主成分分析

評価と視覚化

村田 昇

2020.11.06

講義の予定

- 第 1 日: 主成分分析の考え方
- **第 2 日: 分析の評価と視覚化**

主成分分析の復習

主成分分析

- 多数の変量のもつ情報の分析・視覚化
  - 変量を効率的に縮約して少数の特徴量を構成する
  - 変量の間の関係を明らかにする
- 分析の方針: (以下は同値)
  - データの情報を最大限保持する変量の線形結合を構成
  - データの情報を最大限反映する座標 (方向) を探索

分析の考え方

- 1 変量データ $\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_1, \dots, \boldsymbol{a}^\mathsf{T}\boldsymbol{x}_n$ を構成
  - 観測データ $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ のもつ情報を最大限保持するベクトル $\boldsymbol{a}$ を **うまく** 選択
  - $\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_1, \dots, \boldsymbol{a}^\mathsf{T}\boldsymbol{x}_n$ のばらつきが最も大きい方向を選択
- **最適化問題**: 制約条件 $\|\boldsymbol{a}\| = 1$ の下で以下の関数を最大化せよ

$$f(\boldsymbol{a}) = \sum_{i=1}^{n} (\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_i - \boldsymbol{a}^\mathsf{T}\bar{\boldsymbol{x}})^2, \quad \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$$

行列による表現

- 中心化したデータ行列

$$X = \begin{pmatrix} \boldsymbol{x}_1^\mathsf{T} - \bar{\boldsymbol{x}}^\mathsf{T} \\ \vdots \\ \boldsymbol{x}_n^\mathsf{T} - \bar{\boldsymbol{x}}^\mathsf{T} \end{pmatrix} = \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1p} - \bar{x}_p \\ \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{np} - \bar{x}_p \end{pmatrix}$$

- 評価関数 $f(\boldsymbol{a})$ は行列 $X^\mathsf{T}X$ の二次形式

$$f(\boldsymbol{a}) = \boldsymbol{a}^\mathsf{T} X^\mathsf{T} X \boldsymbol{a}$$

固有値問題

- 最適化問題

$$\text{maximize} \quad f(\boldsymbol{a}) = \boldsymbol{a}^\mathsf{T} X^\mathsf{T} X \boldsymbol{a} \quad \text{s.t.} \quad \boldsymbol{a}^\mathsf{T}\boldsymbol{a} = 1$$

- $f(\boldsymbol{a})$ の極大値を与える $\boldsymbol{a}$ は $X^\mathsf{T}X$ の固有ベクトル

$$X^\mathsf{T} X \boldsymbol{a} = \lambda \boldsymbol{a}$$

主成分負荷量と主成分得点

- **主成分負荷量** (principal component loading): $\boldsymbol{a}$
- **主成分得点** (principal component score): $\boldsymbol{x}_i^\mathsf{T}\boldsymbol{a}$
- 第 1 主成分負荷量は $X^\mathsf{T}X$ の第 1(最大) 固有値 λ_1 に対応する固有ベクトル $\boldsymbol{a}_1$
- 同様に第 k 主成分負荷量は $X^\mathsf{T}X$ の第 k 固有値 λ_k に対応する固有ベクトル $\boldsymbol{a}_k$

寄与率

寄与率の考え方

- 回帰分析で考察した **寄与率** の一般形

$$(\text{寄与率}) = \frac{(\text{その方法で説明できるばらつき})}{(\text{データ全体のばらつき})}$$

- 主成分分析での定義 (proportion of variance)

$$(\text{寄与率}) = \frac{(\text{主成分のばらつき})}{(\text{全体のばらつき})}$$

Gram 行列のスペクトル分解

- 行列 $X^\mathsf{T}X$ (非負値正定対称行列) のスペクトル分解

$$\begin{aligned} X^\mathsf{T} X &= \lambda_1 \boldsymbol{a}_1 \boldsymbol{a}_1^\mathsf{T} + \lambda_2 \boldsymbol{a}_2 \boldsymbol{a}_2^\mathsf{T} + \cdots + \lambda_p \boldsymbol{a}_p \boldsymbol{a}_p^\mathsf{T} \\ &= \sum_{k=1}^{p} \lambda_k \boldsymbol{a}_k \boldsymbol{a}_k^\mathsf{T} \end{aligned}$$

  固有値と固有ベクトルによる行列の表現

- 主成分のばらつきの評価

$$f(\boldsymbol{a}_k) = \boldsymbol{a}_k^\mathsf{T} X^\mathsf{T} X \boldsymbol{a}_k = \lambda_k$$

  固有ベクトル (単位ベクトル) の直交性を利用

寄与率の計算

- 主成分と全体のばらつき

$$(\text{主成分}) = \sum_{i=1}^{n} (\boldsymbol{a}_k^{\mathsf{T}} \boldsymbol{x}_i - \boldsymbol{a}_k^{\mathsf{T}} \bar{\boldsymbol{x}})^2 = \boldsymbol{a}_k^{\mathsf{T}} X^{\mathsf{T}} X \boldsymbol{a}_k = \lambda_k$$
$$(\text{全体}) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2 = \sum_{l=1}^{p} \boldsymbol{a}_l^{\mathsf{T}} X^{\mathsf{T}} X \boldsymbol{a}_l = \sum_{l=1}^{p} \lambda_l$$

- 寄与率の固有値による表現:

$$(\text{寄与率}) = \frac{\lambda_k}{\sum_{l=1}^{p} \lambda_l}$$

累積寄与率

- **累積寄与率** (cumulative proportion) : 第 k 主成分までのばらつきの累計

$$(\text{累積寄与率}) = \frac{\sum_{l=1}^{k} \lambda_l}{\sum_{l=1}^{p} \lambda_l}$$

  第 1 から第 k までの寄与率の総和
  - 累積寄与率はいくつの主成分を用いるべきかの基準
  - 一般に累積寄与率が 80%程度までの主成分を用いる

R: 主成分分析の評価

- 分析結果の評価を行う関数:
  - `summary()`: 主成分負荷量や寄与率を表示
  - `plot()`: 寄与率を図示

```
## データフレームを分析
est <- prcomp( ~ x1 の変数名 + ... + xp の変数名, data = データフレーム)
## 主成分負荷量と寄与率を確認する
summary(est)
## 寄与率を図示する
plot(est)
```

データセットの準備

- 以下の 2 つのデータセットを使用します
  - `japan_social.csv` (先週から使用)

    総務省統計局より取得した都道府県別の社会生活統計指標の一部
  - `MASS::UScereal`

    Nutritional and Marketing Information on US Cereals
    The UScereal data frame has 65 rows and 11 columns. The data come from the 1993 ASA Statistical Graphics Exposition, and are taken from the mandatory F&DA food label. The data have been normalized here to a portion of one American cup.

```
library(MASS) # パッケージの読み込み
help(UScereal) # 変数名などの詳細はヘルプを参照して下さい
```

練習問題

- それぞれのデータにおいて，正規化の有無の違いで寄与率・累積寄与率がどのように異なるか確認しなさい．

```
prcomp(データフレーム) # 正規化を行わない場合
prcomp(データフレーム, scale.=TRUE) # 正規化を行う場合
## 正式なオプション名は "scale." であるが, "sc=TRUE" などでも可
```

  - japan_social.csv

```
JS.data <- read.csv("data/japan_social.csv", row.names=1)
```

  - MASS::UScereal

```
## UScereal にはカテゴリカル変数が含まれるので以下のように処理すると良い
str(UScereal) # 各変数の属性を確認する．factor/int が不要
UC.data <- UScereal[sapply(UScereal, is.double)]
```

主成分負荷量再考

主成分負荷量と主成分得点

- 負荷量 (得点係数) の大きさ: 変数の貢献度
- **問題点**:
  - 変数のスケールによって係数の大きさは変化する
  - 変数の正規化 (標本平均 0，不偏分散 1) がいつも妥当とは限らない
- スケールによらない変数と主成分の関係: **相関係数**

相関係数

- $\boldsymbol{e}_l$: 第 l 成分は 1，それ以外は 0 のベクトル
- $X\boldsymbol{e}_l$: 第 l 変数ベクトル
- $X\boldsymbol{a}_k$: 第 k 主成分得点ベクトル
- 主成分と変数の相関係数:

$$\begin{aligned}\mathrm{Cor}(X\boldsymbol{a}_k, X\boldsymbol{e}_l) &= \frac{\boldsymbol{a}_k^{\mathsf{T}} X^{\mathsf{T}} X \boldsymbol{e}_l}{\sqrt{\boldsymbol{a}_k^{\mathsf{T}} X^{\mathsf{T}} X \boldsymbol{a}_k}\sqrt{\boldsymbol{e}_l^{\mathsf{T}} X^{\mathsf{T}} X \boldsymbol{e}_l}} \\ &= \frac{\lambda_k \boldsymbol{a}_k^{\mathsf{T}} \boldsymbol{e}_l}{\sqrt{\lambda_k}\sqrt{(X^{\mathsf{T}}X)_{ll}}}\end{aligned}$$

正規化データの場合

- $X^{\mathsf{T}}X$ の対角成分は全て $n-1$ ($(X^{\mathsf{T}}X)_{ll} = n-1$)
- 第 k 主成分に対する第 l 変数の相関係数:

$$(\boldsymbol{r}_k)_l = \sqrt{\lambda_k/(n-1)} \cdot (\boldsymbol{a}_k)_l$$

- 第 k 主成分に対する相関係数ベクトル:

$$\boldsymbol{r}_k = \sqrt{\lambda_k/(n-1)} \cdot \boldsymbol{a}_k$$

- **主成分負荷量**
  - 同じ主成分への各変数の影響は固有ベクトルの成分比
  - 同じ変数の各主成分への影響は固有値の平方根で重みづけ

バイプロット

特異値分解

- 階数 r の $n \times p$ 型行列 X の分解:

$$X = U\Sigma V^{\mathsf{T}}$$

  - U は $n \times n$ 型直交行列, V は $p \times p$ 型直交行列
  - Σ は $n \times p$ 型行列

$$\Sigma = \begin{pmatrix} D & O_{r,p-r} \\ O_{n-r,r} & O_{n-r,m-r} \end{pmatrix}$$

    * $O_{s,t}$ は $s \times t$ 型零行列
    * D は $\sigma_1 \geq \sigma_2 \geq \sigma_r > 0$ を対角成分とする $r \times r$ 型対角行列

- D の対角成分: X の **特異値** (singular value)

特異値分解による Gram 行列の表現

- Gram 行列の展開:

$$\begin{aligned} X^{\mathsf{T}}X &= (U\Sigma V^{\mathsf{T}})^{\mathsf{T}}(U\Sigma V^{\mathsf{T}}) \\ &= V\Sigma^{\mathsf{T}}U^{\mathsf{T}}U\Sigma V^{\mathsf{T}} \\ &= V\Sigma^{\mathsf{T}}\Sigma V^{\mathsf{T}} \end{aligned}$$

- 行列 $\Sigma^{\mathsf{T}}\Sigma$ は対角行列

$$\Sigma^{\mathsf{T}}\Sigma = \begin{pmatrix} \sigma_1^2 & & & & & \\ & \ddots & & & & \\ & & \sigma_r^2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

特異値と固有値の関係

- 行列 V の第 k 列ベクトル $\boldsymbol{v}_k$
- 特異値の平方

$$\lambda_k = \begin{cases} \sigma_k^2, & k \leq r \\ 0, & k > r \end{cases}$$

- Gram 行列の固有値問題

$$X^\mathsf{T} X \boldsymbol{v}_k = V \Sigma^\mathsf{T} \Sigma V^\mathsf{T} \boldsymbol{v}_k = \lambda_k \boldsymbol{v}_k$$

  - $X^\mathsf{T} X$ の固有値は行列 X の特異値の平方
  - 固有ベクトルは行列 V の列ベクトル $\boldsymbol{a}_k = \boldsymbol{v}_k$

データ行列の近似表現

- 行列 U の第 k 列ベクトル $\boldsymbol{u}_k$
- データ行列の特異値分解: (**注意** Σ は対角行列)

$$X = U \Sigma V^\mathsf{T} = \sum_{k=1}^{r} \boldsymbol{u}_k \sigma_k \boldsymbol{v}_k^\mathsf{T}$$

- 第 k 主成分と第 l 主成分を用いた行列 X の近似 X'

$$X \simeq X' = \boldsymbol{u}_k \sigma_k \boldsymbol{v}_k^\mathsf{T} + \boldsymbol{u}_l \sigma_l \boldsymbol{v}_l^\mathsf{T}$$

- **バイプロット**: 上記の分解を利用した散布図

バイプロット

- X のばらつきを最大限保持する近似は $k = 1, l = 2$
- $0 \leq s \leq 1$ として

$$X' = GH^\mathsf{T}, \qquad G = \begin{pmatrix} \sigma_k^{1-s} \boldsymbol{u}_k & \sigma_l^{1-s} \boldsymbol{u}_l \end{pmatrix}, \quad H = \begin{pmatrix} \sigma_k^{s} \boldsymbol{v}_k & \sigma_l^{s} \boldsymbol{v}_l \end{pmatrix}$$

  - 行列 G の各行は各データの 2 次元座標
  - 行列 H の各行は各変量の 2 次元座標
- 関連がある 2 枚の散布図を 1 つの画面に表示する散布図を一般に**バイプロット** (biplot) と呼ぶ
- パラメタ s は 0, 1 または 1/2 が主に用いられる

R: 関数 `biplot()` の使い方

- R の標準関数: `biplot()`
- 主成分分析の結果に対して表示:

```
## データフレームを分析
est <- prcomp( ~ x1 の変数名 + ... + xp の変数名, data = データフレーム)
## 第 1 と第 2 主成分を利用した散布図
biplot(est)
## 第 2 と第 3 主成分を利用した散布図
biplot(est, choices = c(2,3))
## パラメタ s を変更 (既定値は 1)
biplot(est, scale=0)
```

練習問題

- それぞれのデータの主成分分析の結果を利用してバイプロットによる可視化を行いなさい.
  - 正規化したデータでの主成分分析を行いなさい
  - 第 1 主成分と第 2 主成分でのバイプロットを描きなさい
  - 第 2 主成分と第 3 主成分でのバイプロットを描きなさい

```
biplot(prcomp の結果, choices=c(x 軸成分,y 軸成分)) # 主成分の指定
```