R パッケージ “Noise-Reduction” マニュアル

(Last Modified: August 9, 2019)

1　パッケージの説明

与えられるデータ行列 $\boldsymbol{X}$ に対して，ノイズ掃き出し法 (Noise-Reduction methodology, 略して NR 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．

NRM(X)

入力:

- “X”: $d \times n$ データ行列．ここで，d はデータの次元数，n (≥ 3) は標本数.

出力:

- “$values$”: NR 法による第 $(n-2)$ 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.）
- “$vectors$”: NR 法による第 $(n-2)$ 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.）
- “$scores$”: NR 法による第 $(n-2)$ 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　ノイズ掃き出し法による固有値・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 $\boldsymbol{\mu}$ と，未知の d 次共分散行列 $\boldsymbol{\Sigma}$（非負定値対称行列）をもつとする．$\boldsymbol{\Sigma}$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_d$ (≥ 0) とし，各固有値 λ_i に対する固有ベクトルを $\boldsymbol{h}_i$ とする．ここで，$\boldsymbol{h}_1, ..., \boldsymbol{h}_d$ は正規直交基底をなすとする．母集団から n (≥ 3) 個の d 次データベクトル $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ を無作為に抽出して，大きさ $d \times n$ のデータ行列 $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ を構成する．そのとき，第 i 主成分スコアは，$j = 1, ..., n$ に対して

$$\boldsymbol{h}_i^T(\boldsymbol{x}_j - \boldsymbol{\mu}) \ (= s_{ij} \text{とおく})$$

で定義される．

標本平均ベクトル $\overline{\boldsymbol{x}} = n^{-1}\sum_{j=1}^{n} \boldsymbol{x}_j$ を n 個並べて，$\overline{\boldsymbol{X}} = [\overline{\boldsymbol{x}}, ..., \overline{\boldsymbol{x}}]$ とおく．標本共分散行列を

$$\boldsymbol{S} = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^T$$

とする．$\boldsymbol{S}$ の固有値を $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$ (≥ 0) とし，各固有値 $\hat{\lambda}_i$ に対する固有ベクトルを $\hat{\boldsymbol{h}}_i$ とする．ここで，$\hat{\boldsymbol{h}}_1, ..., \hat{\boldsymbol{h}}_d$ は正規直交基底をなすとする．そのとき，$\boldsymbol{S}$ の双対標本共分散行列は

$$\boldsymbol{S}_D = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^T(\boldsymbol{X} - \overline{\boldsymbol{X}})$$

で与えられる．$\boldsymbol{S}_D$ は $\boldsymbol{S}$ と正の固有値を共有する．$\boldsymbol{S}_D$ の固有値 $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{n-1}$ (≥ 0) について，各固有値 $\hat{\lambda}_i$ に対する固有ベクトルを $\hat{\boldsymbol{u}}_i$ とする．ここで，$\hat{\boldsymbol{u}}_1, ..., \hat{\boldsymbol{u}}_{n-1}$ は互いに正規直交とする．そのとき，$\boldsymbol{S}$ と $\boldsymbol{S}_D$ の固有ベクトルには，次の双対関係がある．

$$\hat{\boldsymbol{h}}_i = \frac{\boldsymbol{X} - \overline{\boldsymbol{X}}}{\sqrt{(n-1)\hat{\lambda}_i}}\hat{\boldsymbol{u}}_i \quad (i = 1, ..., \min\{d, n-1\})$$

これらを用いると，第 i 主成分スコアは次のように推定される．

$$\hat{\boldsymbol{h}}_i^T(\boldsymbol{x}_j - \overline{\boldsymbol{x}}) = \hat{u}_{ij}\sqrt{(n-1)\hat{\lambda}_i}\ (= \hat{s}_{ij}\text{とおく})$$

ここで，$\hat{\boldsymbol{u}}_i = (\hat{u}_{i1}, ..., \hat{u}_{in})^T$ である．

多変量解析における主成分分析 (PCA) は，$\boldsymbol{\Sigma}$ の固有値・固有ベクトルと主成分スコアを，$\hat{\lambda}_i$，$\hat{\boldsymbol{h}}_i$，$\hat{s}_{ij}$ で推定する．高次元データを扱う場合，$d >> n$ になると，これらの推定は次元の呪いを受けて間違った解析結果をもたらす．高次元統計解析には，新しい PCA が必要となる．Yata and Aoshima [7, 8] は，$\boldsymbol{S}_D$ の幾何学的表現に基づいて，ノイズ掃き出し法という高次元 PCA を考案した．高次元統計解析について，青嶋 [1]，青嶋・矢田 [2]，Aoshima et al. [3] を参照のこと．

ノイズ掃き出し法は，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．

[**ノイズ掃き出し法の計算アルゴリズム**]

(Step 1) 入力されたデータ行列 $\boldsymbol{X}$ に対して，$\boldsymbol{S}_D = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^T(\boldsymbol{X} - \overline{\boldsymbol{X}})$ を計算する．

(Step 2) $\boldsymbol{S}_D$ の固有値 $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{n-2}\ (\geq 0)$ と，対応する固有ベクトル $\hat{\boldsymbol{u}}_i\ (i = 1, ..., n-2)$ を計算する．

(Step 3) 固有値 $\lambda_i\ (i = 1, ..., n-2)$ を，次のように推定する．

$$\tilde{\lambda}_i = \hat{\lambda}_i - \frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{j=1}^{i}\hat{\lambda}_j}{n-1-i}$$

(Step 4) 固有ベクトル $\boldsymbol{h}_i\ (i = 1, ..., n-2)$ を，次のように推定する[1]．

$$\tilde{\boldsymbol{h}}_i = \frac{\boldsymbol{X} - \overline{\boldsymbol{X}}}{\sqrt{(n-1)\tilde{\lambda}_i}}\hat{\boldsymbol{u}}_i$$

(Step 5) 第 i 主成分スコアを，次のように推定する[2]．

$$\tilde{s}_{ij} = \hat{u}_{ij}\sqrt{(n-1)\tilde{\lambda}_i}$$

3　ノイズ掃き出し法の応用例

ノイズ掃き出し法は，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) プリンストン大学の Fan 教授らは，Wang and Fan [6] において高次元共分散行列の推定を考え，ノイズ掃き出し法とスパースモデリングを融合させた S-POET を提案した．

(2) Ishii, Yata and Aoshima [4, 5] は，ノイズ掃き出し法を応用した高次元共分散行列の同等性検定を提案した．Ishii, Yata and Aoshima [4] は，ノイズ掃き出し法を応用して寄与率の信頼区間を構築した．

(3) Yata and Aoshima [9] は，高次元低ランク行列を扱い，ノイズ掃き出し法を応用した高次元信号行列の推定を提案した．

---

[1] $\tilde{\boldsymbol{h}}_i = \hat{\boldsymbol{h}}_i\sqrt{\hat{\lambda}_i/\tilde{\lambda}_i}$ のように計算してもよい．

[2] 青嶋・矢田 [2] の 4 章で，$\tilde{\lambda}_i$，$\tilde{\boldsymbol{h}}_i$，$\tilde{s}_{ij}$ の漸近的性質を解説している．

References

[1] 青嶋 誠 (2018). 日本統計学会賞受賞者特別寄稿論文: 高次元統計解析: 理論と方法論の新しい展開, 日本統計学会誌, **48**, 89-111.

[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., Shen, D., Shen, H., Yata, K., Zhou, Y.-H. and Marron, J. S. (2018). A survey of high dimension low sample size asymptotics, *Australian and New Zealand Journal of Statistics, Special Issue: in Honour of Peter Gavin Hall*, **60**, 4-19.

[4] Ishii, A., Yata, K. and Aoshima, M. (2016). Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context, *Journal of Statistical Planning and Inference*, **170**, 186-199.

[5] Ishii, A., Yata, K. and Aoshima, M. (2019). Equality tests of high-dimensional covariance matrices under the strongly spiked eigenvalue model, *Journal of Statistical Planning and Inference*, **202**, 99-111.

[6] Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance, *The Annals of Statistics*, **45**, 1342-1374.

[7] Yata, K. and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations, *Journal of Multivariate Analysis*, **105**, 193-215.

[8] Yata, K. and Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings, *Journal of Multivariate Analysis*, **122**, 334-354.

[9] Yata, K. and Aoshima, M. (2016). Reconstruction of a high-dimensional low-rank matrix, *Electronic Journal of Statistics*, **10**, 895-917.