数理解析研究所講究録
第1954巻 2015年 51-60

拡張クロスデータ行列法と共分散行列関数の不偏推定

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

高次元データの解析には，母集団に正規分布を仮定しない方法論が必要になる．さらに，膨大なデータを処理するために，低い計算コストで高精度な解析結果を出力できるようなアルゴリズムが求められる．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法を考案した．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．Aoshima and Yata (2011) は，高次元データの統計的推測に幾何学的表現を導入し，クロスデータ行列法に基づいた各種方法論を考案し，統計量の高次元漸近正規性，標本数の設計，推測の精度保証に至るまでの一連の基礎理論を与えた．Yata and Aoshima (2013) は，漸近最適な組み合わせに基づいてクロスデータ行列法を拡張し「拡張クロスデータ行列法 (ECDM)」を提案して，相関係数ベクトルに関する推定・検定を構築した．さらに，Yata and Aoshima (2015) は，ECDMによる推定量・検定統計量を一般化した形で与え，それらが高次元のもとで一致性と漸近正規性を有することを示し，共分散構造に関する推測に応用した．

本論文は，ECDMが高い汎用性をもつことを示すものである．ECDMに基づいて，共分散行列に関する各種特徴量に不偏推定量を導き，高次元における漸近的性質を論じる．さらに，固定次元数を扱う通常の多変量解析の枠組みでも，ECDMが有効に機能することを示す．平均にp次ベクトル$\boldsymbol{\mu}$，共分散行列にp次の非負定値対称行列$\boldsymbol{\Sigma}(\neq \boldsymbol{O})$をもつ母集団を考える．母集団から$n$ (≥ 4)個のp次データベクトル$\boldsymbol{x}_1, ..., \boldsymbol{x}_n$を無作為に抽出する．次のモデルを考える．

$$\boldsymbol{x}_j = \boldsymbol{\Gamma}\boldsymbol{w}_j + \boldsymbol{\mu}, \quad j = 1, ..., n.$$

ここで，$\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_r)$は$\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \boldsymbol{\Sigma}$なる$p \times r$行列，$\boldsymbol{w}_j$は$E(\boldsymbol{w}_j) = \boldsymbol{0}$, $\mathrm{Var}(\boldsymbol{w}_j) = \boldsymbol{I}_r$なる$r$次確率ベクトルとする．このモデルは，Bai and Saranadasa (1996), Chen and Qin (2010), Aoshima and Yata (2013) 等で解析された．いま，$\boldsymbol{w}_j =$

$(w_{1j},...,w_{rj})^T$, $M_i=\mathrm{Var}(w_{ij}^2)$, $i=1,...,r$ とおき，$\limsup_{p\to\infty} M_i<\infty$, $i=1,...,r$ であることを仮定する．母集団分布には，必要な箇所でその都度，次を仮定する．

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

(A-i) は正規分布を緩めた仮定になっている．

2節以降で用いるECDMを，ここで簡単に纏めておく．いま，$n_{(1)}=\lceil n/2\rceil$, $n_{(2)}=n-n_{(1)}$ とおく．$\lceil x\rceil$ は x 以上の最小の整数を表す．2つの集合 $\boldsymbol{V}_{n(1)(k)}$, $\boldsymbol{V}_{n(2)(k)}$ $(k=3,...,2n-1)$ を次のように定義する．

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

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

ここで，$\lfloor x\rfloor$ は x 以下の最大の整数を表す．そのとき，$k=3,...,2n-1$ について，$\#\boldsymbol{V}_{n(l)(k)}=n_{(l)}$, $l=1,2$, $\boldsymbol{V}_{n(1)(k)}\cap\boldsymbol{V}_{n(2)(k)}=\emptyset$, $\boldsymbol{V}_{n(1)(k)}\cup\boldsymbol{V}_{n(2)(k)}=\{1,...,n\}$ 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，$\#\boldsymbol{S}$ は集合 $\boldsymbol{S}$ の要素の個数を表す．$\boldsymbol{V}_{n(1)(i+j)}$ と $\boldsymbol{V}_{n(2)(i+j)}$ に基づいて不偏推定量を構築する手法が，拡張クロスデータ行列法 (ECDM) である．

2 共分散行列に関する不偏推定量

本節では，共分散行列 $\boldsymbol{\Sigma}$ に関する不偏推定量をECDMによって導く．

2.1 $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ の不偏推定量

高次元データの推測に精度を保証するための鍵となるパラメータの1つが，$\mathrm{tr}(\boldsymbol{\Sigma}^2)$ $(=\delta)$ である．例えば，Aoshima and Yata (2011, 2013)，青嶋・矢田 (2013) を参照のこと．標本共分散行列 $\boldsymbol{S}_n$ を用いた単純な推定量 $\mathrm{tr}(\boldsymbol{S}_n^2)$ は，高次元データに対して非常に大きなバイアスをもち役に立たない．Aoshima and Yata (2011) は，クロスデータ行列法を用いて δ の推定を考えた．標本を2分割し，各分割から標本共分散行列 $\boldsymbol{S}_{n(i)}$, $i=1,2$ を計算し，δ の不偏推定量 $\mathrm{tr}(\boldsymbol{S}_{n(1)}\boldsymbol{S}_{n(2)})$ を与えた．

一方で，Bai and Saranadasa (1996)，Srivastava (2005) は，推定量

$$\widehat{\delta}_{BS}=\frac{(n-1)^2}{(n-2)(n+1)}\Big(\mathrm{tr}(\boldsymbol{S}_n^2)-\frac{\mathrm{tr}(\boldsymbol{S}_n)^2}{n-1}\Big)$$

を与えた．母集団に正規分布を仮定すれば，$E(\widehat{\delta}_{BS})=\delta$ なる不偏性をもち，$p\to\infty$，$n\to\infty$ のとき

$$\mathrm{Var}\Big(\frac{\widehat{\delta}_{BS}}{\delta}\Big)=\Big(\frac{8\mathrm{tr}(\boldsymbol{\Sigma}^4)}{n\delta^2}+\frac{4}{n^2}\Big)\{1+o(1)\}\to 0$$

となる．しかし，母集団に正規分布を仮定できないと，$\widehat{\delta}_{BS}$ 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が生じる．さらに，$\boldsymbol{w}_j$ の成分に8次モーメントの一様有界性が仮定できないと，$\mathrm{Var}(\widehat{\delta}_{BS}/\delta)<\infty$ さえ保証できない．

ECDM を使えば，δ の不偏推定量は次のように導かれる．各 k $(=3,...,2n-1)$ について，2分割した集合の標本平均を

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

とし，δ の1つの不偏推定量として $u_n\{(\boldsymbol{x}_i-\overline{\boldsymbol{x}}_{(1)(i+j)})^T(\boldsymbol{x}_j-\overline{\boldsymbol{x}}_{(2)(i+j)})\}^2$ $(i<j)$ を計算する．ただし，$u_n=n_{(1)}n_{(2)}/\{(n_{(1)}-1)(n_{(2)}-1)\}$ である．すべての組合せで平均をとり

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

を定義する．このとき，$\widehat{\delta}$ は母集団分布に依らずに不偏性 $E(\widehat{\delta})=\delta$ をもつ．さらに，$P(\widehat{\delta}\geq 0)=1$ が成り立つ．これは，$\widehat{\delta}$ が母数空間に値をもつことを意味する．Aoshima and Yata (2013), Yata and Aoshima (2013) から，次の結果を得る．

定理1. 母集団分布に (A-i) を仮定する．$n\to\infty$，かつ，$p\to\infty$ もしくは $p<\infty$ のとき，次が成り立つ．

$$\mathrm{Var}\Big(\frac{\widehat{\delta}}{\delta}\Big)=\Big\{\frac{4}{n\delta^2}\Big(2\mathrm{tr}(\boldsymbol{\Sigma}^4)+\sum_{i=1}^{r}(M_i-2)(\boldsymbol{\gamma}_i^T\boldsymbol{\Sigma}\boldsymbol{\gamma}_i)^2\Big)+\frac{4}{n^2}\Big\}\{1+o(1)\}\to 0.$$

母集団に正規分布を仮定すると，$M_i=2$，$i=1,...,r$ であることに注意すれば，$p\to\infty$，$n\to\infty$ において，$\widehat{\delta}$ と $\widehat{\delta}_{BS}$ の漸近分散は同等であることが分かる．

注意1. 次の手順は，$\widehat{\delta}$ の計算コストが $O(pn^2)$ のオーダーになり効率がよい．

(手順1) $\overline{\boldsymbol{x}}_{(l)(k)}$，$l=1,2$ を各 k $(=3,...,2n-1)$ で計算する．

(手順2) すべての i,j $(1\leq i<j\leq n)$ について手順1の $\overline{\boldsymbol{x}}_{(l)(i+j)}$ を代入して $u_n\{(\boldsymbol{x}_i-\overline{\boldsymbol{x}}_{(1)(i+j)})^T(\boldsymbol{x}_j-\overline{\boldsymbol{x}}_{(2)(i+j)})\}^2$ を計算し，それらの平均をとって $\widehat{\delta}$ を得る．

注意 2. $\widehat{\delta}$ の計算アルゴリズム (Mathematica code) は次の通りである.

Input: Sample size n; $n \times p$ data matrix X such as $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^T$.

Mathematica code:

- $n1$ =Ceiling[$n/2$]; $n2 = n - n1$; $u = 2 * n1 * n2/((n1 - 1) * (n2 - 1) * n * (n - 1))$
- V[1, $k\_$, $X\_$] :=If [Floor[$k/2$] $\geq n1$, Take[X, {Floor[$k/2$] $- n1 + 1$, Floor[$k/2$]}], Join[Take[X, {1, Floor[$k/2$]}], Take[X, {Floor[$k/2$] $+ n2 + 1, n$}]]]
- V[2, $k\_$, $X\_$] :=If [Floor[$k/2$] $\leq n1$, Take[X, {Floor[$k/2$] $+ 1$, Floor[$k/2$] $+ n2$}], Join[Take[X, {1, Floor[$k/2$] $- n1$}, Take[X, {Floor[$k/2$] $+ 1, n$}]]]
- Do[M[i, k] =Mean[$V[i, k, X]$], $\{k, 3, 2 * n - 1\}$, $\{j, 1, 2\}$]
- $T = u$*Sum[((Part[X, i]−M[$1, i + j$]).(Part[X, j]−M[$2, i + j$]))^2, $\{j, 2, n\}$, $\{i, 1, j - 1\}$]

そのとき, $\widehat{\delta} = T$ を得る.

Chen et al. (2010) は, U-統計量に基づいて, δ の不偏推定量を次のように与えた.

$$\widehat{\delta}_C = \sum_{i \neq j}^{n} \frac{(\boldsymbol{x}_i^T \boldsymbol{x}_j)^2}{n(n-1)} - 2 \sum_{i \neq j \neq k}^{n} \frac{\boldsymbol{x}_i^T \boldsymbol{x}_j \boldsymbol{x}_j^T \boldsymbol{x}_k}{n(n-1)(n-2)} + \frac{\sum_{i \neq j \neq k \neq l}^{n} \boldsymbol{x}_i^T \boldsymbol{x}_j \boldsymbol{x}_k^T \boldsymbol{x}_l}{n(n-1)(n-2)(n-3)}.$$

これは, $\widehat{\delta}$ と同等の漸近分散をもつが, 計算コストが $O(pn^4)$ と非常に大きく実用には向かない. さらに, $P(\widehat{\delta}_C \geq 0) = 1$ が保証されない. 最近になって, Srivastava et al. (2014) は $\widehat{\delta}_C$ を式変形して, 計算コストを $O(pn^2)$ に抑える

$$\begin{aligned}\widehat{\delta}_C &= \widehat{\delta}_C(\boldsymbol{Y}) \\ &= \frac{1}{n(n-1)(n-2)(n-3)} \Big((n-1)(n-2)\mathrm{tr}(\boldsymbol{M}^2) - n(n-1)\mathrm{tr}(\boldsymbol{D}^2) + \mathrm{tr}(\boldsymbol{D})^2 \Big)\end{aligned}$$

なる書き換えを考えた. ここで, $\boldsymbol{Y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_n)$, $\boldsymbol{y}_i = \boldsymbol{x}_i - \overline{\boldsymbol{x}}_n$, $i = 1, ..., n$, $\overline{\boldsymbol{x}}_n = n^{-1} \sum_{j=1}^{n} \boldsymbol{x}_j$, $\boldsymbol{M} = \boldsymbol{Y}^T \boldsymbol{Y}$, $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{y}_1^T \boldsymbol{y}_1, ..., \boldsymbol{y}_n^T \boldsymbol{y}_n)$ である. Yata and Aoshima (2013) の ECDM は, 計算コストを意識して開発された先行する方法論であり, δ に限らず共分散行列の関数に不偏推定量を導くことができ, 高い汎用性を有することが特徴である.

2.2 Σ^2 の不偏推定量

Σ^2 の不偏推定量は, ECDM を使えば次のように導かれる.

$$\widehat{\Sigma}^2 = \frac{u_n}{n(n-1)} \sum_{i<j}^{n} (\widehat{\boldsymbol{\Sigma}}_{ij(1)} \widehat{\boldsymbol{\Sigma}}_{ij(2)} + \widehat{\boldsymbol{\Sigma}}_{ij(2)} \widehat{\boldsymbol{\Sigma}}_{ij(1)}).$$

ただし，$u_n = n_{(1)}n_{(2)}/\{(n_{(1)}-1)(n_{(2)}-1)\}$,

$$\widehat{\boldsymbol{\Sigma}}_{ij(1)} = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{(1)(i+j)})(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{(1)(i+j)})^T,$$
$$\widehat{\boldsymbol{\Sigma}}_{ij(2)} = (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_{(2)(i+j)})(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_{(2)(i+j)})^T$$

である．このとき，$E(\widehat{\boldsymbol{\Sigma}}^2) = \boldsymbol{\Sigma}^2$ となる．

2.3 $\mathrm{tr}(\boldsymbol{\Sigma})^2$ の不偏推定量

$\mathrm{tr}(\boldsymbol{\Sigma})^2$ $(=\sigma)$ の不偏推定量は，ECDM を使えば次のように導かれる．

$$\widehat{\sigma} = \frac{2u_n}{n(n-1)} \sum_{i<j}^{n} \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{ij(1)})\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{ij(2)}).$$

ここで，u_n, $\widehat{\boldsymbol{\Sigma}}_{ij(1)}$, $\widehat{\boldsymbol{\Sigma}}_{ij(2)}$ は，2.2 節と同じものである．このとき，$E(\widehat{\sigma}) = \sigma$ となる．

ちなみに，標本共分散行列 $\boldsymbol{S}_n$ による単純な推定量 $\mathrm{tr}(\boldsymbol{S}_n)^2$ を使った場合，(A-i) のもとで $E\{\mathrm{tr}(\boldsymbol{S}_n)^2/\sigma\} = 1+O\{\delta/(n\sigma)\}$ となる．つまり，δ/σ が大きいか n が小さいとき，$\mathrm{tr}(\boldsymbol{S}_n)^2$ は大きなバイアスをもつ．簡単なシミュレーション実験で検証する．母集団分布は $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ とし，$\boldsymbol{\Sigma} = 0.5\boldsymbol{I}_p + 0.5\boldsymbol{1}_p\boldsymbol{1}_p^T$ なる級内相関モデルを考える．ただし，$\boldsymbol{1}_p$ は $\boldsymbol{1}_p = (1, ..., 1)^T$ なる p 次ベクトルである．このとき，最大固有値は $\lambda_{\max}(\boldsymbol{\Sigma}) = 0.5(p+1)$ となり，$\mathrm{tr}(\boldsymbol{\Sigma}) = p$ となる．よって，$\delta/\sigma \geq \lambda_{\max}(\boldsymbol{\Sigma})^2/\sigma \geq 0.25$ となり，$p \to \infty$ でも δ/σ は 0 に収束しない．いま，$p = 2^s$, $n = 2s$, $s = 2, ..., 8$ と設定する．図 1 は $\widehat{\sigma}/\sigma$ と $\mathrm{tr}(\boldsymbol{S}_n)^2/\sigma$ について，2000 回のシミュレーションによる平均と，さらに，1 との平均二乗誤差を $\mathrm{MSE}(\widehat{\sigma}/\sigma)$ と $\mathrm{MSE}(\mathrm{tr}(\boldsymbol{S}_n)^2/\sigma)$ について与えている．この実験に関して，$\widehat{\sigma}$ は次元数 p が小さい場合にも有効に機能していることが分かる．割愛するが，他の設定でも同様に，ECDM の性能が確認できる．

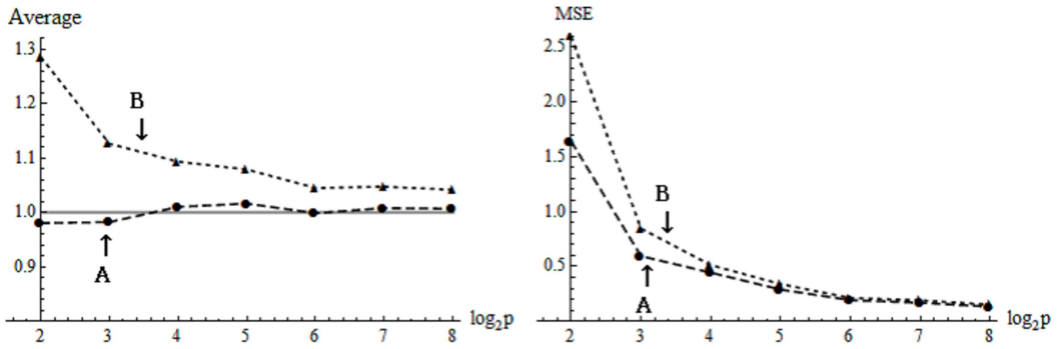

図 1　A: $\widehat{\sigma}/\sigma$ と B: $\mathrm{tr}(\boldsymbol{S}_n)^2/\sigma$ について，2000 回のシミュレーションによる平均 (左図)，1 との平均二乗誤差 A: $\mathrm{MSE}(\widehat{\sigma}/\sigma)$，B: $\mathrm{MSE}(\mathrm{tr}(\boldsymbol{S}_n)^2/\sigma)$ の結果 (右図)．

3 共分散構造に関する不偏推定量

各$\boldsymbol{x}$をp_1次ベクトル$\boldsymbol{x}_1$と$p_2\,(=p-p_1)$次ベクトル$\boldsymbol{x}_2$に分割し，$\boldsymbol{x}_j^T=(\boldsymbol{x}_{1j}^T,\boldsymbol{x}_{2j}^T)$, $j=1,...,n$と表記する．各iで，平均ベクトルを$E(\boldsymbol{x}_{ij})=\boldsymbol{\mu}_i$，共分散行列を$\mathrm{Var}(\boldsymbol{x}_{ij})=\boldsymbol{\Sigma}_i$とおき，相互共分散行列を$\mathrm{Cov}(\boldsymbol{x}_{1j},\boldsymbol{x}_{2j})=\boldsymbol{\Sigma}_*$とおく．すなわち，$\boldsymbol{\mu}=(\boldsymbol{\mu}_1^T,\boldsymbol{\mu}_2^T)^T$,

$$\boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_* \\ \boldsymbol{\Sigma}_*^T & \boldsymbol{\Sigma}_2\end{pmatrix}$$

と表記する．本節では，共分散構造に関する不偏推定量をECDMによって導く．

3.1 相互共分散行列に関する不偏推定量

$||\boldsymbol{\Sigma}_*||_F^2=\mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}_*^T)\ (=\kappa)$の推定量を考える．$\boldsymbol{S}_*=\sum_{j=1}^n(\boldsymbol{x}_{1j}-\overline{\boldsymbol{x}}_{1n})(\boldsymbol{x}_{2j}-\overline{\boldsymbol{x}}_{2n})^T/(n-1)$, $\overline{\boldsymbol{x}}_{in}=n^{-1}\sum_{j=1}^n\boldsymbol{x}_{ij}$, $i=1,2$を用いた単純な推定量$\mathrm{tr}(\boldsymbol{S}_*\boldsymbol{S}_*^T)$は，(A-i)のもと

$$E\{\mathrm{tr}(\boldsymbol{S}_*\boldsymbol{S}_*^T)\}=\kappa+O\Big(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_1)\mathrm{tr}(\boldsymbol{\Sigma}_2)}{n}\Big)$$

となる．これは，高次元データに対して非常に大きなバイアスをもつため，役に立たない．Srivastava and Reid (2012)は，母集団に正規分布を仮定して

$$\widehat{\kappa}_{SR}=\frac{(n-1)^2}{(n-2)(n+1)}\Big(\mathrm{tr}(\boldsymbol{S}_*\boldsymbol{S}_*^T)-\frac{\mathrm{tr}(\boldsymbol{S}_{1n})\mathrm{tr}(\boldsymbol{S}_{2n})}{n-1}\Big)$$

なるκの推定量を考えた．ただし，$\boldsymbol{S}_{in}$は$\boldsymbol{x}_i$の標本共分散行列である．母集団に正規分布を仮定できれば$E(\widehat{\kappa}_{SR})=\kappa$となるが，母集団に正規分布を仮定できないと$\widehat{\kappa}_{SR}$は高次元において非常に大きなバイアスが生じる．

ECDMを使えば，κの不偏推定量が次のように導かれる．各$k\ (=3,...,2n-1)$で2分割した集合について，$\boldsymbol{x}_{1j}$と$\boldsymbol{x}_{2j}$の標本平均を

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

で求め，ECDMによるκの推定量を

$$\widehat{\kappa}=\frac{2u_n}{n(n-1)}\sum_{i<j}^{n}(\boldsymbol{x}_{1i}-\overline{\boldsymbol{x}}_{1(1)(i+j)})^T(\boldsymbol{x}_{1j}-\overline{\boldsymbol{x}}_{1(2)(i+j)})$$
$$\times(\boldsymbol{x}_{2i}-\overline{\boldsymbol{x}}_{2(1)(i+j)})^T(\boldsymbol{x}_{2j}-\overline{\boldsymbol{x}}_{2(2)(i+j)})$$

と定義する．このとき，$\widehat{\kappa}$は母集団の分布型に依らずに不偏性$E(\widehat{\kappa})=\kappa$を主張できる．いま，$\boldsymbol{\Gamma}=(\boldsymbol{\Gamma}_1^T,\boldsymbol{\Gamma}_2^T)^T$，$\boldsymbol{\Gamma}_i=(\boldsymbol{\gamma}_{i1},...,\boldsymbol{\gamma}_{ir})$, $i=1,2$とおく．Yata and Aoshima (2015)から，以下の結果を得る．

補題1. 母集団分布に (A-i) を仮定する．$n \to \infty$，かつ，$p \to \infty$ もしくは $p < \infty$ のとき，次が成り立つ．

$$\mathrm{Var}(\widehat{\kappa}) = \Big\{4\frac{\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_*^T) + \mathrm{tr}\{(\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}_*^T)^2\} + \sum_{i=1}^r (M_i-2)(\boldsymbol{\gamma}_{1i}^T\boldsymbol{\Sigma}_*\boldsymbol{\gamma}_{2i})^2}{n} + 2\frac{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\mathrm{tr}(\boldsymbol{\Sigma}_2^2) + \kappa^2}{n^2}\Big\}\{1+o(1)\} + O\Big(\frac{\{\mathrm{tr}(\boldsymbol{\Sigma}_1^4)\mathrm{tr}(\boldsymbol{\Sigma}_2^4)\}^{1/2}}{n^2}\Big).$$

定理2. 母集団分布に (A-i) を仮定する．さらに，次を仮定する．

(A-ii) $\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\mathrm{tr}(\boldsymbol{\Sigma}_2^2)}{n^2\kappa^2} \to 0$, $n \to \infty$，かつ，$p \to \infty$ もしくは $p < \infty$.

そのとき，$n \to \infty$，かつ，$p \to \infty$ もしくは $p < \infty$ において，次が成り立つ．

$$\frac{\widehat{\kappa}}{\kappa} = 1 + o_P(1).$$

注意3. Yata and Aoshima (2015) は $\widehat{\kappa}$ の漸近正規性も示し，無相関性の仮説

$$H_0 : \mathrm{Corr}(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}) = \boldsymbol{O} \quad \text{vs.} \quad H_1 : \mathrm{Corr}(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}) \neq \boldsymbol{O}.$$

に対する検定方式を構築した．

注意4. $\widehat{\kappa}$ の計算アルゴリズム (Mathematica code) を次の通りである．

Input: Sample size n; $n \times p_i$ data matrices $X[i\,]$, $i = 1, 2$, such as $X[i\,] = (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in})^T$.

Mathematica code:

- $n1$ =Ceiling[$n/2$]; $n2 = n - n1$; $u = 2 * n1 * n2/((n1-1) * (n2-1) * n * (n-1))$
- V[1, $k\_$, $X\_$] :=If [Floor[$k/2$] $\geq n1$, Take[X, {Floor[$k/2$] $- n1 + 1$, Floor[$k/2$]}], Join[Take[X, {1, Floor[$k/2$]}], Take[X, {Floor[$k/2$] $+ n2 + 1, n$}]]]
- V[2, $k\_$, $X\_$] :=If [Floor[$k/2$] $\leq n1$, Take[X, {Floor[$k/2$] $+ 1$, Floor[$k/2$] $+ n2$}], Join[Take[X, {1, Floor[$k/2$] $- n1$}, Take[X, {Floor[$k/2$] $+ 1, n$}]]]
- Do[M[i, j, k] =Mean[$V[j, k, X[i]\,]$], {$k, 3, 2 * n - 1$}, {$i, 1, 2$}, {$j, 1, 2$}]
- $T = u$*Sum[(Part[$X[1], i$]−M[$1, 1, i + j$]).(Part[$X[1], j$]−M[$1, 2, i + j$]) *(Part[$X[2], i$]−M[$2, 1, i + j$]).(Part[$X[2], j$]−M[$2, 2, i + j$]), {$j, 2, n$}, {$i, 1, j - 1$}]

そのとき，$\widehat{\kappa} = T$ を得る．

注意5. $\widehat{\kappa}$ 以外にも，例えば $\boldsymbol{Y}_i = (\boldsymbol{y}_{i1}, ..., \boldsymbol{y}_{in})$, $\boldsymbol{y}_{ij} = \boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in}$, $j = 1, ..., n$; $i = 1, 2$ とおき，$\widehat{\kappa}_* = (\widehat{\delta}_C(\boldsymbol{Y}) - \sum_{i=1}^2 \widehat{\delta}_C(\boldsymbol{Y}_i))/2$ とすれば，$E(\widehat{\kappa}_*) = \kappa$ となる．そのとき，$\widehat{\kappa}_*$ の漸近分散は $\widehat{\kappa}$ と同等である．

3.2 共分散行列 Σ_i に関する不偏推定量

共分散行列 Σ_i に関する不偏推定量を考える．まず，$\mathrm{tr}(\Sigma_i^2)$ $(=\delta_i)$，$i=1,2$ の不偏推定量は，2.1 節と同様に ECDM を用いれば

$$\widehat{\delta}_i = \frac{2u_n}{n(n-1)} \sum_{s<t}^{n} \{(\boldsymbol{x}_{is} - \overline{\boldsymbol{x}}_{i(1)(s+t)})^T (\boldsymbol{x}_{it} - \overline{\boldsymbol{x}}_{i(2)(s+t)})\}^2, \quad i=1,2$$

で与えられる．そのとき，$E(\widehat{\delta}_i)=\delta_i$，$i=1,2$ となる．

注意 6. $\widehat{\kappa} = (\widehat{\delta} - \sum_{i=1}^{2} \widehat{\delta}_i)/2$ と表記できる．

また，$\mathrm{tr}(\Sigma_1)\mathrm{tr}(\Sigma_2)$ $(=\tau)$ の不偏推定量は，ECDM を使って次のように導かれる．

$$\widehat{\tau} = \frac{u_n}{n(n-1)} \sum_{i<j}^{n} (\widehat{\tau}_{ij(1)} + \widehat{\tau}_{ij(2)}).$$

ただし，$i<j$ について

$$\widehat{\tau}_{ij(1)} = (\boldsymbol{x}_{1i} - \overline{\boldsymbol{x}}_{1(1)(i+j)})^T (\boldsymbol{x}_{1i} - \overline{\boldsymbol{x}}_{1(1)(i+j)})(\boldsymbol{x}_{2j} - \overline{\boldsymbol{x}}_{2(2)(i+j)})^T (\boldsymbol{x}_{2j} - \overline{\boldsymbol{x}}_{2(2)(i+j)})$$
$$\widehat{\tau}_{ij(2)} = (\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_{2(1)(i+j)})^T (\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_{2(1)(i+j)})(\boldsymbol{x}_{1j} - \overline{\boldsymbol{x}}_{1(2)(i+j)})^T (\boldsymbol{x}_{1j} - \overline{\boldsymbol{x}}_{1(2)(i+j)})$$

である．そのとき，$E(\widehat{\tau})=\tau$ となる．

3.3 RV 係数

Robert and Escoufier (1974) 等で与えられる RV 係数について推定を考える．真の RV 係数を $\rho = \kappa/(\delta_1\delta_2)^{1/2}$ とおく．ただし，$\rho \in [0,1]$ である．RV 係数は，相関係数を多次元に拡張した統計量となっており，高次元の枠組みで重要な指標となる．ECDM を用いて，3.1 節で与えた $\widehat{\kappa}$ と 3.2 節で与えた $\widehat{\delta}_1$, $\widehat{\delta}_2$ に基づいて，$\widehat{\rho} = \widehat{\kappa}/(\widehat{\delta}_1\widehat{\delta}_2)^{1/2}$ を定義する．そのとき，Yata and Aoshima (2015) から次の結果を得る．

系 1. 母集団分布に (A-i) を仮定する．$n \to \infty$，かつ，$p \to \infty$ もしくは $p < \infty$ のとき，次が成り立つ．

$$\widehat{\rho} = \rho + O_P(n^{-1/2}).$$

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」，学術研究助成基金助成金 挑戦的萌芽研究 26540010 研究代表者: 青嶋 誠「ビッグデータの統計学: 理論の開拓と 3V への挑戦」，および，若手研究 (B) 26800078 研究代表者: 矢田 和善「高次元漸近理論の統一的研究」から研究助成を受けています.

参考文献

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

[2] M. Aoshima, K. Yata, Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions, Methodol. Comput. Appl. Probab. (2013), in press. doi: 10.1007/s11009-013-9370-7.

[3] 青嶋　誠, 矢田和善, 日本統計学会研究業績賞受賞者特別寄稿論文：高次元データの統計的方法論, 日本統計学会誌 **43** (2013), 123-150.

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

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

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

[7] P. Robert, Y. Escoufier, A unifying tool for linear multivariate statistical methods: the RV-coefficient, J. R. Stat. Soc. Ser. C 25 (1976) 257-265.

[8] M.S. Srivastava, Some tests concerning the covariance matrix in high dimensional data, J. Japan Statist. Soc. 35 (2005) 251-272.

[9] M.S. Srivastava, N. Reid, Testing the structure of the covariance matrix with fewer observations than the dimension, J. Multivariate Anal. 112 (2012) 156-171.

[10] M.S. Srivastava, H. Yanagihara, T. Kubokawa, Tests for covariance matrices in high dimension with less sample size, J. Multivariate Anal. 130 (2014) 289-309.

[11] K. Yata, 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.

[12] K. Yata, M. Aoshima, Correlation tests for high-dimensional data using extended cross-data-matrix methodology, J. Multivariate Anal. 117 (2013) 313-331.

[13] K. Yata, M. Aoshima, High-dimensional inference on covariance structures via the extended cross-data-matrix methodology, submitted (2015). arXiv:1503.06492.