多変量解析
第 6 回: 主成分分析 (2) 解析手法

藤本 衡

2018 年 5 月 16 日

本日の内容

- 主成分分析 (N 変量) の解法
- 寄与率
- 主成分得点
- 負荷量

2 変量の解法

$$\begin{pmatrix} \mathbb{V}\left[X_1\right] & \mathrm{Cov}\left[X_1,\ X_2\right] \\ \mathrm{Cov}\left[X_1,\ X_2\right] & \mathbb{V}\left[X_2\right] \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \lambda \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

つまり

$$\boldsymbol{C}\boldsymbol{w} = \lambda \boldsymbol{w}$$

の形の固有値問題を解けばよい。C は X_1, X_2 の分散共分散行列であり、λ は C の固有値、w は対応する右固有ベクトル。

N 変量の主成分分析

- やることは同じ（行列が 2 次でなく N 次になるだけ）

$$
\begin{pmatrix}
\mathbb{V}\left[X_1\right] & \mathrm{Cov}\left[X_1,\ X_2\right] & \cdots & \mathrm{Cov}\left[X_1,\ X_N\right] \\
\mathrm{Cov}\left[X_2,\ X_1\right] & \mathbb{V}\left[X_2\right] & \cdots & \mathrm{Cov}\left[X_2,\ X_N\right] \\
\vdots & \vdots & \ddots & \vdots \\
\mathrm{Cov}\left[X_N,\ X_1\right] & \mathrm{Cov}\left[X_N,\ X_2\right] & \cdots & \mathbb{V}\left[X_N\right]
\end{pmatrix}
\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{pmatrix}
= \lambda
\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{pmatrix}
$$

N 次正方行列の固有値

- N 次の固有方程式の解を求める？
  - 2 次なら簡単、3 次もなんとか、4 次は？（できます）
  - 5 次以上は代数的に解を求めるための一般的な方法がない（できない）
- 代数的に無理なら数値計算で求めよう
  - べき乗法（最大固有値のみ）
  - QR 法

べき乗法 (power method)

- **絶対値最大の固有値が重解だと適用できない**

べき乗法のアルゴリズム

1. $||\boldsymbol{w}^{(0)}||_2 = 1$ **なる初期ベクトル** $\boldsymbol{w}^{(0)}$ **を定める**. $i = 1$.
2. $\boldsymbol{v}^{(i)} \leftarrow \boldsymbol{C}\boldsymbol{w}^{(i-1)}, \quad \lambda^{(i)} \leftarrow \boldsymbol{w}^{(i-1)\mathsf{T}}\boldsymbol{v}^{(i)}, \quad \boldsymbol{w}^{(i)} \leftarrow \dfrac{\boldsymbol{v}^{(i)}}{||\boldsymbol{v}^{(i)}||_2}$
3. $\forall j, \left|\dfrac{w_j^{(i)}}{w_j^{(i-1)}}\right| < \varepsilon$ **?** $\longrightarrow$ **真ならば収束, さもなくば** 2 **へ戻る**
4. $\lambda^{(i)}, \ \boldsymbol{w}^{(i)}$ **が絶対値最大固有値** λ_1 **と対応する右固有ベクトル** w_1 **を与える**

ヴィーラント減次 (Wielandt deflation)

- べき乗法は絶対値最大の固有値しか得られない？
- 対称行列であれば下記の方法で他の固有値も得られる
  - 非対称でも少し形を変えるとなんとかなる

ヴィーラント減次による λ_J の導出

1. $C_1 = C$ とおく. $j = 1$.
2. C_j の絶対値最大固有値 λ_j と対応する右固有ベクトル w_j を求める.
3. $j < N$ ならば $C_{j+1} = C_j - \lambda_j w_j w_j^\intercal, \quad j \leftarrow j + 1$ として 2 に戻る.

QR 法

- 三角行列の対角成分が固有値になることを利用
- QR 分解を繰り返して、三角行列に収束させる
- 固有ベクトルを求めるには別途計算が必要

QR 法のアルゴリズム

1. $\boldsymbol{C}_1 = \boldsymbol{C},\ i = 1.$
2. $\boldsymbol{C}_i \to \boldsymbol{Q}_i \boldsymbol{R}_i$
3. $\boldsymbol{C}_{i+1} \leftarrow \boldsymbol{R}_i \boldsymbol{Q}_i = \boldsymbol{Q}_i^\intercal \boldsymbol{C}_i \boldsymbol{Q}_i$
4. $\boldsymbol{C}_{i+i}$ の下三角成分が十分小さくなれば収束，さもなくば $i \leftarrow i + 1$ として 2 へ戻る

寄与率 (contribution)

- 主成分はどの程度データのばらつきを表現できているか？
- 第 $2, 3, \ldots$ 主成分を求める必要があるのか？
- λ は主成分の分散に比例している
  - 実際、データを標準化し相関行列で求めた場合、$\lambda = \mathbb{V}[Z]$ である

寄与率

第 j 主成分の寄与率 r_j は以下の式で与えられる

$$r_j = \frac{\lambda_j}{\sum_{k=1}^{N} \lambda_k} = \frac{\lambda_j}{\sum_{k=1}^{N} \mathbb{V}[X_k]}$$

寄与率と求めるべき主成分の数

- 第 j 主成分までの累積寄与率は $\sum_{k=1}^{j} r_k$ で与えられる
- 求めるべき主成分の基準は必ずしも明確ではない
- 累積寄与率 80%〜90%あたりがひとつの目安
- または第 r_j に対して r_{j+1} が大きく減少しているあたりで打ち止め

主成分得点

- 各個体ごとに、主成分の式にデータを代入したもの

主成分得点 (主成分スコア)

第 j 主成分

$$Z_j = w_{j1}X_1 + w_{j2}X_2 + \cdots w_{jN}X_N$$

に対して、各個体の主成分得点は以下のように得られる。

$$\begin{aligned}
z_{j1} &= w_{j1}x_{11} + w_{j2}x_{12} + \cdots w_{jN}x_{1N} \\
z_{j2} &= w_{j1}x_{21} + w_{j2}x_{22} + \cdots w_{jN}x_{2N} \\
&\vdots \\
z_{jn} &= w_{j1}x_{n1} + w_{j2}x_{n2} + \cdots w_{jN}x_{nN}
\end{aligned}$$

例: 数学と国語

- 不偏分散・不偏共分散を用い標準化しない場合
  - $\lambda_1 = 284.7, \quad w_{11} = 0.8718, \; w_{12} = -0.4899, r_1 = 0.92$
  - $\lambda_2 = 24.15, \quad w_{21} = -0.4899, \; w_{22} = 0.8718, r_2 = 0.08$

学生	数学 (x_{i1})	国語 (x_{i2})	第 1 主成分得点 ($w_{11}x_{i1} + w_{12}x_{i2}$)	第 2 主成分得点 ($w_{21}x_{i1} + w_{22}x_{i2}$)
A	88	65	44.87	-13.55
B	62	77	16.33	-36.75
C	90	52	52.98	-1.24
D	57	70	15.40	-33.10
E	72	70	28.47	-25.75

負荷量 (loading)

- 変量 X_k と Z_j の相関係数は、X_k が主成分に与える影響の大きさとその方向を示す

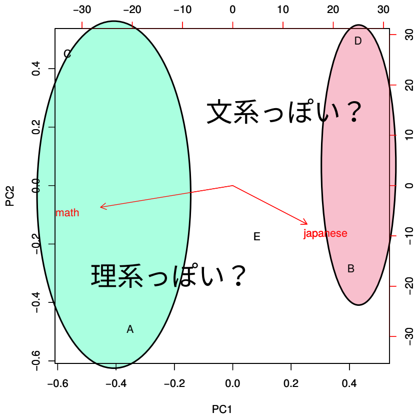
主成分負荷量 (因子負荷量)

X_k と Z_j の相関係数は以下の式で与えられる

$$\frac{\sqrt{\lambda_j} w_{jk}}{\sqrt{\mathbb{V}\left[X_k\right]}}$$

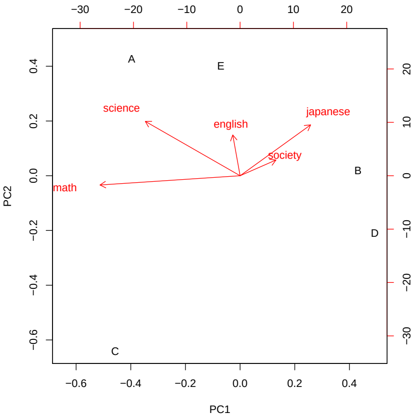
これを X_k の主成分負荷量と呼ぶ。(因子分析とも関係するため、因子負荷量と呼ばれることもある)

主成分得点と負荷量の散布図



- 主成分得点で個体をプロット
- 主成分負荷量で変量をプロット
- 傾向や影響を読み取る

主成分得点と負荷量の散布図 (5 科目版)



- $r_1 = 0.74,\ r_2 = 0.16$
- **何か読み取れるでしょうか？**