

R パッケージ “Cross-Data-Matrix” マニュアル

(Last Modified: August 20, 2019)

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, クロスデータ行列法 (Cross-Data-Matrix methodology, 略して CDM 法) による固有値・固有ベクトル・主成分スコアの計算を実行する. 以下の関数を定義する.

CDM(X, split)

入力:

- “X”: $d \times n$ データ行列. ここで, d はデータの次元数, $n (\geq 4)$ は標本数.
- “split”: ‘True’ または ‘False’. ‘True’ の場合, X を無作為に 2 分割する. ‘False’ の場合, X を単純に $X_1 = [x_1, \dots, x_{n_{(1)}}]$, $X_2 = [x_{n_{(1)}+1}, \dots, x_n]$ と分割する. デフォルトは ‘False’ になっている.

出力:

- “values”: CDM 法による第 $(n_{(2)} - 1)$ 固有値までの計算結果.
(リストの i 番目は, i 番目に大きい固有値を表す.)
- “vectors”: CDM 法による第 $(n_{(2)} - 1)$ 固有ベクトルまでの計算結果.
(リストの i 列目は, i 番目の固有値に対応する固有ベクトルを表す.)
- “scores”: CDM 法による第 $(n_{(2)} - 1)$ 主成分スコアまでの計算結果.
(リストの (i, j) 成分は, 第 i 主成分について j 番目の標本に対するスコアを表す.)

2 クロスデータ行列法による固有値・固有ベクトル・主成分スコアの推定

母集団が, 未知の d 次平均ベクトル μ と, 未知の d 次共分散行列 Σ (非負定値対称行列) をもつとする. Σ の固有値を $\lambda_1 \geq \dots \geq \lambda_d (\geq 0)$ とし, 各固有値 λ_i に対する固有ベクトルを h_i とする. ここで, $h_1, \dots, h_d$ は正規直交基底をなすとする. 母集団から $n (\geq 4)$ 個の d 次データベクトル $x_1, \dots, x_n$ を無作為に抽出して, 大きさ $d \times n$ のデータ行列 $X = [x_1, \dots, x_n]$ を構成する. そのとき, 第 i 主成分スコアは, $j = 1, \dots, n$ に対して

$$h_i^T (x_j - \mu) (= s_{ij} \text{とおく})$$

で定義される.

データ行列 X を無作為に 2 つに分割して, $d \times n_{(l)}$ 部分行列

$$X_l = [x_{1,l}, \dots, x_{n_{(l)},l}], \quad l = 1, 2$$

を定義する¹. ここで, $n_{(1)} = \lceil n/2 \rceil$, $n_{(2)} = n - n_{(1)}$ であり, $\lceil x \rceil$ は x 以上の最小の整数を表す. 各 l について, 標本平均ベクトル $\bar{x}_l = n_{(l)}^{-1} \sum_{j=1}^{n_{(l)}} x_{j,l}$ を $n_{(l)}$ 個並べて, $\bar{X}_l = [\bar{x}_1, \dots, \bar{x}_{n_{(l)}}]$ とおく. そのとき, $n_{(1)} \times n_{(2)}$ 行列

$$S_{D(1)} = \frac{(X_1 - \bar{X}_1)^T (X_2 - \bar{X}_2)}{\sqrt{(n_{(1)} - 1)(n_{(2)} - 1)}}$$

¹単純に $X_1 = [x_1, \dots, x_{n_{(1)}}]$, $X_2 = [x_{n_{(1)}+1}, \dots, x_n]$ と分割してもよい.

をクロスデータ行列とよぶ．いま， $S_{D(1)}$ の特異値分解を

$$S_{D(1)} = \sum_{i=1}^{n_{(2)}-1} \lambda_i \dot{u}_{i,1} \dot{u}_{i,2}^T$$

とおく．ここで， $\lambda_1 \geq \dots \geq \lambda_{n_{(2)}-1} (\geq 0)$ は $S_{D(1)}$ の特異値， $\dot{u}_{i,1} = (\dot{u}_{i1,1}, \dots, \dot{u}_{in_{(1)},1})^T$ は左特異ベクトル， $\dot{u}_{i,2} = (\dot{u}_{i1,2}, \dots, \dot{u}_{in_{(2)},2})^T$ は右特異ベクトルである．Yata and Aoshima [5, 7] は，クロスデータ行列法という高次元主成分分析を考案した．クロスデータ行列法は，固有値・固有ベクトル・主成分スコアを，次のように推定する．

[クロスデータ行列法の計算アルゴリズム]

(Step 1) 入力されたデータ行列 X に対して， $S_{D(1)} = \{(n_{(1)}-1)(n_{(2)}-1)\}^{-1/2} (X_1 - \bar{X}_1)^T (X_2 - \bar{X}_2)$ を計算する．

(Step 2) $S_{D(1)}$ の特異値 $\lambda_1 \geq \dots \geq \lambda_{n_{(2)}-1} (\geq 0)$ と，対応する左特異ベクトル $\dot{u}_{i,1}$ ($i = 1, \dots, n_{(2)} - 1$) と右特異ベクトル $\dot{u}_{i,2}$ ($i = 1, \dots, n_{(2)} - 1$) を計算する．

(Step 3) 固有値 λ_i ($i = 1, \dots, n_{(2)} - 1$) を λ_i で推定する．

(Step 4) 固有ベクトル h_i ($i = 1, \dots, n_{(2)} - 1$) を $\dot{h}_i = \dot{h}_i / \|\dot{h}_i\|$ で推定する²．ここで，

$$\dot{h}_i = \frac{1}{2\sqrt{\lambda_i}} \left(\frac{(X_1 - \bar{X}_1)\dot{u}_{i,1}}{\sqrt{n_{(1)}-1}} + \frac{(X_2 - \bar{X}_2)\dot{u}_{i,2}}{\sqrt{n_{(2)}-1}} \right)$$

(Step 5) 第 i 主成分スコアを， $x_{j,l}$ ($j = 1, \dots, n_{(l)}$; $l = 1, 2$) に対して

$$\dot{s}_{ij,l} = \dot{u}_{ij,l} \sqrt{(n_{(l)}-1)\lambda_i}$$

で推定する．各 i で， $\dot{s}_{i1,1}, \dots, \dot{s}_{in_{(1)},1}$, $\dot{s}_{i1,2}, \dots, \dot{s}_{in_{(2)},2}$ に $\dot{s}_{ij}$, $j = 1, \dots, n$ という通し番号を付ける³．

3 クロスデータ行列法の応用例

クロスデータ行列法は，高次元データの次元削減だけでなく，様々な場面に応用できる．

- (1) Aoshima and Yata [1] 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クロスデータ行列法を応用した．
- (2) Aoshima and Yata [3] と Yata and Aoshima [6] は，高次元データの潜在空間の次元推定にクロスデータ行列法を応用した．
- (3) Yata and Aoshima [5] は，マイクロアレイデータのクラスター分析に，クロスデータ行列法を応用した．

² $\|\cdot\|$ はユークリッドノルムを表す．

³ 青嶋・矢田 [2, 4] で， λ_i , $\dot{h}_i$, $\dot{s}_{ij}$ の漸近的性質を解説している．

References

- [1] Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data, *Sequential Analysis (Editor's special invited paper)*, **30**, 356-399.
- [2] 青嶋 誠 , 矢田和善 (2013). 論説: 高次元小標本における統計的推測, *数学* , **65**, 225-247.
- [3] Aoshima, M. and Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models, *Statistica Sinica*, **28**, 43-62.
- [4] 青嶋 誠 , 矢田和善 (2019). 高次元の統計学 , 共立出版 , 東京.
- [5] Yata, K. and Aoshima, M. (2010a). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix, *Journal of Multivariate Analysis*, **101**, 2060-2077.
- [6] Yata, K. and Aoshima, M. (2010b). Intrinsic dimensionality estimation of high-dimension, low sample size data with d -asymptotics, *Communications in Statistics. Theory and Methods, Special Issue: Honoring Akahira, M. (ed. Aoshima, M.)*, **39**, 1511-1521.
- [7] Yata, K. and Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings, *Journal of Multivariate Analysis*, **122**, 334-354.