主成分分析

基本的な考え方

村田 昇

2020.10.30

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

主成分分析の考え方

主成分分析

- **PCA** (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
  - 特徴量に関与する変量間の関係を明らかにする

分析の枠組み

- $X_1, \dots, X_p$: **変数**
- $Z_1, \dots, Z_d$: **特徴量** ($d \leq p$)
- 変数と特徴量の関係: (線形結合)

$$Z_k = a_{1k}X_1 + \cdots + a_{pk}X_p \quad (k = 1, \dots, d)$$

- 特徴量は定数倍の任意性があるので以下を仮定:

$$\|\boldsymbol{a}_k\|^2 = \sum_{j=1}^{p} a_{jk}^2 = 1$$

主成分分析の用語

- 特徴量 Z_k:
  第 k **主成分得点** (principal component score)
  または
  第 k **主成分**
- 係数ベクトル $\boldsymbol{a}_k$:
  第 k **主成分負荷量** (principal component loading)
  または
  第 k **主成分方向** (principal component direction)

分析の目的

- 目的:

  主成分得点 $Z_1, \dots, Z_d$ が変数 $X_1, \dots, X_p$ の情報を効率よく反映するように主成分負荷量 $\boldsymbol{a}_1, \dots, \boldsymbol{a}_d$ を観測データから **うまく** 決定する

- 分析の方針: (以下は同値)
  - データの情報を最も保持する変量の **線形結合を構成**
  - データの情報を最も反映する **座標軸を探索**
- **教師なし学習** の代表的手法の 1 つ:
  - 次元縮約: 入力をできるだけ少ない変数で表現
  - 特徴抽出: 情報処理に重要な特性を変数に凝集

R: 主成分分析を実行する関数

- R の標準的な関数:
  - `prcomp()`
  - `princomp()`
- 計算法に若干の違いがある
  - 数値計算の観点からみると `prcomp()` が優位
  - `princomp()` は S 言語 (商用) との互換性を重視した実装
- 本講義では `prcomp()` を利用

R: 関数 `prcomp()` の使い方

- データフレームの全ての列を用いる場合:

```
prcomp(x = データフレーム)
## x: 必要な変数を含むデータフレーム
```

- 列名を指定する (formula を用いる) 場合:

```
prcomp(formula = ~ x1の変数名 + ... + xpの変数名,
       data = データフレーム)
## formula: ~ 変数名 (解析の対象を + で並べる) 左辺はないので注意すること
## data: 必要な変数を含むデータフレーム
```

- 関数の返り値は `help(prcomp)` を参照

練習問題

- 数値実験により主成分分析の考え方を確認しなさい
  - 以下のモデルに従う人工データを生成する

    ```
    a <- c(1, 2)/sqrt(5) # 主成分負荷量 (単位ベクトル)
    ## 観測データ (2次元) の作成 (aのスカラー倍に正規乱数を重畳)
    n <- 100 # データ数
    myData <- data.frame(runif(n,-1,1) %o% a + rnorm(2*n, sd=0.3))
    ```

  - 観測データの散布図を作成
  - 観測データから第 1 主成分負荷量を推定

    ```
    prcomp(myData) # 全ての主成分を計算する
    ahat <- prcomp(myData)$rotation[,1] # 負荷量 (rotation) の1列目が第1主成分
    ```

  - 散布図上に主成分負荷量を描画

    ```
    abline(切片, 傾き) # 指定の直線を追加できる
    ```

第 1 主成分の計算

記号の準備

- 変数: $x_1, \dots, x_p$ (p 次元)
- 観測データ: n 個の $(x_1, \dots, x_p)$ の組

$$\{(x_{i1}, \dots, x_{ip})\}_{i=1}^{n}$$

- $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})^\top$:
  i 番目の観測データ (p 次元空間内の 1 点)
- $\boldsymbol{a} = (a_1, \dots, a_p)^\top$:
  長さ 1 の p 次元ベクトル

係数ベクトルによる射影

- データ $\boldsymbol{x}_i$ の $\boldsymbol{a}$ 方向成分の長さ:

$$\boldsymbol{a}^\top \boldsymbol{x}_i \quad (\text{スカラー})$$

- 方向ベクトル $\boldsymbol{a}$ をもつ直線上への点 $\boldsymbol{x}_i$ の直交射影

$$(\boldsymbol{a}^\top \boldsymbol{x}_i)\,\boldsymbol{a} \quad (\text{スカラー} \times \text{ベクトル})$$

幾何学的描像

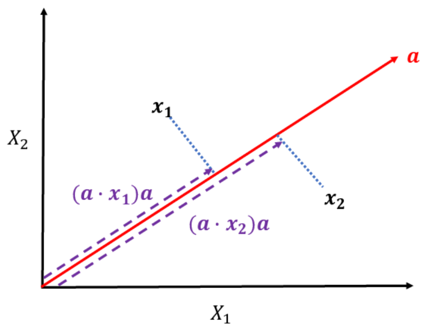

図 1: 観測データの直交射影 ($p = 2, n = 2$ の場合)

ベクトル a の選択の指針

- 線形結合での見方
  ベクトル $\boldsymbol{a}$ を **うまく** 選んで観測データ $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ の情報を最も保持する 1 変量データを構成:

$$\boldsymbol{a}^\top \boldsymbol{x}_1, \boldsymbol{a}^\top \boldsymbol{x}_2, \dots, \boldsymbol{a}^\top \boldsymbol{x}_n$$

- 座標軸での見方
  観測データの **ばらつき**を最も反映するベクトル $\boldsymbol{a}$ を選択:

$$\arg\max_{\boldsymbol{a}} \sum_{i=1}^{n} (\boldsymbol{a}^\top \boldsymbol{x}_i - \boldsymbol{a}^\top \bar{\boldsymbol{x}})^2, \quad \bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i,$$

ベクトル $\boldsymbol{a}$ の最適化

- 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:

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

- この最大化問題は必ず解をもつ:
  - $f(\boldsymbol{a})$ は連続関数
  - 集合 $\{\boldsymbol{a} \in \mathbb{R}^p : \|\boldsymbol{a}\| = 1\}$ はコンパクト (有界閉集合)

行列による表現

- 中心化したデータ行列:

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

  回帰分析のデザイン行列を参照

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

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

  回帰分析の Gram 行列を参照

ベクトル $\boldsymbol{a}$ の解

- 最適化問題

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

- 固有値問題

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

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

第 1 主成分

- 求める $\boldsymbol{a}$ は行列 $X^\top X$ の最大固有ベクトル (長さ 1)
- $f(\boldsymbol{a})$ は行列 $X^\top X$ の最大固有値

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

- **第 1 主成分負荷量**: ベクトル $\boldsymbol{a}$
- **第 1 主成分得点**:

$$z_{i1} = a_1 x_{i1} + \cdots + a_p x_{ip} \quad (i = 1, \ldots, n)$$

練習問題

- 第 1 主成分と Gram 行列の固有ベクトルの関係を調べなさい
  - 人工データを生成する
  - 主成分分析を実行する
  - Gram 行列を計算し固有値・固有ベクトルを求める

```
## 中心化を行う
X <- scale(myData, scale=FALSE) # help(scale) でオプション scale を確認
## Gram 行列を計算する
G <- crossprod(X)
## 固有値・固有ベクトルを求める
eig <- eigen(G) # help(eigen) で返り値を確認
```

第 2 主成分以降の計算

Gram 行列の性質

- $X^\mathsf{T}X$ は非負定値対称行列
- $X^\mathsf{T}X$ の固有値は 0 以上の実数
  - 固有値を重複を許して降順に並べる

$$\lambda_1 \geq \cdots \geq \lambda_p \quad (\geq 0)$$

- 固有値 λ_j に対する固有ベクトルを $\boldsymbol{a}_j$(長さ 1) とする

$$\|\boldsymbol{a}_j\| = 1 \quad (j = 1, \dots, p)$$

Gram 行列のスペクトル分解

- $\boldsymbol{a}_1, \dots, \boldsymbol{a}_p$ は **互いに直交** するようとることができる

$$j \neq k \quad \Rightarrow \quad \boldsymbol{a}_j^\mathsf{T}\boldsymbol{a}_k = 0$$

- 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

第 2 主成分の考え方

- 第 1 主成分:
  - 主成分負荷量: ベクトル $\boldsymbol{a}_1$
  - 主成分得点: $\boldsymbol{a}_1^\mathsf{T}\boldsymbol{x}_i$ $(i = 1, \dots, n)$
- 第 1 主成分負荷量に関してデータが有する情報:

$$(\boldsymbol{a}_1^\top \boldsymbol{x}_i)\, \boldsymbol{a}_1 \quad (i = 1, \dots, n)$$

- 第 1 主成分を取り除いた観測データ: (分析対象)

$$\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - (\boldsymbol{a}_1^\top \boldsymbol{x}_i)\, \boldsymbol{a}_1 \quad (i = 1, \dots, n)$$

第 2 主成分の最適化

- 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:

$$\tilde{f}(\boldsymbol{a}) = \sum_{i=1}^{n} (\boldsymbol{a}^\top \tilde{\boldsymbol{x}}_i - \boldsymbol{a}^\top \bar{\tilde{\boldsymbol{x}}})^2 \quad \text{ただし} \quad \bar{\tilde{\boldsymbol{x}}} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i$$

行列による表現

- 中心化したデータ行列:

$$\tilde{X} = \begin{pmatrix} \tilde{\boldsymbol{x}}_1^\top - \bar{\tilde{\boldsymbol{x}}}^\top \\ \vdots \\ \tilde{\boldsymbol{x}}_n^\top - \bar{\tilde{\boldsymbol{x}}}^\top \end{pmatrix} = X - X\boldsymbol{a}_1\boldsymbol{a}_1^\top$$

- Gram 行列

$$\begin{aligned} \tilde{X}^\top \tilde{X} &= (X - X\boldsymbol{a}_1\boldsymbol{a}_1^\top)^\top (X - X\boldsymbol{a}_1\boldsymbol{a}_1^\top) \\ &= X^\top X - \lambda_1 \boldsymbol{a}_1\boldsymbol{a}_1^\top \\ &= \sum_{k=2}^{p} \lambda_k \boldsymbol{a}_k \boldsymbol{a}_k^\top \end{aligned}$$

第 2 主成分の解

- Gram 行列 $\tilde{X}^\top \tilde{X}$ の固有ベクトル $\boldsymbol{a}_1$ の固有値は 0
- Gram 行列 $\tilde{X}^\top \tilde{X}$ の最大固有値は λ_2
- 解は第 2 固有値 λ_2 に対応する固有ベクトル $\boldsymbol{a}_2$
- 以下同様に第 k 主成分負荷量は $X^\top X$ の第 k 固有値 λ_k に対応する固有ベクトル $\boldsymbol{a}_k$

データセットの準備

- 主成分分析では以下のデータセットを使用します
  - `japan_social.csv`

    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
    * Pref: 都道府県名
    * Forest: 森林面積割合 (%) 2014 年
    * Agri: 就業者 1 人当たり農業産出額 (販売農家) (万円) 2014 年
    * Ratio: 全国総人口に占める人口割合 (%) 2015 年
    * Land: 土地生産性（耕地面積 1 ヘクタール当たり）(万円) 2014 年
    * Goods: 商業年間商品販売額［卸売業＋小売業］（事業所当たり）(百万円) 2013 年

  https://www.e-stat.go.jp/SG1/estat/List.do?bid=000001083999&cycode=0

練習問題

- 前掲のデータを用いて主成分分析を行いなさい
  - 都道府県名を行名としてデータを読み込む

    ```
    JS.data <- read.csv("data/japan_social.csv", row.names=1)
    ```
  - データの散布図行列を描く
  - 各データの箱ひげ図を描き，変数の大きさを確認する
  - 主成分負荷量を計算する

    ```
    JS.pca <- prcomp(JS.data, scale=TRUE)
    ## scale=TRUE とすると変数を正規化してから解析する
    ```