

# 多变量解析

## 第 5 回: 主成分分析 (1) 基礎概念

藤本 衡

2018 年 5 月 9 日

# 本日の内容

- 多変量データを「より少ない尺度で」比較する
- 主成分分析の考え方
- 2 変数での例

# 多変量データの比較

- 多変量を 1 個ずつ比較するのは容易だが、順序・優劣・分類が明確でない
- できれば 1 つか 2 つの尺度で並べ替えができないだろうか？

# 例：誰が理系で、誰が文系？

- あるいは、「どの科目が理系的で、どの科目が文系的なのか」

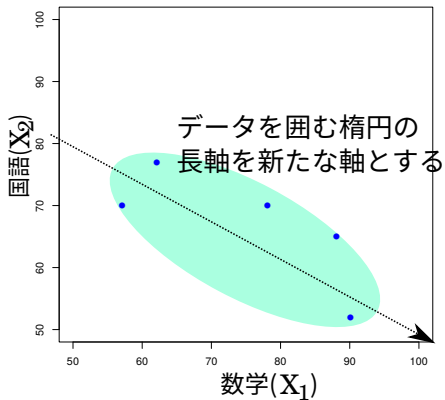
学生	数学	英語	国語	理科	社会
A	88	75	65	90	70
B	62	60	77	70	78
C	90	58	52	82	65
D	57	68	70	64	72
E	72	64	70	90	70

# 主成分分析 (Principal Component Analysis: PCA)

- $N$  個の変量を持つ個体の群を 1 つの指標で並べ替えたい
- できるだけ多くの変量から情報を取り込み、重み付けして新たな指標を合成する  
= 主成分

$$Z = w_1X_1 + w_2X_2 + \cdots w_NX_N$$

## 2変数での例: 数学と国語



- 数学の点数で並べ替える手もある
- でも国語の点数も加味したい
- 軸を変えてしまえばいいのでは？

# 軸の回転

## 回転行列

$$\begin{pmatrix} Z \\ Z' \end{pmatrix} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

新しい軸に沿った指標  $Z$  は

$$Z = w_1 X_1 + w_2 X_2 = X_1 \cos \theta - X_2 \sin \theta$$

となる。

ここで

$$w_1^2 + w_2^2 = \cos^2 \theta + \sin^2 \theta = 1$$

である (重みの平方和が 1 となる) 点に注意。

# 主成分を求める基準 (1)

- 似た個体は近い場所に、異なる個体は遠くに「ばらけて」ほしい
- $Z$  の分散が最大になるように  $(w_1, w_2)$  を選ばばよい？

## 制約条件付き最大化問題

$$\begin{aligned} \text{maximize } \mathbb{V}[Z] &= \mathbb{V}[w_1 X_1 + w_2 X_2] \\ &= w_1^2 \mathbb{V}[X_1] + 2w_1 w_2 \text{Cov}[X_1, X_2] + w_2^2 \mathbb{V}[X_2] \quad (1) \\ \text{subject to } w_1^2 + w_2^2 - 1 &= 0 \end{aligned}$$



# ラグランジュの未定乗数法

(1) 式はそのままラグランジュの未定乗数法が適用可能  
 $\lambda$  をラグランジュ乗数とおき、

$$L = \mathbb{V}[Z] - \lambda(w_1^2 + w_2^2 - 1)$$

$$\frac{\partial L}{\partial w_1} = 2w_1 \mathbb{V}[X_1] + 2w_2 \text{Cov}[X_1, X_2] - 2\lambda w_1 = 0 \quad (2)$$

$$\frac{\partial L}{\partial w_2} = 2w_2 \mathbb{V}[X_2] + 2w_1 \text{Cov}[X_1, X_2] - 2\lambda w_2 = 0 \quad (3)$$

# 固有値問題

(2),(3) 式を行列で表現すると

$$\begin{pmatrix} \mathbb{V}[X_1] & \text{Cov}[X_1, X_2] \\ \text{Cov}[X_1, X_2] & \mathbb{V}[X_2] \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \lambda \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \quad (4)$$

つまり

$$Cw = \lambda w$$

の形の固有値問題となる。

$C$  は  $X_1, X_2$  の分散共分散行列であり、 $\lambda$  は  $C$  の固有値、 $w$  は対応する右固有ベクトル。

# 固有値を求める

2 次正方行列なら固有値は簡単に求まるよね？

## 特性方程式 (固有方程式)

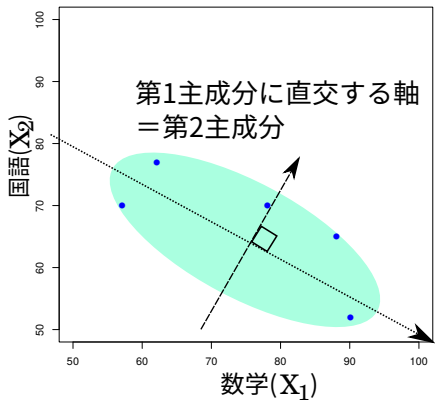
$$\begin{aligned}|C - \lambda I| &= \begin{vmatrix} \mathbb{V}[X_1] - \lambda & \text{Cov}[X_1, X_2] \\ \text{Cov}[X_1, X_2] & \mathbb{V}[X_2] - \lambda \end{vmatrix} \\ &= \lambda^2 - (\mathbb{V}[X_1] + \mathbb{V}[X_2])\lambda + \mathbb{V}[X_1]\mathbb{V}[X_2] - (\text{Cov}[X_1, X_2])^2 \\ &= 0\end{aligned}$$

$$\lambda = \frac{1}{2} \left\{ (\mathbb{V}[X_1] + \mathbb{V}[X_2]) \pm \sqrt{(\mathbb{V}[X_1] - \mathbb{V}[X_2])^2 + 4(\text{Cov}[X_1, X_2])^2} \right\}$$

# 最大固有値と固有ベクトル

- $C$  は半正定値行列なので固有値はすべて非負
- $\lambda$  の解は 2 つ得られる 大きい方 (最大固有値) を  $\lambda_1$  とおく
- $\lambda_1$  に対応する右固有ベクトル (第 1 固有ベクトル) を求める  
     $\mathbb{V}[Z]$  を最大化する  $(w_1, w_2)$  が得られる
  - また連立一次方程式です
  - ただし制約  $w_1^2 + w_2^2 = 1$  がないと比例関係しか出てこないので注意
- 第 1 固有ベクトルから得られる主成分を、第 1 主成分と呼ぶ

## 第2主成分



- もう1個の固有値は使えないの？
- 楕円の短軸を表す = 第2主成分
- これも使える「かも」しれない

# 変量の標準化

- この例では  $X_1$  と  $X_2$  はともに 100 点満点（同じスケール）
- もし一方が 10 点満点だったら？ 単位を合わせたい

## 標準化 (standardization)

$$T_j = \frac{X_j - \mathbb{E}[X_j]}{\sqrt{\mathbb{V}[X_j]}}$$

$\mathbb{E}[T_j] = 0$ ,  $\mathbb{V}[T_j] = 1$  となる。また  $\text{Cov}[T_1, T_2]$  は  $X_1$  と  $X_2$  の相関係数に等しいので、標準化したとき (4) 式の  $C$  は相関係数行列  $R$  に置き換えられる。

$$Rw = \lambda w$$