R パッケージ “Extended-Cross-Data-Matrix” マニュアル

(Last Modified: June 20, 2020)

1 パッケージの説明

与えられるデータ行列 $X = [x_1, \dots, x_n]$ に対して、拡張クロスデータ行列法 (Extended-Cross-Data-Matrix Methodology, 略して ECDM 法) による $\text{tr}(\Sigma^2)$ の推定、および、高次元無相関性の検定を実行する。ただし、 $x_1, \dots, x_n$ は、平均に p 次のベクトル μ 、共分散行列に p 次の非負定値対称行列 Σ をもつ母集団からの無作為標本とする。

1.1 $\text{tr}(\Sigma^2)$ の推定

データ行列 X に対して、ECDM 法による $\text{tr}(\Sigma^2)$ の推定を実行する。以下の関数を定義する。

$$W(X)$$

入力:

- “ X ”: $p \times n$ データ行列。ここで、 p はデータの次元数、 $n(\geq 4)$ は標本数。

出力:

- “ W_n ”: $\text{tr}(\Sigma^2)$ の推定量 W_n 。

なお、 W_n の詳細は、3 節を確認のこと。

1.2 高次元無相関性の検定

各 x_j を p_1 次ベクトル x_{1j} と $p_2 (= p - p_1)$ 次ベクトル x_{2j} に分割し¹、 $x_j^T = (x_{1j}^T, x_{2j}^T)$ とおき、各 i で $X_i = [x_{i1}, \dots, x_{in}]$ とする。 x_{1j} と x_{2j} の相関行列を $\text{Corr}(x_{1j}, x_{2j}) = P_*$ とおく。そのとき、データ行列 X_1 と X_2 に対して、ECDM 法による検定

$$H_0 : P_* = O \quad \text{vs.} \quad H_1 : P_* \neq O \quad (1)$$

を実行する。以下の関数を定義する。

$$\text{Tecdm}(X_1, X_2)$$

入力:

- “ X_i ”: $p_i \times n$ データ行列 ($i = 1, 2$)。ここで、 p_i はデータの次元数、 $n(\geq 4)$ は標本数。

出力:

- “TestStatistics”: 検定統計量 ($T_n/\hat{\delta}$)。
- “pvalue”: p 値。

¹ $p_1 = 1$ 、または、 $p_2 = 1$ の場合も含む。

- “Tn”: ECDM 法による Δ の不偏推定量 T_n .

- “delta”: T_n の標準偏差の推定量 $\hat{\delta}$.

上記の4つを, 上から順にリストとして出力する. なお, Δ と $\hat{\delta}$ の詳細は, 4節を確認のこと.

2 拡張クロスデータ行列法

Yata and Aoshima [5] は, 拡張クロスデータ行列法 (ECDM 法) という方法論を開発した. 2つの集合 $\mathcal{V}_{n(1)(k)}$, $\mathcal{V}_{n(2)(k)}$ ($k = 3, \dots, 2n - 1$) を次のように定義する.

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

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

ここで, $n_{(1)} = \lceil n/2 \rceil$, $n_{(2)} = n - n_{(1)}$, $\lfloor x \rfloor$ は x 以下の最大の整数, $\lceil x \rceil$ は x 以上の最小の整数を表す. そのとき, $k = 3, \dots, 2n - 1$ について, $\#\mathcal{V}_{n(l)(k)} = n_{(l)}$, $l = 1, 2$, $\mathcal{V}_{n(1)(k)} \cap \mathcal{V}_{n(2)(k)} = \emptyset$, $\mathcal{V}_{n(1)(k)} \cup \mathcal{V}_{n(2)(k)} = \{1, \dots, n\}$ となること, 及び, $j < k$ ($\leq n$) について

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

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. $\mathcal{V}_{n(1)(j+k)}$ と $\mathcal{V}_{n(2)(j+k)}$ 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1つ計算する. これを, $j < k$ ($\leq n$) のすべての組合せで繰り返し, 得られる不偏推定量の平均をとる. これが拡張クロスデータ行列法である.

拡張クロスデータ行列法は, 高次元統計解析の様々な場面で応用できる. 例えば, Aoshima and Yata [3, 4] は, 高次元における推定・検定・判別分析などの各種統計的推測を考え, それらの高次元統計量の構成に拡張クロスデータ行列法を応用した. さらに, Yata and Aoshima [5, 6] は, 高次元パスウェイ解析に拡張クロスデータ行列法を応用した.

3 拡張クロスデータ行列法による $\text{tr}(\Sigma^2)$ の推定

$\text{tr}(\Sigma^2)$ は, 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である. それゆえ, このパラメータの推定は重要である. 詳細は, 青嶋・矢田 [1, 2], Yata and Aoshima [5, 6] を参照のこと. 拡張クロスデータ行列法を使えば, $\text{tr}(\Sigma^2)$ の不偏推定量は次のように構築できる. 各 k ($= 3, \dots, 2n - 1$) で2分割した集合について, 標本平均を

$$\bar{\mathbf{x}}_{n(1)(k)} = \frac{1}{n_{(1)}} \sum_{j \in \mathcal{V}_{n(1)(k)}} \mathbf{x}_j, \quad \bar{\mathbf{x}}_{n(2)(k)} = \frac{1}{n_{(2)}} \sum_{j \in \mathcal{V}_{n(2)(k)}} \mathbf{x}_j$$

とし, ある j, k ($j < k$) について $\text{tr}(\Sigma^2)$ の不偏推定量

$$c_n \{ (\mathbf{x}_j - \bar{\mathbf{x}}_{n(1)(j+k)})^T (\mathbf{x}_k - \bar{\mathbf{x}}_{n(2)(j+k)}) \}^2$$

を計算する．ここで， $c_n = (n_{(1)} - 1)^{-1}(n_{(2)} - 1)^{-1}n_{(1)}n_{(2)}$ である．すべての組合せの平均をとり

$$W_n = \frac{2c_n}{n(n-1)} \sum_{j,k=1(j<k)}^n \{(\mathbf{x}_j - \bar{\mathbf{x}}_{n(1)(j+k)})^T(\mathbf{x}_k - \bar{\mathbf{x}}_{n(2)(j+k)})\}^2 \quad (2)$$

を定義する． W_n は，母集団分布に依らずに，不偏性 $E(W_n) = \text{tr}(\Sigma^2)$ をもつ．

[拡張クロスデータ行列法による $\text{tr}(\Sigma^2)$ の推定量 W_n の計算アルゴリズム]

(Step 1) $\bar{\mathbf{x}}_{n(l)(k)}$, $l = 1, 2$ を各 $k (= 3, \dots, 2n - 1)$ で計算する．

(Step 2) すべての j, k ($1 \leq j < k \leq n$) について (Step 1) で計算した $\bar{\mathbf{x}}_{n(l)(j+k)}$ を代入して $c_n \{(\mathbf{x}_j - \bar{\mathbf{x}}_{n(1)(j+k)})^T(\mathbf{x}_k - \bar{\mathbf{x}}_{n(2)(j+k)})\}^2$ を計算し，それらの平均をとった W_n で $\text{tr}(\Sigma^2)$ を推定する．

なお， W_n の漸近的性質は Aoshima and Yata [3]，Yata and Aoshima [6] を参照のこと．

4 拡張クロスデータ行列法による高次元無相関性の検定

第一種の過誤 $\alpha \in (0, 1/2)$ を設定し，高次元のもと検定 (1) を考える．なお，検定 (1) は $\mathbf{x}_{1j}$ と $\mathbf{x}_{2j}$ における偏相関に関する無相関性の検定にもなっている²．

いま， $\mathbf{x}_j^T = (\mathbf{x}_{1j}^T, \mathbf{x}_{2j}^T)$ について，各 i の共分散行列を $\text{Var}(\mathbf{x}_{ij}) = \Sigma_i$ とおき，相互共分散行列を $\text{Cov}(\mathbf{x}_{1j}, \mathbf{x}_{2j}) = \Sigma_*$ とおく．すなわち，

$$\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_* \\ \Sigma_*^T & \Sigma_2 \end{pmatrix}$$

と表記する．行列のフロベニウスノルムを用いて，

$$\Delta = \|\Sigma_*\|_F^2 = \text{tr}(\Sigma_*\Sigma_*^T)$$

とおく． H_0 のもと $\Delta = 0$ であることに注意する．拡張クロスデータ行列法を使えば， Δ の不偏推定量は次のように構築できる．各 $k (= 3, \dots, 2n - 1)$ で 2 分割した集合について，各 i で標本平均を

$$\bar{\mathbf{x}}_{i,n(1)(k)} = \frac{1}{n_{(1)}} \sum_{j \in \mathcal{V}_{n(1)(k)}} \mathbf{x}_{ij}, \quad \bar{\mathbf{x}}_{i,n(2)(k)} = \frac{1}{n_{(2)}} \sum_{j \in \mathcal{V}_{n(2)(k)}} \mathbf{x}_{ij}$$

とし，

$$T_n = \frac{2c_n}{n(n-1)} \sum_{j,k=1(j<k)}^n (\mathbf{x}_{1j} - \bar{\mathbf{x}}_{1,n(1)(j+k)})^T (\mathbf{x}_{1k} - \bar{\mathbf{x}}_{1,n(2)(j+k)}) \\ \times (\mathbf{x}_{2j} - \bar{\mathbf{x}}_{2,n(1)(j+k)})^T (\mathbf{x}_{2k} - \bar{\mathbf{x}}_{2,n(2)(j+k)}) \quad (3)$$

を定義する． T_n は，母集団分布に依らずに，不偏性 $E(T_n) = \Delta$ をもつ． T_n の分散の推定量として，

$$\hat{\delta}^2 = 2 \frac{W_{1,n}W_{2,n}}{n^2}$$

²Yata and Aoshima [6] の 1 節を参照のこと．

を考える．ここで， $W_{i,n}$ は (2) 式に従って計算する $\text{tr}(\Sigma_i^2)$ の不偏推定量である．高次元 ($p \rightarrow \infty$) のもと $T_n/\hat{\delta}$ は漸近的に正規分布に従う³．

[仮説 (1) の検定方式の構築]

(Step 1) $\bar{\mathbf{x}}_{i,n(l)(k)}$, $l = 1, 2$ を各 $k (= 3, \dots, 2n - 1)$ と $i (= 1, 2)$ で計算する．

(Step 2) すべての j, k ($1 \leq j < k \leq n$) について (Step 1) で計算した $\bar{\mathbf{x}}_{i,n(l)(j+k)}$ を代入して $c_n(\mathbf{x}_{1j} - \bar{\mathbf{x}}_{1,n(1)(j+k)})^T(\mathbf{x}_{1k} - \bar{\mathbf{x}}_{1,n(2)(j+k)})(\mathbf{x}_{2j} - \bar{\mathbf{x}}_{2,n(1)(j+k)})^T(\mathbf{x}_{2k} - \bar{\mathbf{x}}_{2,n(2)(j+k)})$ を計算し，それらの平均をとった T_n を求める．

(Step 3) W_n の計算アルゴリズムと同様に，各 i で $W_{i,n}$ を計算し， $\hat{\delta}$ を求める．

(Step 4) 標準正規分布の上側 α 点 z_α を用いて，仮説 (1) を次のように検定する⁴．

$$H_0 \text{ を棄却} \iff \frac{T_n}{\hat{\delta}} > z_\alpha \quad (4)$$

Yata and Aoshima [6] は，上記の検定方式を用いて，マイクロアレイデータのパスウェイ解析を行った．

References

- [1] 青嶋誠，矢田和善 (2013). 日本統計学会研究業績賞受賞者特別寄稿論文：高次元データの統計的方法論，日本統計学会誌，**43**, 123-150.
- [2] 青嶋誠，矢田和善 (2019). 高次元の統計学，共立出版，東京.
- [3] Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics*, **66**, 983-1010.
- [4] Aoshima, M. and Yata, K. (2015). Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions, *Methodology and Computing in Applied Probability*, **17**, 419-439.
- [5] Yata, K. and Aoshima, M. (2013). Correlation tests for high-dimensional data using extended cross-data-matrix methodology, *Journal of Multivariate Analysis*, **117**, 313-331.
- [6] Yata, K. and Aoshima, M. (2016). High-dimensional inference on covariance structures via the extended cross-data-matrix methodology, *Journal of Multivariate Analysis*, **151**, 151-166.

³Yata and Aoshima [6] の 3 節を参照のこと．

⁴本検定方式の検出力は，Yata and Aoshima [6] の 4 節を参照のこと．