}$ に $(*)$ と $\lambda_1 > \cdots > \lambda_m$ なる仮定と前述の (YA-i)–(YA-ii) なる条件のもとで,

$$\text{Angle}(\hat{\boldsymbol{h}}_i, \boldsymbol{h}_i) \to 0 \quad (i = 1, \cdots, m) \tag{2.4}$$

が主張できる. いま, (2.3) に基づいて $\acute{\boldsymbol{h}}_i = (n\acute{\lambda}_i)^{-1/2} \boldsymbol{X} \hat{\boldsymbol{u}}_i$ による固有ベクトルの推定を考える. このとき, 次の定理が成り立つ.

定理 3 (Yata and Aoshima, 2010b). $\lambda_1 > \cdots > \lambda_m$ を仮定する. $\boldsymbol{Z}$ に $(*)$ を仮定する. 定理 1 の条件 (i)-(ii) のもとで,

$$\acute{\boldsymbol{h}}_i^T \boldsymbol{h}_i = 1 + o_p(1) \quad (i = 1, ..., m)$$

が成り立つ.

次に, 主成分スコアの推定を考える. データ $\boldsymbol{x}_j$ の第 i 主成分スコアは, $\boldsymbol{h}_i^T \boldsymbol{x}_j = \sqrt{\lambda_i} z_{ij}$ $(= s_{ij})$ である. いま, $\boldsymbol{S}_D$ の固有ベクトルの成分を $\hat{\boldsymbol{u}}_i = (\hat{u}_{i1}, \cdots, \hat{u}_{in})^T$ で表せば, 第 i 主成分スコアは $\hat{\boldsymbol{h}}_i^T \boldsymbol{x}_j = \hat{u}_{ij} \sqrt{n\hat{\lambda}_i}$ $(= \hat{s}_{ij})$ で推定される. Yata and

Aoshima (2009) の Corollary 5.1 から，$\boldsymbol{Z}$ に (*) と $\lambda_1 > \cdots > \lambda_m$ なる仮定と前述の (YA-i)-(YA-ii) なる条件のもとで，第 i 主成分スコアの平均二乗誤差 $MSE(\hat{s}_i) = n^{-1}\sum_{j=1}^{n}(\hat{s}_{ij} - s_{ij})^2$ に

$$\frac{MSE(\hat{s}_i)}{\lambda_i} = o_p(1) \quad (i = 1, \cdots, m) \tag{2.5}$$

が主張できる．いま，(2.3) に基づいて $\hat{u}_{ij}\sqrt{n\hat{\lambda}_i}$ $(= \acute{s}_{ij})$ による主成分スコアの推定を考える．このとき，次の定理が成り立つ．

定理 4 (Yata and Aoshima, 2010b). $\lambda_1 > \cdots > \lambda_m$ を仮定する．$\boldsymbol{Z}$ に (*) を仮定する．$MSE(\acute{s}_i) = n^{-1}\sum_{j=1}^{n}(\acute{s}_{ij} - s_{ij})^2$ とおく．定理 1 の条件 (i)-(ii) のもとで，

$$\frac{MSE(\acute{s}_i)}{\lambda_i} = o_p(1) \quad (i = 1, \cdots, m)$$

が成り立つ．

注意 1. $\boldsymbol{Z}$ に (*) を仮定できない状況において，$\acute{\lambda}_i$ $(i = 1, \cdots, m)$ について，定理 1-4 は，条件，

(YA-i') $\alpha_i > 1$ のとき $d \to \infty$, $n \to \infty$;

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

で保証される．

系 1 (Yata and Aoshima, 2010b). 平均が $\mathbf{0}$ 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，$\overline{\boldsymbol{X}} = [\bar{\boldsymbol{x}}, ..., \bar{\boldsymbol{x}}]$, $\bar{\boldsymbol{x}} = \sum_{k=1}^{n}\boldsymbol{x}_k/n$ 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，上の 4 つの定理が成り立つ．

3. クロスデータ行列法による PCA の改良

平均が $\mathbf{0}$ の d 次元分布をもつ母集団があるとする．分布は特に仮定しなくてもよい．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して，$\boldsymbol{\Sigma}$ の固有値には (2.1) のモデルを仮定する．母集団から n 個のデータベクトルを無作為に抽出して，データ行列 $\boldsymbol{X}$: $d \times n$ を定義する．そのとき，Yata and Aoshima (2009) の Theorem 3.1 から，(2.2) が主張できる．ここで収束は，(YA-i')-(Ya-ii') で保証される．この結果から，$\alpha_i \in (0, 1]$ のとき，HDLSS データに対して，従来型の PCA が必ずしも一致性を保証しないことが分かる．

近年，Yata and Aoshima (2010a) は，クロスデータ行列法とよばれる新しい PCA を提唱した．標本を 2 つに分割して得られるデータ行列 $\boldsymbol{X}_l$: $d \times (n/2) = [\boldsymbol{x}_{l1}, \cdots, \boldsymbol{x}_{ln/2}]$ $(l = 1, 2)$ を用いて，クロスデータ行列 $\boldsymbol{S}_{D(1)} = (n/2)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2$ を

定義する．ただし，表記を簡単にするために n は偶数とする．n が奇数の場合は $(n/2+1/2,\ n/2-1/2)$ などに分割する．いま，$\boldsymbol{S}_{D(1)}$ 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．このとき，次の定理が成り立つ．

定理 5 (Yata and Aoshima, 2010a). $\tilde{\lambda}_i\ (i=1,\cdots,m)$ について，

$$\frac{\tilde{\lambda}_i}{\lambda_i} = 1 + o_p(1). \tag{3.1}$$

ここで，収束は，

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

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

で保証される．

上の固有値推定では，一致性をもつための条件が緩まっていることに注意する．

系 2 (Yata and Aoshima, 2010a). $\boldsymbol{Z}$ に $(*)$ を仮定する．そのとき，$\tilde{\lambda}_i\ (i=1,\cdots,m)$ について，定理 1 の条件 (i)-(ii) のもとで，(3.1) が成り立つ．

定理 6 (Yata and Aoshima, 2010a). $\lambda_1 > \cdots > \lambda_m$ と仮定する．$V(z_{ik}^2) = M_i,\ i=1,...,d\ (k=1,...,n)$ とおく．そのとき，$\tilde{\lambda}_i\ (i=1,\cdots,m)$ について，定理 5 の条件 (i)–(ii) のもとで，

$$\sqrt{\frac{n}{M_i}}\left(\frac{\tilde{\lambda}_i}{\lambda_i} - 1\right) \Rightarrow N(0,1) \tag{3.2}$$

が成り立つ．

系 3 (Yata and Aoshima, 2010a). $\lambda_1 > \cdots > \lambda_m$ と仮定する．$\boldsymbol{Z}$ に $(*)$ を仮定する．そのとき，$\tilde{\lambda}_i\ (i=1,\cdots,m)$ について，定理 2 の条件 (i)–(ii) のもとで，(3.2) が成り立つ．

固有ベクトルの推定には，$\boldsymbol{S}_{D(1)}$ の特異値 $\tilde{\lambda}_i$ と特異ベクトル $\tilde{\boldsymbol{u}}_{i(j)},\ j=1,2$ に基づく $\tilde{\boldsymbol{h}}_i = (\tilde{\lambda}_i n/2)^{-1/2}(\boldsymbol{X}_1 \tilde{\boldsymbol{u}}_{i(1)} + \boldsymbol{X}_2 \tilde{\boldsymbol{u}}_{i(2)})/2$ を考える．

定理 7 (Yata and Aoshima, 2010a). $\lambda_1 > \cdots > \lambda_m$ を仮定する．定理 5 の条件 (i)-(ii) のもとで，

$$\tilde{\boldsymbol{h}}_i^T \boldsymbol{h}_i = 1 + o_p(1) \quad (i=1,...,m) \tag{3.3}$$

が成り立つ.

系 4 (Yata and Aoshima, 2010a). $\lambda_1 > \cdots > \lambda_m$ を仮定する. $\boldsymbol{Z}$ に (*) を仮定する. 定理 1 の条件 (i)-(ii) のもとで, (3.3) が成り立つ.

次に, 主成分スコアの推定を考える. $\boldsymbol{S}_{D(1)}$ の特異ベクトルの成分を $\tilde{\boldsymbol{u}}_{i(l)} = (\tilde{u}_{i1(l)}, ..., \tilde{u}_{in/2(l)})^T$ $(l = 1, 2)$ で表す. いま, $\boldsymbol{x}_{lj}$ $(l = 1, 2)$ の第 i 主成分スコアを, $\boldsymbol{S}_{D(1)}$ の特異値と特異ベクトルに基づいて $\tilde{u}_{ij(l)}\sqrt{(n/2)\tilde{\lambda}_i}$ $(= \tilde{s}_{ij(l)})$ で推定することを考える. これ以降は, $\tilde{s}_{ij(1)} = \tilde{s}_{ij}$, $\tilde{s}_{ij(2)} = \tilde{s}_{ij+n/2}$, $j = 1, ..., n/2$ と表記する. いま, $MSE(\tilde{s}_i) = n^{-1}\sum_{j=1}^{n}(\tilde{s}_{ij} - s_{ij})^2$ とおく. そのとき, 次の定理が成り立つ.

定理 8 (Yata and Aoshima, 2010a). $\lambda_1 > \cdots > \lambda_m$ を仮定する. 定理 5 の条件 (i)-(ii) のもとで,

$$\frac{MSE(\tilde{s}_i)}{\lambda_i} = o_p(1) \quad (i = 1, \cdots, m) \tag{3.4}$$

が成り立つ.

系 5 (Yata and Aoshima, 2010a). $\lambda_1 > \cdots > \lambda_m$ を仮定する. $\boldsymbol{Z}$ に (*) を仮定する. 定理 1 の条件 (i)-(ii) のもとで, (3.4) が成り立つ.

系 6 (Yata and Aoshima, 2010a). 平均が $\boldsymbol{0}$ でなければ $\boldsymbol{S}_{oD(1)} = (n/2)^{-1}(\boldsymbol{X}_1 - \overline{\boldsymbol{X}}_1)^T(\boldsymbol{X}_2 - \overline{\boldsymbol{X}}_2)$ とおく. ここで, $\overline{\boldsymbol{X}}_l = [\bar{\boldsymbol{x}}_l, ..., \bar{\boldsymbol{x}}_l]$, $\bar{\boldsymbol{x}}_l = (n/2)^{-1}\sum_{k=1}^{n/2}\boldsymbol{x}_{lk}$ $(l = 1, 2)$ 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, 上の 4 つの定理が成り立つ.

4. ノイズ掃き出し法とクロスデータ行列法の比較

ノイズ掃き出し法は, データに正規性を緩めた条件 (*) を仮定できるセミパラメトリックな状況で有効な手法である. 一方, クロスデータ行列法は母集団分布に仮定がないノンパラメトリックな状況で有効な手法である. $\boldsymbol{Z}$ に (*) を仮定できないノンパラメトリックな状況においては, 注意 1 と定理 5-8 より, ノイズ掃き出し法に比べ, クロスデータ行列法における PCA の方が有効な手法といえる. 次に, 条件 (*) が仮定できるセミパラメトリックな状況のもとで, ノイズ掃き出し法とクロスデータ行列法を比較する.

いま, $\boldsymbol{Z}$ に (*) を仮定できるとき, 定理 2 と系 3 より, 定理 2 の収束条件 (i)-(ii) のもとで固有値の漸近分布は,

$$\sqrt{\frac{n}{M_i}}\left(\frac{\tilde{\lambda}_i}{\lambda_i} - 1\right) \Rightarrow N(0, 1), \quad \sqrt{\frac{n}{M_i}}\left(\frac{\tilde{\lambda}_i}{\lambda_i} - 1\right) \Rightarrow N(0, 1) \tag{4.1}$$

となり，漸近的に二つの固有値の推定量は同じ分布をもち，同等な推定量だといえる．それに対して，本論文では主成分スコアの推定量における平均二乗誤差について，新たな漸近理論を導出する．

いま，ノイズ掃き出し法における主成分スコアの平均二乗誤差 $MSE(\acute{s}_i)$ について，次の定理が成り立つ．

定理 9. $\lambda_1 > \cdots > \lambda_m$ を仮定する．$\boldsymbol{Z}$ に $(*)$ を仮定する．定理 2 の条件 (i)-(ii) のもとで，

$$\frac{MSE(\acute{s}_i)}{\lambda_i} = o_p(n^{-1/2}) \quad (i = 1, \cdots, m)$$

が成り立つ．

一方，クロスデータ行列法における主成分スコアの平均二乗誤差 $MSE(\tilde{s}_i)$ について，次の定理が成り立つ．

定理 10. $\lambda_1 > \cdots > \lambda_m$ を仮定する．$\boldsymbol{Z}$ に $(*)$ を仮定する．定理 2 の条件 (i)-(ii) のもとで，

$$\frac{MSE(\tilde{s}_i)}{\lambda_i} = o_p((n/2)^{-1/2}) = o_p(n^{-1/2}) \quad (i = 1, \cdots, m)$$

が成り立つ．

それゆえ，定理 9-10 より，条件 $(*)$ のもとでノイズ掃き出し法における主成分スコアの平均二乗誤差 $MSE(\acute{s}_i)$ とクロスデータ行列法のものと比較すると，$n^{-1/2}$ のオーダーまでは漸近的にどちらも 0 となり，等しくなる．しかしながら，クロスデータ行列法においては，$n/2 \to \infty$ なる条件のもとでの結果であることに注意すると，標本数 n が小さくなるにつれ，主成分スコアの平均二乗誤差はクロスデータ行列法の方が大きくなると予想される．

5. シミュレーション

本論文では，Yata and Aoshima (2010ab) に基づいた HDLSS データに対する新しい PCA 手法であるノイズ掃き出し法とクロスデータ行列法を紹介し，さらに主成分スコアにおける漸近的性質を導出した．これらの結果により，データに正規性を緩めた条件 $(*)$ を仮定できるセミパラメトリックな状況では，(4.1) に見られるように，ノイズ掃き出し法とクロスデータ行列法に基づく固有値の推定量は同じ分布に従い，同等な推定量といえる．対して，主成分スコアの推定に関しては，ある程度大きな標本数 n のもとであれば，定理 9-10 より，ノイズ掃き出し法とクロスデータ行列法においては，共に良い推定量といえる．一方，少ない標本数 n の

もとでは，ノイズ掃き出し法の方がクロスデータ行列法に比べ，有効な主成分スコアの推定量になると予想される．

それに対して，$\boldsymbol{Z}$ に $(*)$ を仮定できないノンパラメトリックな状況では，クロスデータ行列法は通常の PCA に対して，有効であることが理論的に保証される．一方，ノイズ掃き出し法については，注意 1 より，通常の PCA に比べても改良できているとはいえない．本節では，これらをモンテカルロ・シミュレーションで確認する．

下の図 1-1（第 1 固有値），図 1-2（第 2 固有値），図 1-3（第 3 固有値）は，$d = 1600$ 次元の正規乱数 $N_d(\mathbf{0}, \boldsymbol{\Sigma})$ を生成して，標本数 $n \in [20, 120]$ における A : $\hat{\lambda}_i/\lambda_i$，B : $\acute{\lambda}_i/\lambda_i$，C : $\tilde{\lambda}_i/\lambda_i$ の値について，それぞれ 1000 回のシミュレーション実験を行い，その平均値をプロットしたものである．ここでは，(2.1) のモデルにおいて，パラメータを $\lambda_1 = d^{4/5}$，$\lambda_2 = d^{3/5}$，$\lambda_3 = d^{2/5}$，$\lambda_4 = \cdots = \lambda_d = 1$ と設定した．

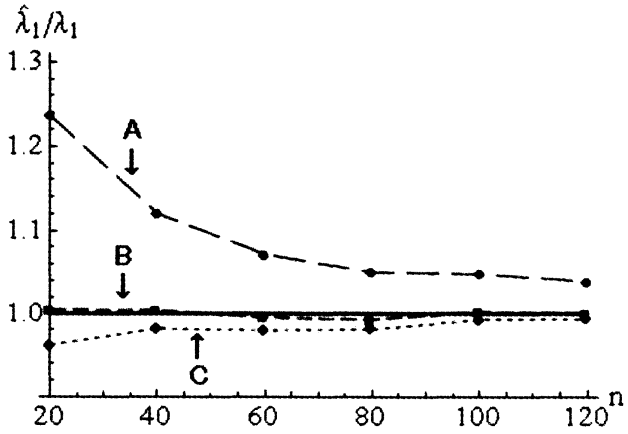

図 1-1. 第 1 固有値

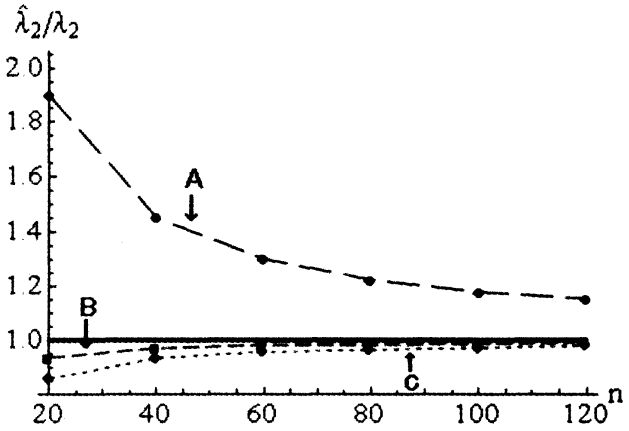

図 1-2. 第 2 固有値

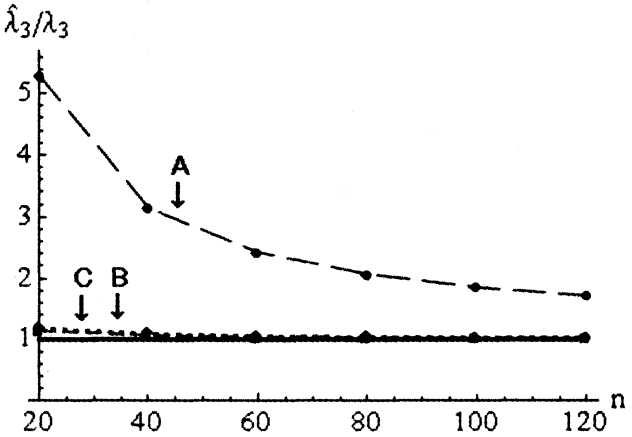

図 1-3. 第 3 固有値

これらの図から分かるように，固有値の推定について，B と C で与えた固有値の推定量 $\acute{\lambda}_i$，$\tilde{\lambda}_i$ が通常の固有値の推定量 $\hat{\lambda}_i$ に比べ，良い推定量になっている．ここで，$\acute{\lambda}_i$ と $\tilde{\lambda}_i$ は推定に大きな差がなく，同等な推定量だといえる．さらに，下の図 2-1（第 1 固有値の分散），図 2-2（第 2 固有値の分散），図 2-3（第 3 固有値の分散）は，A : $\hat{\lambda}_i/\lambda_i$ B : $\acute{\lambda}_i/\lambda_i$，C : $\tilde{\lambda}_i/\lambda_i$ の値について，先ほどのシミュレーション

実験における不偏分散の値をプロットしたものである.

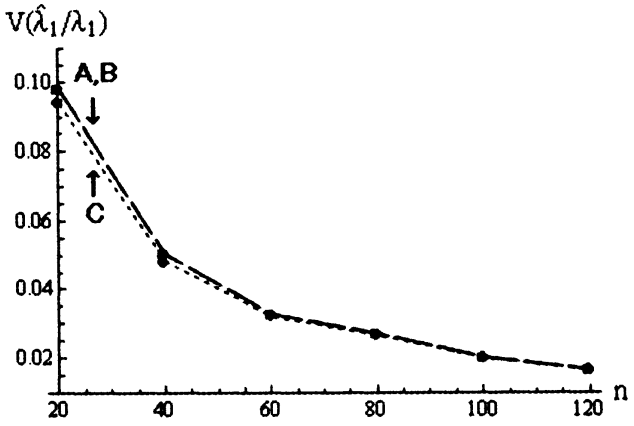

図 2-1. 第 1 固有値の分散

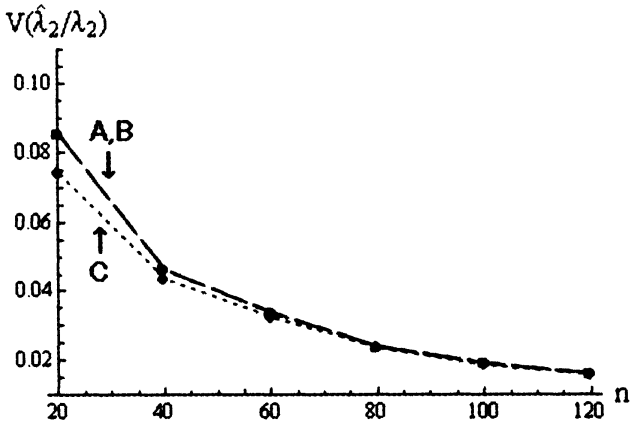

図 2-2. 第 2 固有値の分散

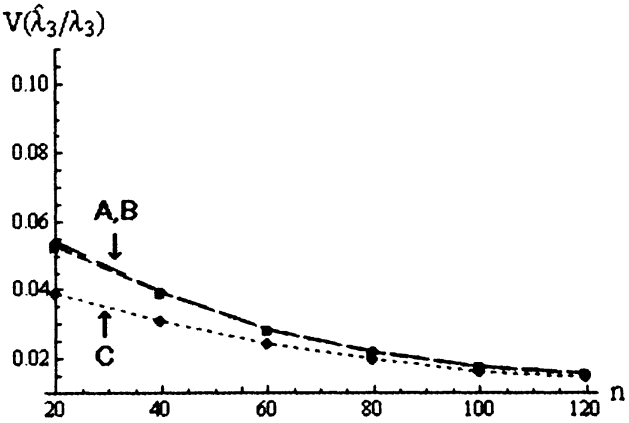

図 2-3. 第 3 固有値の分散

C で与えたクロスデータ行列法に基づく推定では，標本を 2 分割して推定量を定義するので，λ_i の推定が A や B と比べて不安定になるのではと危惧される．しかしながら，これらの図から分かるように，第 1 固有値から第 3 固有値の何れも，A, B と C による推定の分散は，ほぼ等しくなっている．それは (4.1) からも各推定量の分散が漸近的に等しくなることは理論的に正しい結果であり，その安定した挙動がシミュレーションで確認されたということである．

次に図 3-1（第 1 主成分スコア），図 3-2（第 2 主成分スコア），図 3-3（第 3 主成分スコア) は各主成分スコアの平均二乗誤差，A : $MSE(\hat{s}_i)/\lambda_i$, B : $MSE(\acute{s}_i)/\lambda_i$, C : $MSE(\tilde{s}_i)/\lambda_i$ の値について，それぞれ 1000 回のシミュレーション実験を行い，その平均値をプロットしたものである．

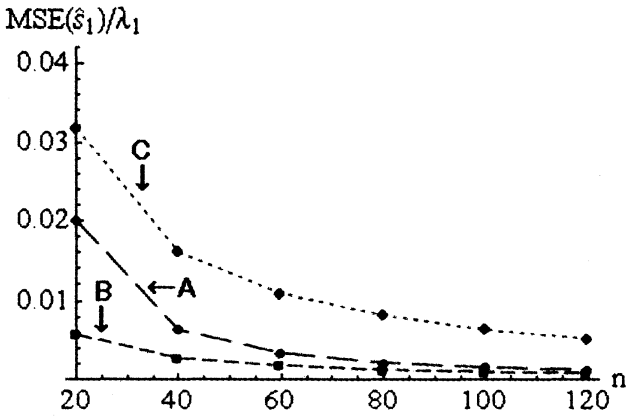

図 3-1. 第 1 主成分スコア

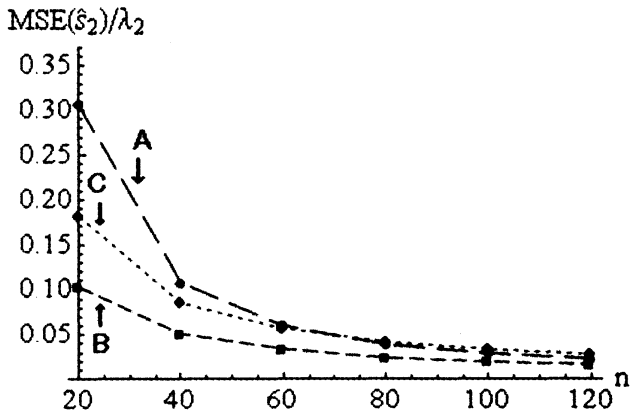

図 3-2. 第 2 主成分スコア

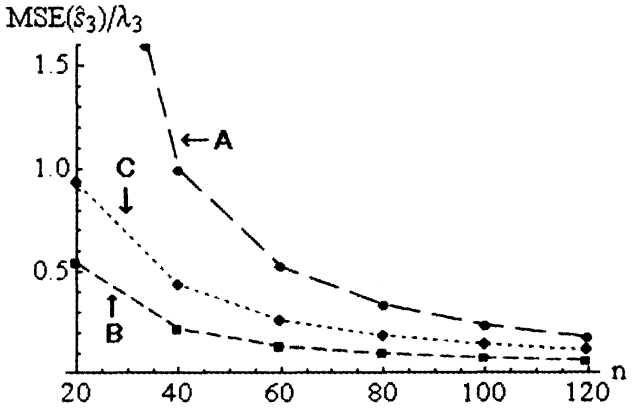

図 3-3. 第 3 主成分スコア

この結果から，ノイズ掃き出し法の方がクロスデータ行列法により，有効に主成分スコアを推定できることが数値的にも保証された．特に少ない標本数において，ノイズ掃き出し法が有効だといえる．

次にデータに (*) が仮定できない非正規分布のもとで，シミュレーション実験を行う．平均 $\mathbf{0}$，共分散行列 $\mathbf{\Sigma}$，自由度 ν の $d = 1600$ 次元の t 分布の乱数を生成する．下の図 4-1（第 1 固有値），図 4-2（第 2 固有値），図 4-3（第 3 固有値）は標本数 60，自由度 $\nu \in [5, 25]$ における A : $\hat{\lambda}_i/\lambda_i$，B : $\acute{\lambda}_i/\lambda_i$，C : $\tilde{\lambda}_i/\lambda_i$ の値について，それぞれ 1000 回のシミュレーション実験を行い，その平均値をプロットしたものである．ここで，$\mathbf{\Sigma}$ に，$\lambda_1 = d^{4/5}$，$\lambda_2 = d^{3/5}$，$\lambda_3 = d^{2/5}$，$\lambda_4 = \cdots = \lambda_d = 1$ と設定した．

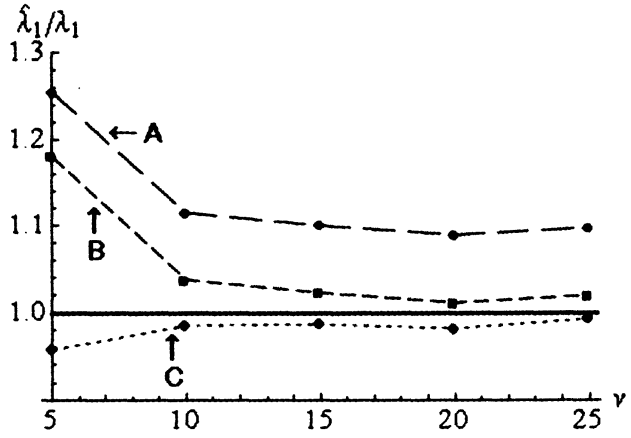

図 4-1. 第 1 固有値

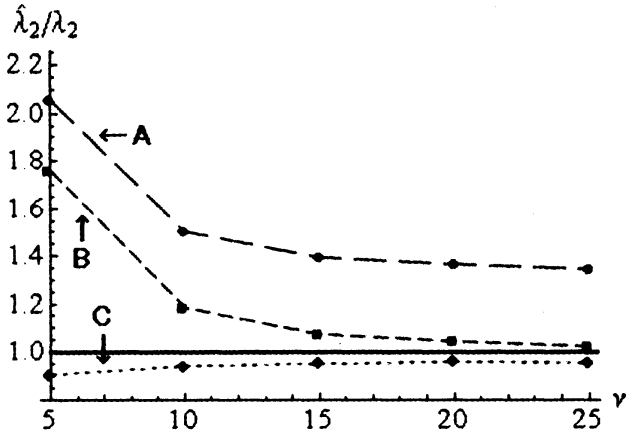

図 4-2. 第 2 固有値

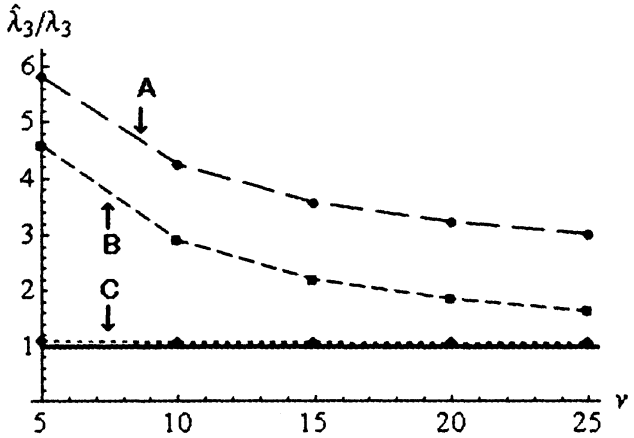

図 4-3. 第 3 固有値

いま，自由度 ν が大きくなるにつれ，t 分布は正規分布に近づくことに注意をすると，これらの図から分かるように，自由度が小さく，正規分布から離れた分布においては，クロスデータ行列法がノイズ掃き出し法に比べて，良い推定になっている．自由度 ν が大きく，正規分布に近い分布においては，ノイズ掃き出し法も良い推定量を構築できていることが分かる．

次に図 5-1（第 1 主成分スコア），図 5-2（第 2 主成分スコア），図 5-3（第 3 主成分スコア）は各主成分スコアの平均二乗誤差，A : $MSE(\hat{s}_i)/\lambda_i$, B : $MSE(\acute{s}_i)/\lambda_i$, C : $MSE(\tilde{s}_i)/\lambda_i$ の値について，それぞれ 1000 回のシミュレーション実験を行い，その平均値をプロットしたものである．

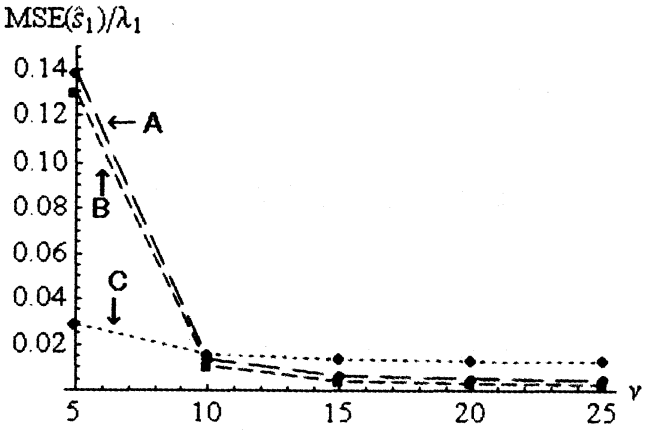

図 5-1. 第 1 主成分スコア

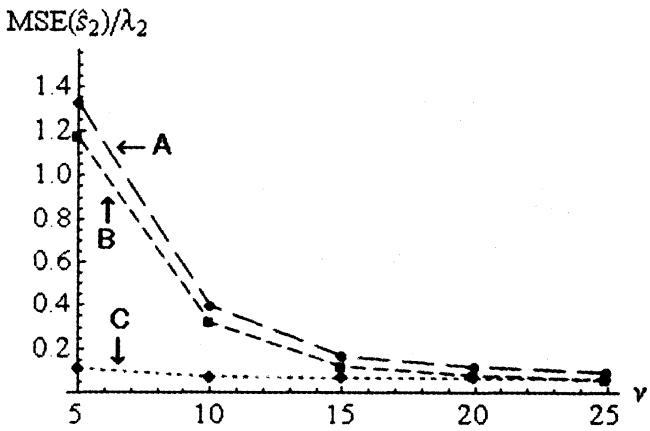

図 5-2. 第 2 主成分スコア

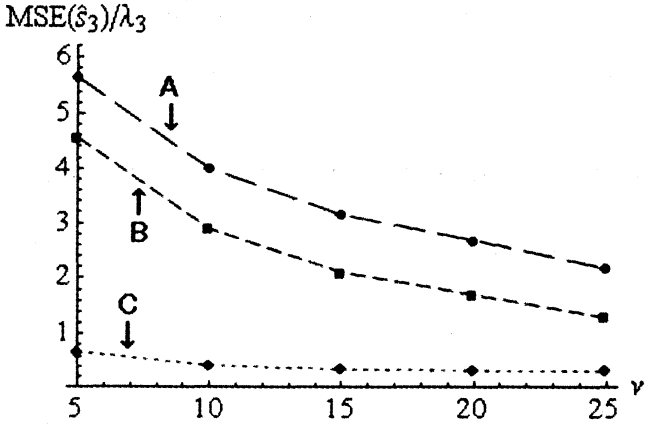

図 5-3. 第 3 主成分スコア

この結果から，クロスデータ行列法の方がノイズ掃き出し法に比べ，特に正規分布から離れた分布においては，主成分スコアを良く推定できることが確認された．

今回の結果を含め多くの実験結果から，HDLSS データに対して，ノイズ掃き出し法は，データが正規性を有するような状況で有効な手法だといえる．それに対して，クロスデータ行列法は，非正規性のもとで有効な手法だといえる．本研究において，母集団分布に正規分布，もしくは，それに近い分布が仮定することが出来るのであれば，ノイズ掃き出し法を用いることを推奨する．それに対して，母集団分布が非正規分布，もしくは，分布が仮定できないノンパラメトリックな状況であれば，クロスデータ行列法を用いることを推奨する．

Appendix

Appendix を通して，$\boldsymbol{z}_{1i} = (z_{i1}, ..., z_{in/2})^T$，$\boldsymbol{z}_{2i} = (z_{in/2+1}, ..., z_{in})^T$ $(i = 1, ..., m)$ とし，

$$\tilde{\boldsymbol{z}}_i = (||n^{-1/2}\boldsymbol{z}_i||)^{-1} n^{-1/2}\boldsymbol{z}_i, \quad \tilde{\boldsymbol{z}}_{i'i} = (||(n/2)^{-1/2}\boldsymbol{z}_{i'i}||)^{-1}(n/2)^{-1/2}\boldsymbol{z}_{i'i}$$
$$(i' = 1, 2;\ i = 1, ..., m)$$

とおく.

次の補題はYata and Aoshima (2010ab) から得る.

補題 1 $\lambda_1 > \cdots > \lambda_m$ と仮定する. $\boldsymbol{Z}$ に (*) を仮定する. そのとき, 定理2の条件 (i)-(ii) のもとで,

$$\frac{\acute{\lambda}_i}{\lambda_i} = ||n^{-1/2}\boldsymbol{z}_i||^2 + o_p(n^{-1/2}) = 1 + o_p(1), \quad \hat{\boldsymbol{u}}_i^T\tilde{\boldsymbol{z}}_i = 1 + o_p(n^{-1/2}) \quad (i = 1, ..., m)$$

が成り立つ.

補題 2 $\lambda_1 > \cdots > \lambda_m$ と仮定する. $\boldsymbol{Z}$ に (*) を仮定する. そのとき, 定理2の条件 (i)-(ii) のもとで,

$$\frac{\tilde{\lambda}_i}{\lambda_i} = \left(||(n/2)^{-1/2}\boldsymbol{z}_{1i}||\right)\left(||(n/2)^{-1/2}\boldsymbol{z}_{2i}||\right) + o_p((n/2)^{-1/2}) = 1 + o_p(1),$$
$$\tilde{\boldsymbol{u}}_{i(i')}^T\tilde{\boldsymbol{z}}_{i'i} = 1 + o_p((n/2)^{-1/2}) \quad (i' = 1, 2;\ i = 1, ..., m)$$

が成り立つ.

定理9の証明 いま, $i\ (= 1, ..., m)$ において, 次のように表記できる

$$\begin{aligned}
\mathrm{MSE}(\acute{s}_i) &= \lambda_i n^{-1}\sum_{k=1}^{n}\left(z_{ik} - \sqrt{n\frac{\acute{\lambda}_i}{\lambda_i}}\hat{u}_{ik}\right)^2 \\
&= \lambda_i\left(n^{-1}\sum_{k=1}^{n}z_{ik}^2 + \frac{\acute{\lambda}_i}{\lambda_i}\sum_{k=1}^{n}\hat{u}_{ik}^2 - 2\sqrt{\frac{\acute{\lambda}_i}{\lambda_i}}\frac{\boldsymbol{z}_i^T}{\sqrt{n}}\hat{\boldsymbol{u}}_i\right) \\
&= \lambda_i\left(||n^{-1/2}\boldsymbol{z}_i||^2 + \frac{\acute{\lambda}_i}{\lambda_i} - 2||n^{-1/2}\boldsymbol{z}_i||\sqrt{\frac{\acute{\lambda}_i}{\lambda_i}}\tilde{\boldsymbol{z}}_i^T\hat{\boldsymbol{u}}_i\right).
\end{aligned}$$

このとき, 補題1より, 定理2の条件 (i)-(ii) のもとで,

$$\mathrm{MSE}(\acute{s}_i)/\lambda_i = 2||n^{-1/2}\boldsymbol{z}_i||^2 - 2||n^{-1/2}\boldsymbol{z}_i||^2\sqrt{1 + o_p(n^{-1/2})} + o_p(n^{-1/2}) = o_p(n^{-1/2})$$

を得る. □

定理10の証明　いま，$i\,(=1,...,m)$において，次のように表記できる

$$
\begin{aligned}
&MSE(\tilde{s}_i)\\
&=\lambda_i n^{-1}\sum_{k=1}^{n/2}\left(z_{ik}-\sqrt{\frac{n\tilde{\lambda}_i}{2\lambda_i}}\tilde{u}_{ik(1)}\right)^2+\lambda_i n^{-1}\sum_{k=n/2+1}^{n}\left(z_{ik}-\sqrt{\frac{n\tilde{\lambda}_i}{2\lambda_i}}\tilde{u}_{ik-n/2(2)}\right)^2\\
&=\frac{\lambda_i}{2}\left((n/2)^{-1}\sum_{k=1}^{n/2}z_{ik}^2+\frac{\tilde{\lambda}_i}{\lambda_i}\sum_{k=1}^{n/2}\tilde{u}_{ik(1)}^2-2\sqrt{\frac{\tilde{\lambda}_i}{\lambda_i}}((n/2)^{-1/2}\boldsymbol{z}_{1i}^T\boldsymbol{u}_{i(1)})\right)\\
&\quad+\frac{\lambda_i}{2}\left((n/2)^{-1}\sum_{k=n/2+1}^{n}z_{ik}^2+\frac{\tilde{\lambda}_i}{\lambda_i}\sum_{k=n/2+1}^{n}\tilde{u}_{ik-n/2(2)}^2-2\sqrt{\frac{\tilde{\lambda}_i}{\lambda_i}}((n/2)^{-1/2}\boldsymbol{z}_{2i}^T\boldsymbol{u}_{i(2)})\right)\\
&=\lambda_i\left(||n^{-1/2}\boldsymbol{z}_i||^2+\frac{\tilde{\lambda}_i}{\lambda_i}-\sqrt{\frac{\tilde{\lambda}_i}{\lambda_i}}\left(||(n/2)^{-1/2}\boldsymbol{z}_{1i}||\tilde{\boldsymbol{z}}_{1i}^T\boldsymbol{u}_{i(1)}+||(n/2)^{-1/2}\boldsymbol{z}_{2i}||\tilde{\boldsymbol{z}}_{2i}^T\boldsymbol{u}_{i(2)}\right)\right)
\end{aligned}
\tag{A.1}
$$

ここで，$n\to\infty$のもとで，

$$
||(n/2)^{-1/2}\boldsymbol{z}_{i'i}||=1+\frac{1}{2}(||(n/2)^{-1/2}\boldsymbol{z}_{i'i}||^2-1)+o_p((n/2)^{-1/2})\quad(i'=1,2),
$$

$$
\begin{aligned}
&\left(||(n/2)^{-1/2}\boldsymbol{z}_{1i}||\right)\left(||(n/2)^{-1/2}\boldsymbol{z}_{2i}||\right)\\
&=\frac{1}{2}\left(||(n/2)^{-1/2}\boldsymbol{z}_{1i}||^2+||(n/2)^{-1/2}\boldsymbol{z}_{2i}||^2\right)+o_p((n/2)^{-1/2})\\
&=||n^{-1/2}\boldsymbol{z}_i||^2+o_p((n/2)^{-1/2})
\end{aligned}
$$

に注意し，(A.1)と補題2より，定理2の条件(i)-(ii)のもとで，

$$
\begin{aligned}
&MSE(\tilde{s}_i)/\lambda_i\\
&=2||n^{-1/2}\boldsymbol{z}_i||^2-\sqrt{||n^{-1/2}\boldsymbol{z}_i||^2}(1+||n^{-1/2}\boldsymbol{z}_i||^2)+o_p((n/2)^{-1/2})\\
&=2||n^{-1/2}\boldsymbol{z}_i||^2-\left(1+\frac{1}{2}(||n^{-1/2}\boldsymbol{z}_i||^2-1)\right)(1+||n^{-1/2}\boldsymbol{z}_i||^2)+o_p((n/2)^{-1/2})\\
&=2||n^{-1/2}\boldsymbol{z}_i||^2-2\left(1+\frac{1}{2}(||n^{-1/2}\boldsymbol{z}_i||^2-1)\right)^2+o_p((n/2)^{-1/2})=o_p((n/2)^{-1/2})
\end{aligned}
$$

を得る．　□

謝辞　本研究は，科学研究費補助金 基盤研究(B) 22300094 研究代表者: 青嶋 誠「高次元データの理論と方法論の総合的研究」から，研究助成を受けています．

References

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. (2009). Eigenvalues estimation for high dimension, Gaussian data and sample size determination. IWSM 2009 Proceedings, in press

Baik, J., Ben Arous, G., and Péché, S. (2005). Phase transition of the largest eigenvalue for non-null complex covariance matrices. *Ann. Probab.*, **33**, 1643–1697.

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

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.

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

Jung, S. and Marron, J. S. (2009). PCA Consistency in High Dimension, Low Sample Size Context. *Annals of Statistics* 37: 4104-4130.

Muller, K. E., Chi, Y.-Y., Ahn, J. and Marron, J. S. (2009). Limitations of high dimension, low sample size principal components for gaussian data. *J. Amer. Statist. Assoc.*, revised.

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

Yata, K. and Aoshima, M. (2009). PCA consistency for non-Gaussian data in high dimension, low sample size context. *Commun. Statist.-Theory Meth.*, **38**, 2634-2652.

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

Yata, K. and Aoshima, M. (2010b). PCA consistency for high-dimension, low sample size data with geometric representations. *J. Mult. Anal.*, submitted.