

## データ解析

<http://coconut.sys.eng.shizuoka.ac.jp/data/>

静岡大学工学部  
安藤和敏

2005.12.14

## 主成分分析のデータ (変数が3個の場合)

| No       | 変数 x     | 変数 y     | 変数 z     |
|----------|----------|----------|----------|
| 1        | $x_1$    | $y_1$    | $z_1$    |
| 2        | $x_2$    | $y_2$    | $z_2$    |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $i$      | $x_i$    | $y_i$    | $z_i$    |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n$      | $x_n$    | $y_n$    | $z_n$    |

## 合成変数 $u$

$$a^2 + b^2 + c^2 = 1$$

を満たす  $a, b, c$  に対して,

$$u = ax + by + cz$$

という変数変換を考える.

ただし, ここで  $u$  の分散  $s_u^2$  が最大になるように,  $a, b, c$  を選びたい.

## $u$ の分散

$$\begin{aligned} s_u^2 &= \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^2 = \frac{1}{n} \sum_{i=1}^n (ax_i + by_i + cz_i - a\bar{x} - b\bar{y} - c\bar{z})^2 \\ &= a^2 \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 + b^2 \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 + c^2 \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 \\ &\quad + 2ab \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + 2bc \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z}) \\ &\quad + 2ca \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x}) \\ &= a^2 s_x^2 + b^2 s_y^2 + c^2 s_z^2 + abs_{xy} + bcs_{yz} + cas_{zx} \end{aligned}$$

## $u$ の分散の最大化

$$\text{条件 } a^2 + b^2 + c^2 = 1$$

を満足する  $[a, b, c]$  のうちで,

$$s_u^2 = a^2 s_x^2 + b^2 s_y^2 + c^2 s_z^2 + abs_{xy} + bcs_{yz} + cas_{zx}$$

を最大にするものを求めたい.

## 主成分の必要条件

$s_u^2$  を最大にする  $[a, b, c]$  は, 以下の方程式の解である必要がある.

$$\begin{bmatrix} s_x^2 & s_{xy} & s_{zx} \\ s_{xy} & s_y^2 & s_{yz} \\ s_{zx} & s_{yz} & s_z^2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \lambda \begin{bmatrix} a \\ b \\ c \end{bmatrix},$$

$$a^2 + b^2 + c^2 = 1$$

つまり, そのような  $[a, b, c]$  は, 分散共分散行列の固有ベクトル(で長さが1のもの)である.

## 主成分の十分条件

$$s_u^2 = a^2 s_x^2 + b^2 s_y^2 + c^2 s_z^2 + abs_{xy} + bcs_{yz} + cas_{zx}$$

$$= [a \quad b \quad c] \begin{bmatrix} s_x^2 & s_{xy} & s_{zx} \\ s_{xy} & s_y^2 & s_{yz} \\ s_{zx} & s_{yz} & s_z^2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

$$= [a \quad b \quad c] \lambda \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \lambda (a^2 + b^2 + c^2) = \lambda.$$

## 主成分の必要十分条件

$u$ の分散  $s_u^2$  の極値を与える  $[a, b, c]$  は,

$$\text{分散共分散行列} \begin{bmatrix} s_x^2 & s_{xy} & s_{zx} \\ s_{xy} & s_y^2 & s_{yz} \\ s_{zx} & s_{yz} & s_z^2 \end{bmatrix} \text{の固有ベクトル}$$

であり, そのとき,  $s_u^2 = \lambda$  (= 固有ベクトル)となる.

したがって,  $s_u^2$  の最大値を与える  $[a, b, c]$  は, 分散共分散行列の最大の固有値に属する固有ベクトル (で, 長さが1のもの)である.

## 主成分分析のデータ (変数が5個の場合)

| No       | 変数 x     | 変数 y     | 変数 z     | 変数 v     | 変数 w     |
|----------|----------|----------|----------|----------|----------|
| 1        | $x_1$    | $y_1$    | $z_1$    | $v_1$    | $w_1$    |
| 2        | $x_2$    | $y_2$    | $z_2$    | $v_2$    | $w_2$    |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $i$      | $x_i$    | $y_i$    | $z_i$    | $v_i$    | $w_i$    |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n$      | $x_n$    | $y_n$    | $z_n$    | $v_n$    | $w_n$    |

## 合成変数 $u$

$$a^2 + b^2 + c^2 + d^2 + e^2 = 1$$

を満たす  $[a, b, c, d, e]$  に対して,

$$u = ax + by + cz + dv + ew$$

という変数変換を考える.

ただし, ここで  $u$  の分散  $s_u^2$  が最大になるように,  $[a, b, c, d, e]$  を選びたい.

## 分散共分散行列

$$S = \begin{bmatrix} s_x^2 & s_{xy} & s_{xz} & s_{xv} & s_{xw} \\ s_{xy} & s_y^2 & s_{yz} & s_{yv} & s_{yw} \\ s_{xz} & s_{yz} & s_z^2 & s_{zv} & s_{zw} \\ s_{xv} & s_{yv} & s_{zv} & s_v^2 & s_{vw} \\ s_{xw} & s_{yw} & s_{zw} & s_{vw} & s_w^2 \end{bmatrix},$$

とする.

## $u$ の分散 $s_u^2$

3変数のときと同様にして,

$$s_u^2 = [a \quad b \quad c \quad d \quad e] S [a \quad b \quad c \quad d \quad e]^T$$

であることが分かる.

さらに,  $u$  の分散を最大化する  $[a, b, c, d, e]$  は,  $S$  の最大の固有値  $\lambda_1$  に属する固有ベクトルである.

## 主成分の必要条件

$s_u^2$  を最大にする  $[a, b, c, d, e]$  は、以下の方程式の解である必要がある。

$$S \begin{bmatrix} a \\ b \\ c \\ d \\ e \end{bmatrix} = \lambda \begin{bmatrix} a \\ b \\ c \\ d \\ e \end{bmatrix}, a^2 + b^2 + c^2 + d^2 + e^2 = 1$$

つまり、そのような  $[a, b, c, d, e]$  は、分散共分散行列  $S$  の固有ベクトル (で長さが1のもの) である。

## 主成分の必要十分条件

$[a, b, c, d, e]$  が、 $S$  の固有値  $\lambda$  に属する固有ベクトルであるならば、

$$\begin{aligned} s_u^2 &= [a \ b \ c \ d \ e] S [a \ b \ c \ d \ e]^T \\ &= [a \ b \ c \ d \ e] \lambda [a \ b \ c \ d \ e]^T \\ &= \lambda (a^2 + b^2 + c^2 + d^2 + e^2) = \lambda. \end{aligned}$$

したがって、 $s_u^2$  の最大値を与える  $[a, b, c, d, e]$  は、分散共分散行列の最大の固有値に属する固有ベクトル (で、長さが1のもの) である。

## 第1主成分

$S$  の最大固有値  $\lambda_1$  に属する固有ベクトル  $[a_1, b_1, c_1, d_1, e_1]$  を係数として得られる合成変数

$$u = a_1 x + b_1 y + c_1 z + d_1 v + e_1 w$$

が主成分である。

以降では、 $u$  を第1主成分と呼んで  $u_1$  と書くことにしよう。  
すなわち、

$$u_1 = a_1 x + b_1 y + c_1 z + d_1 v + e_1 w$$

## 分散共分散行列の固有値

$$S = \begin{bmatrix} s_x^2 & s_{xy} & s_{xz} & s_{xv} & s_{xw} \\ s_{xy} & s_y^2 & s_{yz} & s_{yv} & s_{yw} \\ s_{xz} & s_{yz} & s_z^2 & s_{zv} & s_{zw} \\ s_{xv} & s_{yv} & s_{zv} & s_v^2 & s_{vw} \\ s_{xw} & s_{yw} & s_{zw} & s_{vw} & s_w^2 \end{bmatrix},$$

の固有値を  $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5$  とする。

## 第2主成分

2番目に大きい固有値  $\lambda_2$  に属する固有ベクトル (で長さが1のもの) を  $[a_2, b_2, c_2, d_2, e_2]$  とする。

$[a_2, b_2, c_2, d_2, e_2]$  を係数とする合成変数

$$u_2 = a_2 x + b_2 y + c_2 z + d_2 v + e_2 w$$

は第2主成分と呼ばれる。  $u_2$  の分散は、第1主成分

$$u_1 = a_1 x + b_1 y + c_1 z + d_1 v + e_1 w$$

に次いで2番目に大きい分散を与える。

なぜならば、

## 第2主成分

$[a_2, b_2, c_2, d_2, e_2]$  が、 $S$  の固有値  $\lambda_2$  に属する固有ベクトルであるので、

$$\begin{aligned} s_{u_2}^2 &= [a_2 \ b_2 \ c_2 \ d_2 \ e_2] S [a_2 \ b_2 \ c_2 \ d_2 \ e_2]^T \\ &= [a_2 \ b_2 \ c_2 \ d_2 \ e_2] \lambda_2 [a_2 \ b_2 \ c_2 \ d_2 \ e_2]^T \\ &= \lambda_2 (a_2^2 + b_2^2 + c_2^2 + d_2^2 + e_2^2) = \lambda_2. \end{aligned}$$

## 第3, 第4, 第5主成分

3番目に大きい固有値  $\lambda_3$  に属する固有ベクトル (で長さが1のもの) を  $[a_3, b_3, c_3, d_3, e_3]$  とする.

$[a_3, b_3, c_3, d_3, e_3]$  を係数とする合成変数

$$u_3 = a_3x + b_3y + c_3z + d_3v + e_3w$$

は第3主成分と呼ばれ,  $u_3$  の分散は  $\lambda_3$  となる.

以下同様に, 第4主成分, 第5主成分も定義される.

## 寄与率

第1主成分  $u_1$  の寄与率  $C_1$  は,

$$C_1 = \frac{s_{u_1}^2}{s_x^2 + s_y^2 + s_z^2 + s_v^2 + s_w^2} = \frac{\lambda_1}{s_x^2 + s_y^2 + s_z^2 + s_v^2 + s_w^2}$$

で定義される.

$u_1$  の寄与率は, 与えられた多変量データのもつ情報量のうち,  $u_1$  で表現できる情報量であると解釈できる ( p.12を見よ).

## 寄与率

一般に第  $i$  主成分  $u_i$  の寄与率  $C_i$  は,

$$C_i = \frac{s_{u_i}^2}{s_x^2 + s_y^2 + s_z^2 + s_v^2 + s_w^2} = \frac{\lambda_i}{s_x^2 + s_y^2 + s_z^2 + s_v^2 + s_w^2}$$

で定義される.

$u_i$  の寄与率は, 与えられた多変量データのもつ情報量のうち,  $u_i$  で表現できる情報量であると解釈できる ( p.12を見よ).

## 累積寄与率

第1主成分  $u_1$  の寄与率  $C_1$  と第2主成分  $u_2$  の寄与率  $C_2$  の和

$$C_2' = \frac{s_{u_1}^2 + s_{u_2}^2}{s_x^2 + s_y^2 + s_z^2 + s_v^2 + s_w^2} = \frac{\lambda_1 + \lambda_2}{s_x^2 + s_y^2 + s_z^2 + s_v^2 + s_w^2}$$

は第2主成分までの累積寄与率と呼ばれる.

これは, 第1主成分  $u_1$  と第2主成分  $u_2$  の2つの合成変数によって, 与えられた多変量データの情報をどの程度表現しているかを示す指標である.

## 累積寄与率

以下同様に, 第3主成分までの累積寄与率,

$$C_3' = \frac{s_{u_1}^2 + s_{u_2}^2 + s_{u_3}^2}{s_x^2 + s_y^2 + s_z^2 + s_v^2 + s_w^2} = \frac{\lambda_1 + \lambda_2 + \lambda_3}{s_x^2 + s_y^2 + s_z^2 + s_v^2 + s_w^2}$$

第4主成分までの累積寄与率も同様に定義される.

$$C_4' = \frac{s_{u_1}^2 + s_{u_2}^2 + s_{u_3}^2 + s_{u_4}^2}{s_x^2 + s_y^2 + s_z^2 + s_v^2 + s_w^2} = \frac{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}{s_x^2 + s_y^2 + s_z^2 + s_v^2 + s_w^2}$$

これらの量の意味するところはもはや明らかであろう.

## とりあげる主成分の数

第1主成分  $u_1$  の寄与率  $C_1$  が十分大きいものであれば, もうこれ以上の主成分を調べる必要はないが, 寄与率  $C_1$  の大きさが満足いくものでなければ, 第2主成分まで調べる必要がある. そこで, 第2主成分までの累積寄与率もまだ満足いくものでなければ, 累積寄与率が満足いく数字になるまで, 以降の主成分を調べていく.

ただし, とりあげる主成分の数を増やすことは, 与えられたデータを少ない変数で表現するという, 主成分分析本来の目的に反するので, 望ましいことではない.

## 主成分の解釈

主成分の意味を考えるための手助けになるものとして、**変量プロット**と**主成分得点プロット**がある。両方とも、視覚的に主成分を捕らえるためのものである。

## 変量プロット

第1, 第2主成分 $u_1, u_2$ が, それぞれ以下の式で表されているとする。

$$u_1 = a_1x + b_1y + c_1z + d_1v + e_1w$$

$$u_2 = a_2x + b_2y + c_2z + d_2v + e_2w$$

以下の5つの点を二次元平面にプロットしたものが、**変量プロット**である。

$$(a_1, a_2), (b_1, b_2), (c_1, c_2), (d_1, d_2), (e_1, e_2)$$

## 主成分得点プロット

2つの主成分, 例えば $u_1$ と $u_2$ を考えて, 以下の式によって, 各データ $[x_i, y_i, z_i, v_i, w_i]$ の第1, 第2主成分得点を計算する。

$$\begin{cases} u_{1i} = a_1x_i + b_1y_i + c_1z_i + d_1v_i + e_1w_i \\ u_{2i} = a_2x_i + b_2y_i + c_2z_i + d_2v_i + e_2w_i \end{cases}$$

こうして得られる $n$ 個の点  $(u_{1i}, u_{2i})$  ( $i=1, \dots, n$ )

を2次元平面上にプロットしたものが**主成分得点プロット**である。

## Excelで学ぼう

ファイル: 第3章/3\_3

## 本日のまとめ

- 第 $i$ 主成分 $u_i$ がどのようにして得られるかを理解した。(データの分散共分散行列の番目に大きい固有値に属する固有ベクトルから得られる.)
- 寄与率, 累積寄与率の定義とその意味を理解した。
- Excelを用いて, 第 $i$ 主成分 $u_i$ 主成分を計算する方法, 第 $i$ 主成分得点を計算する方法を理解した。
- 変量プロット, 主成分得点プロットの概念, 及び, Excelを用いてこれらのプロットの求め方を理解した。