# データ行列の類似度に基づく主成分数の選択
# Selection of the number of principal components based on similarity between data matrices

数学専攻　保科　架風
HOSHINA, Ibuki

## 1　研究目的

情報工学, 金融工学, 遺伝子工学, 環境学, 社会学などの広い分野で, 多変量解析は大規模データの要約や表現に用いられている. その中で主成分分析（Principal Component Analysis; PCA）は, 次元削減やノイズ除去などのデータの下処理などの場面で有用かつ標準的な解析方法である.

主成分分析の目的は, データ空間の元の軸よりも効率的にデータの特徴を説明するような, 新しい軸（主成分; 主成分軸; 主成分ベクトル）を変数の線形結合によって再構成することにある. これにより, データ空間の次元よりも少ない数の軸でデータの持つ情報の大部分を保有することが可能となる.

さて, この主成分分析の重要なステップに「主成分数の選択」がある. これは, いくつの主成分によって元のデータを近似するのか, いくつの主成分に元のデータを圧縮するのか, を決定するということである. このための方法として, 統計的推測を用いた方法や記述的なものなど, これまで様々な方法が提案されてきた. 前者の例としては, データの構造にモデルを仮定した上での仮説検定, 一方後者の例としては, 近似によって保有される（近似によって失われる）情報の量を評価する「累積寄与率」が十分大きくなるような主成分数を選ぶ, というものがある.

しかしながら, これらの方法には様々な問題がある. まず, 仮説検定での主成分数を選択では, 設定される仮説が必ずしも情報の量を評価していない. また, 累積寄与率では, 例えば近年開発されたカーネル主成分分析, スパース主成分分析による圧縮と主成分分析に圧縮を直接比較することができない.

そこで本研究では, 主成分分析, カーネル主成分分析のどちらでも情報の量が評価可能であり, また, データのバラつきを考慮に入れて主成分数を選択する方法の提案を目的とする.

## 2　研究内容

カーネル主成分分析とは, データを非線形写像によって特徴空間と呼ばれる空間に写し, その空間内で主成分分析を行い, 結果を元の空間に戻す手法である. したがって, カーネル主成分分析での主成分数の選択に累積寄与率を用いれば, 評価できるのは特徴空間での圧縮行列の保有する情報の量である. よって入力空間での圧縮の精度は評価できず, 主成分数の選択に用いることは適していない.

そこで, 本研究では元のデータ行列と圧縮行列の類似度によって圧縮行列に含まれる情報の量を評価するということを提案する. これにより, 圧縮の際に使用された方法に関わらず情報の量を評価することが可能となり, カーネル主成分分析においても主成分数の選択に用いることが出来る. また, カーネル主成分分析におけるカーネル関数のパラメータ設定においても, 類似度に基づく情報の量の評価は適用可能である.

この行列の類似度を測る指標として Escoufier [1], Robert and Escoufier [3] らで提案されている RV 係数 (RV-coefficient) が挙げられる.

**定義 2.1 (RV 係数)**
$X : n \times p$ を観測次元 $p$ の $n$ 個の個体からなる中心化データ行列とし, $Y : n \times q$ を $X$ と同じ $n$ 個の個体から

の観測次元 $q$ の中心化データ行列とする. このとき, $X$ と $Y$ の間の類似度を測る $RV$ 係数を以下で定める.

$$\mathrm{RV}(X,Y)=\frac{\mathrm{tr}(X'YY'X)}{\sqrt{\mathrm{tr}(X'XX'X)\mathrm{tr}(Y'YY'Y)}}$$

この RV 係数でデータ行列 $X$ と $k$ 主成分による圧縮行列 $X_{(k)}$ を測れば, 第 $k$ 主成分へ圧縮したときの個体間の配置が元の個体間の配置にどのぐらい近いのかを評価することができる. 圧縮したとき, 各個体がもともと持っている情報の多くを保有しているのであれば, 他の個体との配置関係に大きな違いは生じず, 保有する情報の量が少ないのであれば配置関係にも差が生じる. すなわち, $X_{(k)}$ が $X$ が持つ情報の多くを保有するとき, $X$ と $X_{(k)}$ の間の RV 係数の値は大きくなり, そうでないときは小さくなる. したがって, RV 係数で類似度を測ることは情報の量を評価することと等しい. なお, $X$ と $X_{(k)}$ の類似度を RV 係数で測ったものを RV 統計量 $R_k$ として定義する.

**定義 2.2 (RV 統計量)**

$$R_k=\mathrm{RV}(X,X_{(k)})=\frac{\mathrm{tr}(XX'X_{(k)}{X_{(k)}}')}{\sqrt{\mathrm{tr}(XX'XX')\mathrm{tr}(X_{(k)}{X_{(k)}}'X_{(k)}{X_{(k)}}')}} \tag{1}$$

さて, RV 統計量 $R_k$ はデータによって値が決まる. すなわち, データのバラつきを反映する. よって, $R_k$ の値によって主成分数の選択をすると情報の量を誤って評価してしまい, 適切でない選択をしてしまうことが考えられる. そこで本研究では, $R_k$ の分布から求められたパーセント点の値によって主成分数の選択を行うということを提案する.

## 3 研究成果

データに正規性を仮定した下での $R_k$ の極限分布は以下の通りである.

**定理 3.1 (RV 統計量の極限分布)**
$p$ 変量確率ベクトル $\boldsymbol{X}\sim N_p(\mathbf{0},\ \Sigma)$ に対し, 大きさ $N$ の無作為標本を $\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \ldots,\ \boldsymbol{x}_N$, 標本共分散行列を $S$ で与える. ただし,

$$\Sigma=\Gamma\Lambda\Gamma',\ \Lambda=\mathrm{diag}[\lambda_1,\ \ldots,\ \lambda_p], \tag{2}$$

$$S=HLH',\ H=(\boldsymbol{h}_1,\boldsymbol{h}_2,\ldots,\boldsymbol{h}_p),\ L=\mathrm{diag}[\ell_1,\ell_2,\ldots,\ell_p] \tag{3}$$

である. このとき, 統計量 $R_k$ の極限分布は以下で与えられる.

$$\begin{aligned}
&R_k\xrightarrow{d}\ N(\rho_k,\ \frac{\sigma^2}{n})\\
&\sigma^2=2\rho_k^2\mathrm{tr}\left\{\left(\frac{\Lambda^2}{(\mathrm{tr}\Lambda^2)}-\frac{I^{(k)}\Lambda^2}{(\mathrm{tr}I^{(k)}\Lambda^2)}\right)^2\right\},\ \rho_k=\sqrt{\frac{\sum_i^k\lambda_i^2}{\sum_i^p\lambda_i^2}},\\
&I^{(k)}=\begin{pmatrix} I_k & \\ & O\end{pmatrix}:\ p\times p,\quad I_k:\ k\text{ 次単位行列}
\end{aligned} \tag{4}$$

しかし, Josse [2] では任意の 2 つの行列の類似度を RV 係数で測った場合, その値の正規分布への漸近的な収束の悪さが指摘されている. また, 本研究において主成分数の選択に RV 係数を用いる目的は, 主成分分析やカーネル主成分分析などの手法に関係なく, 圧縮具合を評価し, それに基づいて主成分数を決定することにある. そこでブートストラップ法によって標本から分布に依存せずに RV 統計量の分布を計算し, そこからパーセンタイル法によって下側パーセント点を求めるということを提案する. さらに, パーセンタイル法でサンプリングによる偏りのために生じる誤差を補正した $\mathrm{BC_a}$ 法を用いた手法についても提案する.

さて, 以上の 3 つの方法の精度をモンテカルロシミュレーションを用いて平均二乗誤差 (Mean Square Error) を基準に比較する.

母平均は $\mathbf{0}$ とし, 対角化された母分散共分散行列 $\Sigma = \mathrm{diag}[\lambda_1, \lambda_2, ..., \lambda_9]$ の設定は,

- Case 1: $\Sigma = \mathrm{diag}[100, 64, 36, 6, 5, 4, 3, 2, 1]$
- Case 2: $\Sigma = \mathrm{diag}[64, 49, 36, 6, 5, 4, 3, 2, 1]$
- Case 3: $\Sigma = \mathrm{diag}[49, 36, 25, 6, 5, 4, 3, 2, 1]$
- Case 4: $\Sigma = \mathrm{diag}[36, 25, 16, 6, 5, 4, 3, 2, 1]$
- Case 5: $\Sigma = \mathrm{diag}[25, 16, 9, 6, 5, 4, 3, 2, 1]$
- Case 6: $\Sigma = \mathrm{diag}[10, 8.5, 7.25, 6, 5, 4, 3, 2, 1]$

の 6 パターンとする. また, 標本サイズは $N = 50, 100, 500$ の 3 パターンとする.

これらの設定の下で $R_k$ の分布（シミュレーション回数: $1,000,000$）および $R_k$ の真の 5% 点を求めた. また, $N$ 個のデータを発生させ, 極限分布およびブートストラップ分布によって $R_k$ の 5% 点を $10,000$ 回求め, $R_k$, 極限分布, ブートストラップ分布の密度, $R_k$ の 5% 点, 極限分布, パーセンタイル法, $BC_a$ 法から求められた 5% 点, および $R_k$ の真の 5% 点周りでの Coverage Probability の MSE の比較を行った. ただし, ブートストラップ分布を生成するための繰り返し回数は 3000 である.

結果, 表 (1) より, 全ての固有値の設定, 全ての標本サイズの設定で $\mathrm{BC_a}$ 法による 5% パーセント点の Coverage Probability が $1 - 0.05$ の周りでのバラつきが小さいことが分かる.

すなわち, $\mathrm{BC_a}$ 法によって $R_k$ のパーセント点を求めれば, バラつきの中心は極限分布に比べれば真のパーセント点とは多少ずれてはいるが, 誤ったパーセント点を導くリスクを 3 つの方法の中では最小にすることが出来ることがわかった.

## 4 結論

本研究では, 主成分分析によるデータ圧縮に着目し, 圧縮データの保有する情報の量を元のデータ行列と圧縮データ行列の類似度を RV 係数で測ったもので評価する方法を提案した. また, この方法では, 主成分分析だけでなくカーネル主成分分析による圧縮の際にも情報の量を評価することが可能であり, また, 主成分分析による圧縮とカーネル主成分分析による圧縮の精度が比較可能となることを示した.

さらに, 類似度のデータによるバラつきを考慮して, 元のデータ行列と圧縮データ行列の類似度を RV 統計量のパーセント点を基準に圧縮次元の選択を行う方法を提案し, その際に $\mathrm{BC_a}$ 法によってパーセント点を求めればリスクを抑えられることをシミュレーションによって示した.

これからの課題としては, まず別の行列の類似度を測る指標への拡張, 漸近分布によるパーセント点の導出およびその補正, そして, カーネル主成分分析での数値実験などが挙げられる.

表 1　MSE による比較

Case 1

| | N=50 | N=100 | N=500 |
|---|---|---|---|
| Limit | 0.0510 | 0.0465 | 0.0426 |
| Bootstrap (Percentile) | 0.0452 | 0.0395 | 0.0351 |
| Bootstrap (BCa) | 0.0240 | 0.0251 | 0.0286 |

Case 2

| | N=50 | N=100 | N=500 |
|---|---|---|---|
| Limit | 0.0494 | 0.0438 | 0.0404 |
| Bootstrap (Percentile) | 0.0494 | 0.0427 | 0.0382 |
| Bootstrap (BCa) | 0.0203 | 0.0229 | 0.0281 |

Case 3

| | N=50 | N=100 | N=500 |
|---|---|---|---|
| Limit | 0.0477 | 0.0454 | 0.0395 |
| Bootstrap (Percentile) | 0.0499 | 0.0456 | 0.0406 |
| Bootstrap (BCa) | 0.0195 | 0.0232 | 0.0294 |

Case 4

| | N=50 | N=100 | N=500 |
|---|---|---|---|
| Limit | 0.0478 | 0.0461 | 0.0396 |
| Bootstrap (Percentile) | 0.0587 | 0.0484 | 0.0399 |
| Bootstrap (BCa) | 0.0199 | 0.0224 | 0.0275 |

Case 5

| | N=50 | N=100 | N=500 |
|---|---|---|---|
| Limit | 0.0428 | 0.0434 | 0.0393 |
| Bootstrap (Percentile) | 0.0794 | 0.0633 | 0.0427 |
| Bootstrap (BCa) | 0.0167 | 0.0201 | 0.0250 |

Case 6

| | N=50 | N=100 | N=500 |
|---|---|---|---|
| Limit | 0.0236 | 0.0266 | 0.0325 |
| Bootstrap (Percentile) | 0.1932 | 0.1455 | 0.0736 |
| Bootstrap (BCa) | 0.0092 | 0.0096 | 0.0163 |

## 参考文献


[1] Escoufier, Y.(1973). La dépendance de deux aléas vectoriels critéres et visualisation. *Rev. Statist. Appl.* **21**, 5-16.

[2] Josse, J., Pagès, J., Husson, F.(2008). Testing the significance of the RV coefficient. *Comput. Statist. Data Anal.* **53**, 82-91.

[3] Robert, P. and Escoufier, Y.(1976). A unifying tool for linear multivariate statistical methods: The RV- coefficient. *Appl. Statist.* **25**, 257-265.

[4] 保科架風, 酒折文武, 藤越康祝 (2008). 主成分の次元に関する RV 検定統計量について. 日本計算機統計学会第２２回シンポジウム論文集. 113-116.

[5] 保科架風, 酒折文武 (2009). 類似度に基づく主成分数の選択と画像解析への応用. 日本計算機統計学会第２３回大会論文集. 113-116.

[6] 赤穂昭太郎.(2008). カーネル多変量解析-非線形データ解析の新しい展開-. 岩波書店.

[7] 小西貞則, 越智義道, 大森裕浩.(2008). 計算機統計学の方法-ブートストラップ・EM アルゴリズム・MCMC-. 岩波書店.