スピアマンランク行列による主成分分析のロバストネス

千葉大・融合理工学府　渡邊 宏大
千葉大・理学研究院　内藤 貫太

はじめに: ロバストな主成分分析については，これまで多くの手法が提案されている．本発表では，Marden (1999) の Rank から派生したスピアマンランク行列に基づく主成分分析のロバスト性について報告する．Han and Liu (2018) で議論されている Kendall's tau との比較についても考察する．

設定と定義: $\boldsymbol{y} \in \mathbb{R}^d$ の Spatial sign を

$$S(\boldsymbol{y}) = \begin{cases} \dfrac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2} & , \quad \boldsymbol{y} \neq \boldsymbol{0} \\ \boldsymbol{0} & , \quad \boldsymbol{y} = \boldsymbol{0} \end{cases}$$

で定義する．$\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ を有限な分散共分散行列 Σ を持つ d 次元確率分布 F にしたがう互いに独立な確率ベクトルとする．Spatial sign に基づくロバストな主成分分析の手法が，Marden (1999)，Han and Liu (2018) で議論されている．

母集団での Kendall's tau matrix $\boldsymbol{K} \in \mathbb{R}^{d \times d}$ は

$$\boldsymbol{K} = E_{\boldsymbol{X},\boldsymbol{Y}} \left[S(\boldsymbol{X} - \boldsymbol{Y}) S(\boldsymbol{X} - \boldsymbol{Y})^T \right]$$

で定義され，その推定量は

$$\widehat{\boldsymbol{K}} = \frac{2}{n(n-1)} \sum_{j<i} S(\boldsymbol{X}_i - \boldsymbol{X}_j) S(\boldsymbol{X}_i - \boldsymbol{X}_j)^T$$

で定義される．一方，母集団でのスピアマンランク行列 $\Sigma_R \in \mathbb{R}^{d \times d}$ は

$$\Sigma_R = E_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}} \left[S(\boldsymbol{X} - \boldsymbol{Y}) S(\boldsymbol{X} - \boldsymbol{Z})^T \right] = E_{\boldsymbol{X}} \left[E_{\boldsymbol{Y}} \left[S(\boldsymbol{X} - \boldsymbol{Y}) \right] E_{\boldsymbol{Z}} \left[S(\boldsymbol{X} - \boldsymbol{Z}) \right]^T \right]$$

で定義され，その推定量は

$$\widehat{\Sigma}_R = \frac{1}{n(n-1)(n-2)} \sum_{i=1}^{n} \sum_{j \neq i, k \neq i, j \neq k} S(\boldsymbol{X}_i - \boldsymbol{X}_j) S(\boldsymbol{X}_i - \boldsymbol{X}_k)^T$$

で定義される．

問題: 重要な事実として，F が楕円分布のとき，Σ，$\boldsymbol{K}$ および Σ_R は同じ直交行列で対角化されることが知られている (Marden (1999)，Han and Liu (2018) を参照)．Marden (1999) においても，Σ_R の推定量 $\widetilde{\Sigma}_R$ によるロバストな主成分分析が議論されているが，その $\widetilde{\Sigma}_R$ は

$$\widetilde{\Sigma}_R = \frac{1}{n} \widehat{\boldsymbol{K}} + \left(1 - \frac{2}{n} \right) \widehat{\Sigma}_R$$

と分解される．$n \to \infty$ のとき，$\widehat{\boldsymbol{K}}$ は $\boldsymbol{K}$ に，$\widehat{\Sigma}_R$ は Σ_R に確率収束することが示される．このことから，$\widetilde{\Sigma}_R$ と $\widehat{\boldsymbol{K}}$ の比較は，漸近的には $\widehat{\Sigma}_R$ と $\widehat{\boldsymbol{K}}$ の比較となる．我々の問題は特に，$\widehat{\Sigma}_R$ と $\widehat{\boldsymbol{K}}$ はどちらがロバストな主成分分析を提供するのか？ということになる．

影響関数: ロバストネスの指標として影響関数を考える (影響関数については，Hampel et al. (1986) を参照)．F を楕円分布とする．比較の対象として，$\boldsymbol{K}$ と Σ_R の最大固有値に関する影響関数および固有ベクトルに関する影響関数を与える．

$\boldsymbol{K}$ は直交行列 $\Gamma = [\boldsymbol{\gamma}_1 \cdots \boldsymbol{\gamma}_d]$ と対角行列 $A = \mathrm{diag}(a_1, \ldots, a_d)$，$a_1 > \cdots > a_d > 0$ によって $\boldsymbol{K} = \Gamma A \Gamma^T$ とスペクトル分解されるとする．上で述べた重要な事実から，Σ_R も同じ直交行列によって $\Gamma \Lambda \Gamma^T$，$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ とスペクトル分解できる．ただし，対応する固有値については，$\lambda_1 > \cdots > \lambda_d > 0$ となるかは定かではない．したがって，$\boldsymbol{\gamma}_1$ が属している固有値を λ_M，$M \in \{1, \ldots, d\}$，Σ_R の最大固有値と対応する固有ベクトルをそれぞれ λ_{max}，$\boldsymbol{\gamma}_L$，$L \in \{1, \ldots, d\}$ とする．このとき，$\boldsymbol{K}$ と Σ_R の最大固有値に関する影響関数 $IF_{\boldsymbol{K}}(\boldsymbol{x}, a_1, F)$，$IF_{\Sigma_R}(\boldsymbol{x}, \lambda_{max}, F)$ はそれぞれ

$$
\begin{aligned}
IF_{\boldsymbol{K}}(\boldsymbol{x}, a_1, F) &= \boldsymbol{\gamma}_1^T A(\boldsymbol{x}) \boldsymbol{\gamma}_1 - 2a_1, \\
IF_{\Sigma_R}(\boldsymbol{x}, \lambda_{max}, F) &= \boldsymbol{\gamma}_L^T B(\boldsymbol{x}) \boldsymbol{\gamma}_L - 3\lambda_{max}
\end{aligned}
$$

で与えられる．ここで,

$$
\begin{aligned}
A(\boldsymbol{x}) &= 2E_{\boldsymbol{X}} \left[S(\boldsymbol{x} - \boldsymbol{X}) S(\boldsymbol{x} - \boldsymbol{X})^T \right], \\
B(\boldsymbol{x}) &= E_{\boldsymbol{X}, \boldsymbol{Y}} \left[S(\boldsymbol{X} - \boldsymbol{Y}) S(\boldsymbol{X} - \boldsymbol{x})^T + S(\boldsymbol{X} - \boldsymbol{x}) S(\boldsymbol{X} - \boldsymbol{Y})^T + S(\boldsymbol{x} - \boldsymbol{X}) S(\boldsymbol{x} - \boldsymbol{Y})^T \right]
\end{aligned}
$$

である．また，$\boldsymbol{K}$ と Σ_R の固有ベクトル $\boldsymbol{\gamma}_1$ に関する影響関数 $IF_{\boldsymbol{K}}(\boldsymbol{x}, \boldsymbol{\gamma}_1, F)$，$IF_{\Sigma_R}(\boldsymbol{x}, \boldsymbol{\gamma}_1, F)$ はそれぞれ

$$
\begin{aligned}
IF_{\boldsymbol{K}}(\boldsymbol{x}, \boldsymbol{\gamma}_1, F) &= \sum_{k=2}^{d} \frac{1}{a_1 - a_k} \boldsymbol{\gamma}_k^T A(\boldsymbol{x}) \boldsymbol{\gamma}_1 \cdot \boldsymbol{\gamma}_k, \\
IF_{\Sigma_R}(\boldsymbol{x}, \boldsymbol{\gamma}_1, F) &= \sum_{k \neq M} \frac{1}{\lambda_M - \lambda_k} \boldsymbol{\gamma}_k^T B(\boldsymbol{x}) \boldsymbol{\gamma}_1 \cdot \boldsymbol{\gamma}_k
\end{aligned}
$$

で与えられる．

適用例: 影響関数を用いた固有ベクトルのロバストネスの比較，シミュレーション結果と実データへの適用結果についてはシンポジウムにて報告する．

参考文献

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics : The Approach Based on Influence Functions*. Wiley.

Han, F. and Liu, H. (2018). Eca: High-dimensional elliptical component analysis in non-gaussian distributions. *Journal of the American Statistical Association*, 113:252–268.

Marden, J. I. (1999). Some robust estimates of principal components. *Statistics & Probability Letters*, 43:349–359.