

主成分分析

評価と視覚化

村田 昇

2019.11.06

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
 - 変量の間関係を明らかにする
- 分析の方針:
 - 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
 - (データの情報を保持する=データを区別できる)

分析の考え方

- 1変量データ $\mathbf{a}^\top \mathbf{x}_1, \dots, \mathbf{a}^\top \mathbf{x}_n$ を構成
 - 観測データ $\mathbf{x}_1, \dots, \mathbf{x}_n$ のもつ情報を最大限保持するベクトル \mathbf{a} を **うまく** 選択
 - $\mathbf{a}^\top \mathbf{x}_1, \dots, \mathbf{a}^\top \mathbf{x}_n$ の変動 (ばらつき) が最も大きい方向を選択
- **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

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

行列による表現

- 中心化したデータ行列

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

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

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

固有値問題

- 最適化問題

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

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

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

主成分負荷量と主成分得点

- 主成分負荷量 (principal component loading): \mathbf{a}
- 主成分得点 (principal component score): $\mathbf{x}_i^\top \mathbf{a}$
- 第 1 主成分負荷量は $X^\top X$ の第 1(最大) 固有値 λ_1 に対応する固有ベクトル \mathbf{a}_1
- 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 \mathbf{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

$$(\text{寄与率}) = \frac{(\text{その方法で説明できる変動})}{(\text{データ全体の変動})}$$

- 主成分分析での定義 (proportion of variance)

$$(\text{寄与率}) = \frac{(\text{主成分の変動})}{(\text{全体の変動})}$$

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

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

固有値と固有ベクトルによる行列の表現

- 主成分の変動の評価

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

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

- 主成分と全体の変動

$$\begin{aligned}(\text{主成分}) &= \sum_{i=1}^n (\mathbf{a}_k^\top \mathbf{x}_i - \mathbf{a}_k^\top \bar{\mathbf{x}})^2 = \mathbf{a}_k^\top X^\top X \mathbf{a}_k = \lambda_k \\(\text{全体}) &= \sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 = \sum_{l=1}^p \mathbf{a}_l^\top X^\top X \mathbf{a}_l = \sum_{l=1}^p \lambda_l\end{aligned}$$

- 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 主成分までの変動の累計

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

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

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

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

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

正規化データの場合

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

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

- 第 k 主成分に対する相関係数ベクトル:

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

- 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:

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

- 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:

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

- 行列 $\Sigma^T \Sigma$ は対角行列

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

特異値と固有値の関係

- 行列 V の第 k 列ベクトル \mathbf{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

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

- $X^T X$ 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 $\mathbf{a}_k = \mathbf{v}_k$

データ行列の近似表現

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

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

- 第 k 主成分と第 l 主成分を用いた行列 X の近似 X'

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

- **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 の変動を最大限保持する近似は $k=1, l=2$
- $0 \leq s \leq 1$ として

$$X' = G H^T,$$

$$G = (\sigma_k^{1-s} \mathbf{u}_k \quad \sigma_l^{1-s} \mathbf{u}_l), \quad H = (\sigma_k^s \mathbf{v}_k \quad \sigma_l^s \mathbf{v}_l)$$

- 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