主成分分析 - 評価と視覚化

数理科学続論J

(Press `?` for help, `n` and `p` for next and previous slide)

村田 昇

2019.11.15

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 (principal component analysis)

- 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 $\boldsymbol{a}\cdot\boldsymbol{x}_1,\dots,\boldsymbol{a}\cdot\boldsymbol{x}_n$ 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
- $\boldsymbol{a}\cdot\boldsymbol{x}_1,\dots,\boldsymbol{a}\cdot\boldsymbol{x}_n$ 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 $\|\boldsymbol{a}\|=1$ の下で関数

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

を最大化せよ

3 . 3

固有値問題

- 中心化したデータ行列

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

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

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

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

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{a} = \lambda \boldsymbol{a}$$

3 . 4

主成分方向と主成分得点

- **主成分方向** (principal component direction): $\boldsymbol{a}$
- **主成分得点** (principal component score): $\boldsymbol{x}_i^\top \boldsymbol{a}$
- 第1主成分方向は $\boldsymbol{X}^\top \boldsymbol{X}$ 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 主成分方向は $\boldsymbol{X}^\top \boldsymbol{X}$ 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

4.1

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 $X^\top X$ (非負値正定対称行列) のスペクトル分解

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

(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

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

(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

$$
(\text{主成分}) = \sum_{i=1}^{n} (\boldsymbol{a}_k \cdot \boldsymbol{x}_i - \boldsymbol{a}_k \cdot \bar{\boldsymbol{x}})^2 = \boldsymbol{a}_k^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{a}_k = \lambda_k
$$

$$
(\text{全体}) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2 = \sum_{l=1}^{p} \boldsymbol{a}_l^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{a}_l = \sum_{l=1}^{p} \lambda_l
$$

- 寄与率の固有値による表現:

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

4.4

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()` および `plot()`
- データフレームに対する分析:
  - データフレーム `mydata`: 必要な変数を含むデータフレーム
  - 列名: x1の変数名, …, xpの変数名

```
## データフレームを分析
est <- prcomp( ~ x1の変数名 + ... + xpの変数名, data = mydata)
## 主成分方向や寄与率を確認
summary(est)
## 寄与率を図示
plot(est)
```

4.6

演習: 寄与率による分析の評価

- 08-summary.r を確認してみよう

4.7

演習: 実データによる考察

- 累積寄与率から適切な成分数を考察してみよう
  - datasets::USArrests
  - MASS::Cars93
  - MASS::UScereal

4.8

主成分方向再考

5.1

主成分方向と主成分得点

- 得点係数の大きさから変数の貢献度がわかる
- **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を知りたい
- **相関係数** を利用することができる

相関係数

- $X\boldsymbol{a}_k$: 第 k 主成分得点
- $X\boldsymbol{e}_l$: 第 l 変数 ($\boldsymbol{e}_l$ は第 l 成分のみ1のベクトル)
- 主成分と変数の相関係数:

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

正規化データの場合

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

$$(\boldsymbol{l}_k)_l = \sqrt{\lambda_k}(\boldsymbol{a}_k)_l$$

- 第 k 主成分に対する相関係数ベクトル

$$\boldsymbol{l}_k = \sqrt{\lambda_k}\boldsymbol{a}_k$$

- **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

5.4

バイプロット

6.1

特異値分解

- 階数 r の $n \times p$ 型行列 X の分解:

$$X = U\Sigma V^\top$$

  - 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

特異値分解によるGram行列の表現

- Gram行列の展開:

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

6.3

特異値分解とGram行列の関係

- 行列 $\Sigma^\top\Sigma$ は対角行列

6.4

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

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

  - $X^\top X$ 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)

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

- 第 k 主成分と第 l 主成分を用いた行列 X の近似 X'

$$X \simeq X' = \boldsymbol{u}_k \sigma_k \boldsymbol{v}_k^\top + \boldsymbol{u}_l \sigma_l \boldsymbol{v}_l^\top$$

- **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 $k=1, l=2$
- $0 \le s \le 1$ として

$$X' = GH^\top,$$

$$G = \begin{pmatrix} \sigma_k^{1-s}\boldsymbol{u}_k & \sigma_l^{1-s}\boldsymbol{u}_l \end{pmatrix}, \quad H = \begin{pmatrix} \sigma_k^s\boldsymbol{v}_k & \sigma_l^s\boldsymbol{v}_l \end{pmatrix}$$

  - 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

6.7

R: 関数 `biplot()` の使い方

- Rの標準関数: `biplot()`
- データフレームに対する分析:
  - データフレーム `mydata`: 必要な変数を含むデータフレーム
  - 列名: x1の変数名, …, xpの変数名

```
## データフレームを分析
est <- prcomp( ~ x1の変数名 + ... + xpの変数名, data = mydata)
## 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)
```

6.8

演習: 関数 `biplot()` の使い方

- 08-biplot.r を確認してみよう

6.9

演習: 実データへの適用

- バイプロットによる分析結果の図示を行ってみよう
  - datasets::USArrests
  - MASS::Cars93
  - MASS::UScereal

6.10