主成分分析

基本的な考え方

村田 昇

2020.11.03

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

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

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

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

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

- 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

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

幾何学的描像

ベクトル $\boldsymbol{a}$ 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:

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

- 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:

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

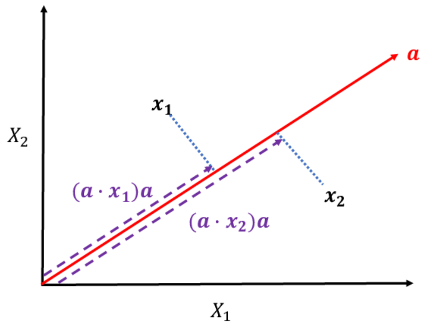

図 1: 観測データの直交射影 ($p=2, n=2$ の場合)

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)

第 1 主成分の解

ベクトル a の解

- 最適化問題

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

- 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)
- このとき $f(\boldsymbol{a})$ 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**:

$$z_{i1}=a_1x_{i1}+\cdots+a_px_{ip} \quad (i=1,\ldots,n)$$

Gram 行列の性質

Gram 行列の固有値

- $X^{\mathsf{T}}X$ 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 $\boldsymbol{a}_1$
  - 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:

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

- 第 1 主成分を取り除いた観測データ: (分析対象)

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

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:

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

第 2 主成分の解

第 2 主成分

- Gram 行列 $\tilde{X}^\mathsf{T}\tilde{X}$ の固有ベクトル $\boldsymbol{a}_1$ の固有値は 0
- Gram 行列 $\tilde{X}^\mathsf{T}\tilde{X}$ 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 $\boldsymbol{a}_2$
- 以下同様に第 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 $\boldsymbol{a}_k$