

主成分分析

基本的な考え方

村田 昇

講義概要

- 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

- 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
 - 特徴量に關与する変量間の關係を明らかにする
- PCA (Principal Component Analysis)
 - 構成する特徴量: **主成分** (principal component)

分析の枠組み

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

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

- 特徴量は定数倍の任意性があるので以下を仮定:

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

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 \mathbf{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 a_1, \dots, 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

演習

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

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

- $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$:
 i 番目の観測データ (p 次元空間内の 1 点)
- $\mathbf{a} = (a_1, \dots, a_p)^\top$:
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 \mathbf{x}_i の \mathbf{a} 方向成分の長さ:

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

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

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

幾何学的描像

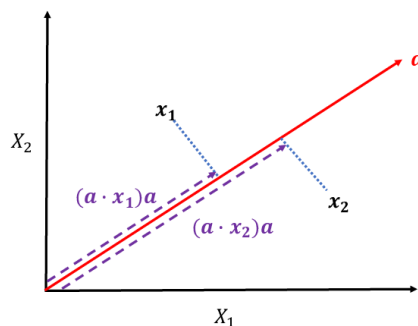


図 1: 観測データの直交射影 ($p = 2, n = 2$ の場合)

ベクトル a の選択の指針

- 射影による特徴量の構成

ベクトル a を **うまく** 選んで観測データ x_1, \dots, x_n の情報を最も保持する 1 変量データを構成:

$$a^T x_1, a^T x_2, \dots, a^T x_n$$

- 特徴量のばらつき

観測データの **ばらつき** を最も反映するベクトル a を選択:

$$\arg \max_a \sum_{i=1}^n (a^T x_i - a^T \bar{x})^2, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

ベクトル a の最適化

- 最適化問題

制約条件 $\|a\| = 1$ の下で以下の関数を最大化せよ:

$$f(a) = \sum_{i=1}^n (a^T x_i - a^T \bar{x})^2$$

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

第1主成分の解

行列による表現

- 中心化したデータ行列:

$$X = \begin{pmatrix} x_1^T - \bar{x}^T \\ \vdots \\ x_n^T - \bar{x}^T \end{pmatrix} = \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1p} - \bar{x}_p \\ \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{np} - \bar{x}_p \end{pmatrix}$$

(回帰分析のデザイン行列を参照)

- 評価関数 $f(a)$ は行列 $X^T X$ の二次形式:

$$f(a) = a^T X^T X a$$

(回帰分析の Gram 行列を参照)

ベクトル a の解

- 最適化問題

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

- 固有値問題

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

$$X^T X a = \lambda a$$

第1主成分

- 求める \mathbf{a} は行列 $X^T X$ の最大固有ベクトル (長さ 1)
- このとき $f(\mathbf{a})$ は行列 $X^T X$ の最大固有値

$$f(\mathbf{a}) = \mathbf{a}^T X^T X \mathbf{a} = \mathbf{a}^T \lambda \mathbf{a} = \lambda$$

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

$$z_{i1} = a_1 x_{i1} + \cdots + a_p x_{ip} = \mathbf{a}^T \mathbf{x}_i, \quad (i = 1, \dots, n)$$

演習

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) で戻り値を確認
```

Gram 行列の性質

Gram 行列の固有値

- $X^T X$ は非負定値対称行列
- $X^T X$ 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_k に対する固有ベクトルを \mathbf{a}_k (長さ 1) とする

$$\|\mathbf{a}_k\| = 1, \quad (k = 1, \dots, p)$$

Gram 行列のスペクトル分解

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

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

- 行列 $X^T X$ (非負定値対称行列) のスペクトル分解:

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

固有値と固有ベクトルによる行列の表現

第2主成分以降の計算

第2主成分の考え方

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

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

- 第1主成分を取り除いた観測データ: (分析対象)

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

第2主成分の最適化

- 最適化問題
制約条件 $\|\mathbf{a}\| = 1$ の下で以下の関数を最大化せよ:

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

第2主成分以降の解

行列による表現

- 中心化したデータ行列:

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

- Gram 行列

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

第2主成分

- Gram 行列 $\tilde{X}^T \tilde{X}$ の固有ベクトル \mathbf{a}_1 の固有値は 0
- Gram 行列 $\tilde{X}^T \tilde{X}$ 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 \mathbf{a}_2
- 以下同様に第 k 主成分負荷量は $X^T X$ の第 k 固有値 λ_k に対応する固有ベクトル \mathbf{a}_k

演習

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
```

次週の予定

- 第1日: 主成分分析の考え方
- 第2日: 分析の評価と視覚化