主成分分析

評価と視覚化

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
  - 変量の間の関係を明らかにする
- 分析の方針
  - **データの情報を保持する = データを区別することができる**
  - 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 変量の特徴量 $\boldsymbol{a}^\top\boldsymbol{x}_1,\dots,\boldsymbol{a}^\top\boldsymbol{x}_n$ 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 **適切に** 選択
  - $\boldsymbol{a}^\top\boldsymbol{x}_1,\dots,\boldsymbol{a}^\top\boldsymbol{x}_n$ 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 $\|\boldsymbol{a}\| = 1$ の下で以下の関数を最大化せよ

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

行列による表現

- 中心化したデータ行列

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

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

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

固有値問題

- 最適化問題

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

- 解の条件
  $f(\boldsymbol{a})$ の極大値を与える $\boldsymbol{a}$ は $X^\top X$ の固有ベクトルである
  $$X^\top X \boldsymbol{a} = \lambda \boldsymbol{a}$$
  - 未定係数法を用いている

主成分負荷量と主成分得点

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

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 $X^\top X$ (半正定値行列) のスペクトル分解

$$X^\top X = \sum_{k=1}^{p} \lambda_k \boldsymbol{a}_k \boldsymbol{a}_k^\top$$

  - 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

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

  - 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

- 主成分と全体の変動

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

- 固有値による寄与率の表現

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

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

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

  - 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

実習

主成分負荷量

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
  - 変数の標準化 (平均 0，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** を考えればよい

相関係数

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

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

相関係数による評価

- 標準化されたデータの場合
  - $X^\top X$ の対角成分は全て $n-1$ ($(X^\top X)_{jj} = n-1$)
  - 第 k 主成分に対する相関係数ベクトル

$$\boldsymbol{r}_k = \sqrt{\lambda_k/(n-1)} \cdot \boldsymbol{a}_k, \quad (\boldsymbol{r}_k)_j = \sqrt{\lambda_k/(n-1)} \cdot (\boldsymbol{a}_k)_j$$

  - **主成分負荷量の比較**
    * 同じ主成分 (k を固定) への各変数の影響は固有ベクトルの成分比
    * 同じ変数 (j を固定) の各主成分への影響は固有値の平方根で重みづけ
- 標準化されていない場合
  - 変数の分散の影響を考慮する必要がある

データ行列の分解表現

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

- 行列 Σ の成分表示

$$\Sigma = \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & O_{r,p-r} \\ & & \sigma_r & \\ & & & \\ & O_{n-r,r} & & O_{n-r,m-r} \end{pmatrix}$$

- 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

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

- 行列 $\Sigma^\top\Sigma$ は対角行列

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

バイプロット

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

$$\lambda_k = \begin{cases} \sigma_k^2, & k \le r \\ 0, & k > r \end{cases}$$

- 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 $\boldsymbol{a}_k = \boldsymbol{v}_k$

データ行列の分解

- 行列 U の第 k 列ベクトル $\boldsymbol{u}_k$
- 行列 V の第 k 列ベクトル $\boldsymbol{v}_k$
- データ行列の特異値分解: (Σ の非零値に注意)

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

データ行列の近似表現

- 第 k 主成分と第 l 主成分を用いた行列 X の近似 X'

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

- 行列の積による表現

$$\begin{aligned} X' =& GH^\top, (0 \le s \le 1) \\ & G = \begin{pmatrix}\sigma_k^{1-s}\boldsymbol{u}_k & \sigma_l^{1-s}\boldsymbol{u}_l\end{pmatrix}, \quad H = \begin{pmatrix}\sigma_k^{s}\boldsymbol{v}_k & \sigma_l^{s}\boldsymbol{v}_l\end{pmatrix} \end{aligned}$$

バイプロット

- 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
- 行列 G, H の各行を 2 次元座標と見なす

$$X' = GH^{\top}$$

  - 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
- パラメタ s は 0, 1 または 1/2 が主に用いられる
- X の変動を最大限保持する近似は $k = 1, l = 2$

実習

次回の予定

- **第 1 日 : 判別分析の考え方**
- 第 2 日 : 分析の評価