05．主成分分析

京都大学　加納 学

Division of Process Control & Process Systems Engineering
Department of Chemical Engineering, Kyoto University

manabu@cheme.kyoto-u.ac.jp
http://www-pse.cheme.kyoto-u.ac.jp/~kano/

Outline

- 主成分分析
- 具体例
- 主成分回帰

標準化

各変数を平均0，分散1の変数に変換する.

$$x_{nm} = \frac{x^*_{nm} - \bar{x}_m}{\sigma_m}$$

変数　　m
サンプル　n

平均

$$\bar{x}_m = \frac{1}{N}\sum_{n=1}^{N} x^*_{nm}$$

分散

$$\sigma_m^2 = \frac{1}{N-1}\sum_{n=1}^{N} (x^*_{nm} - \bar{x}_m)^2$$

準備

入力変数　出力変数

測定データ

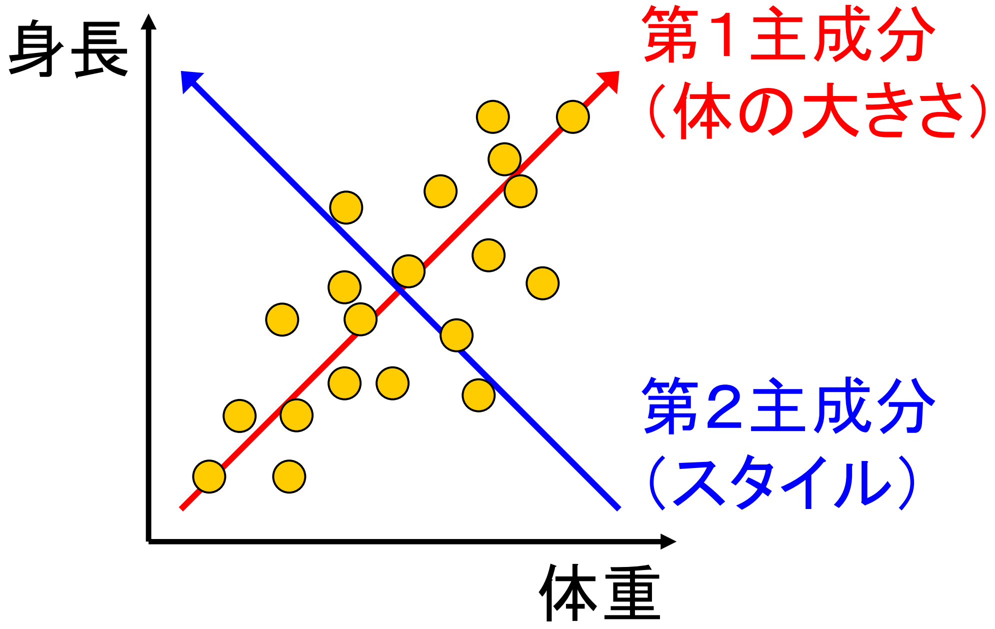
$$X^* = \begin{bmatrix} x_{11}^* & x_{12}^* & \cdots & x_{1M}^* \\ x_{21}^* & x_{22}^* & \cdots & x_{2M}^* \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1}^* & x_{N2}^* & \cdots & x_{NM}^* \end{bmatrix} \qquad Y^* = \begin{bmatrix} y_{11}^* \\ y_{21}^* \\ \vdots \\ y_{N1}^* \end{bmatrix}$$

標準化データ

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{bmatrix} \qquad Y = \begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{N1} \end{bmatrix}$$

主成分分析

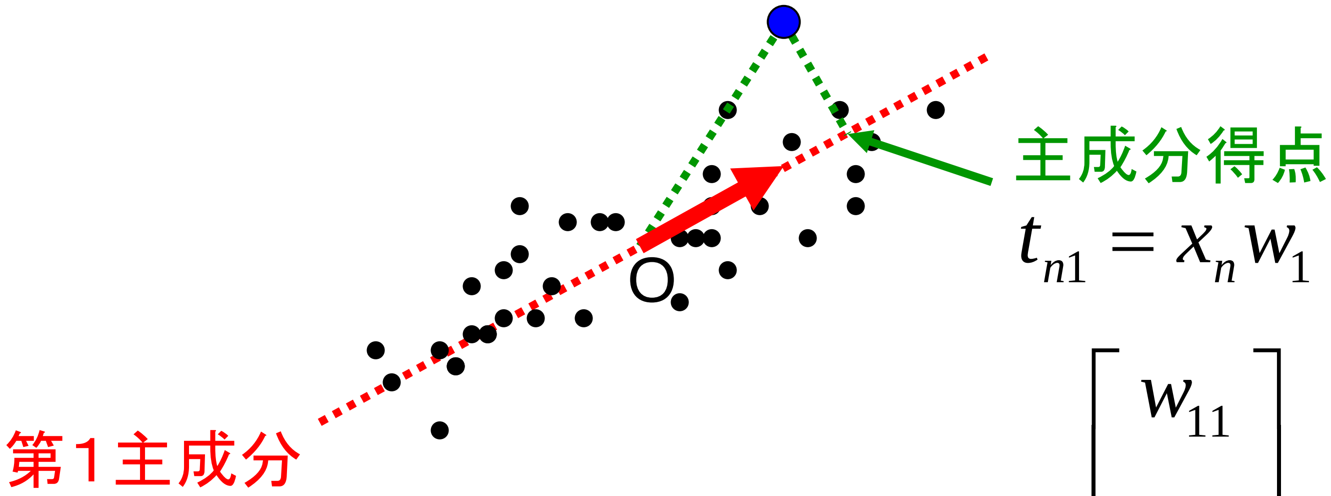
測定データを最もよく表現する軸を新たに作成する.
新しい変数(主成分)は, 入力変数の線形結合として,
その分散が最大となるように決定される. また, 各主
成分は互いに直交するように決定される.



PCA： 主成分の決定1

主成分得点の分散が最大となるように，
大きさ1の主成分ベクトル w_1 を決定する．

測定データ $x_n = \begin{bmatrix} x_{n1} & x_{n2} & \cdots & x_{nM} \end{bmatrix}$



$$t_{n1} = x_n w_1 = x_n \cos\theta$$

主成分ベクトル $w_1 = \begin{bmatrix} w_{11} \\ w_{21} \\ \vdots \\ w_{M1} \end{bmatrix}$

PCA： 主成分の決定2

第1主成分得点 $t_1 = Xw_1$ $\|w_1\| = 1$

分散
$$\frac{1}{N-1} t_1^T t_1 = \frac{1}{N-1}(Xw_1)^T Xw_1$$
$$= w_1^T \frac{1}{N-1} X^T Xw_1$$
$$= w_1^T V w_1$$ ← 最大化

共分散行列 $V = \frac{1}{N-1} X^T X$

PCA：　主成分の決定3

ラグランジュ乗数法

$$J_1 = w_1^T V w_1 - \lambda(w_1^T w_1 - 1)$$

$$\frac{\partial J_1}{\partial w_1} = 2Vw_1 - 2\lambda w_1 = 0$$

分散　$w_1^T V w_1 = \lambda$

$$(V - \lambda I)w_1 = 0$$

← 固有値問題

ラグランジュ乗数 λ および第1主成分ベクトル w_1 は共分散行列 V の固有値および固有ベクトルである．第1主成分得点の分散は最大固有値に等しく，第1主成分ベクトルは最大固有値に対応する固有ベクトルとして与えられる．

PCA: 主成分の決定4

第r主成分得点の分散は共分散行列のr番目に大きな固有値に等しく、第r主成分ベクトルはその固有値に対応する固有ベクトルとして与えられる.

$$VW = W\Lambda$$

共分散行列 V

Loading Matrix

$W = [w_1 \quad w_2 \quad \cdots \quad w_R]$ w_r 第r主成分ベクトル

$\Lambda = diag\{\lambda_1, \lambda_2, \cdots, \lambda_R\}$ λ_r 第r主成分得点の分散

PCA:　主成分の特徴

Loading Matrix　　$W = [w_1 \quad w_2 \quad \cdots \quad w_R]$

直交行列

主成分得点　　$T = XW$

$$T = [t_1 \quad t_2 \quad \cdots \quad t_R]$$

$$\begin{aligned}\frac{1}{N-1}T^T T &= \frac{1}{N-1}(XW)^T XW \\ &= W^T V W \\ &= W^T W \Lambda \\ &= \Lambda\end{aligned}$$

$$VW = W\Lambda$$

主成分は互いに直交する.

PCA： 再構築

Loading Matrix $W = [w_1 \quad w_2 \quad \cdots \quad w_R]$

直交行列

主成分得点 $T = XW$

$T = [t_1 \quad t_2 \quad \cdots \quad t_R]$

入力データの再構築 $\hat{X} = TWW^T$

主成分数 R を適切に選択することにより,
データの主要な変動のみを再構築できる.

PCA： 寄与率

$$tr(V) = \sum_{m=1}^{M} \sigma_{x_m}^2 = \sum_{r=1}^{R} \sigma_{t_r}^2 = \sum_{r=1}^{R} \lambda_r$$

寄与率

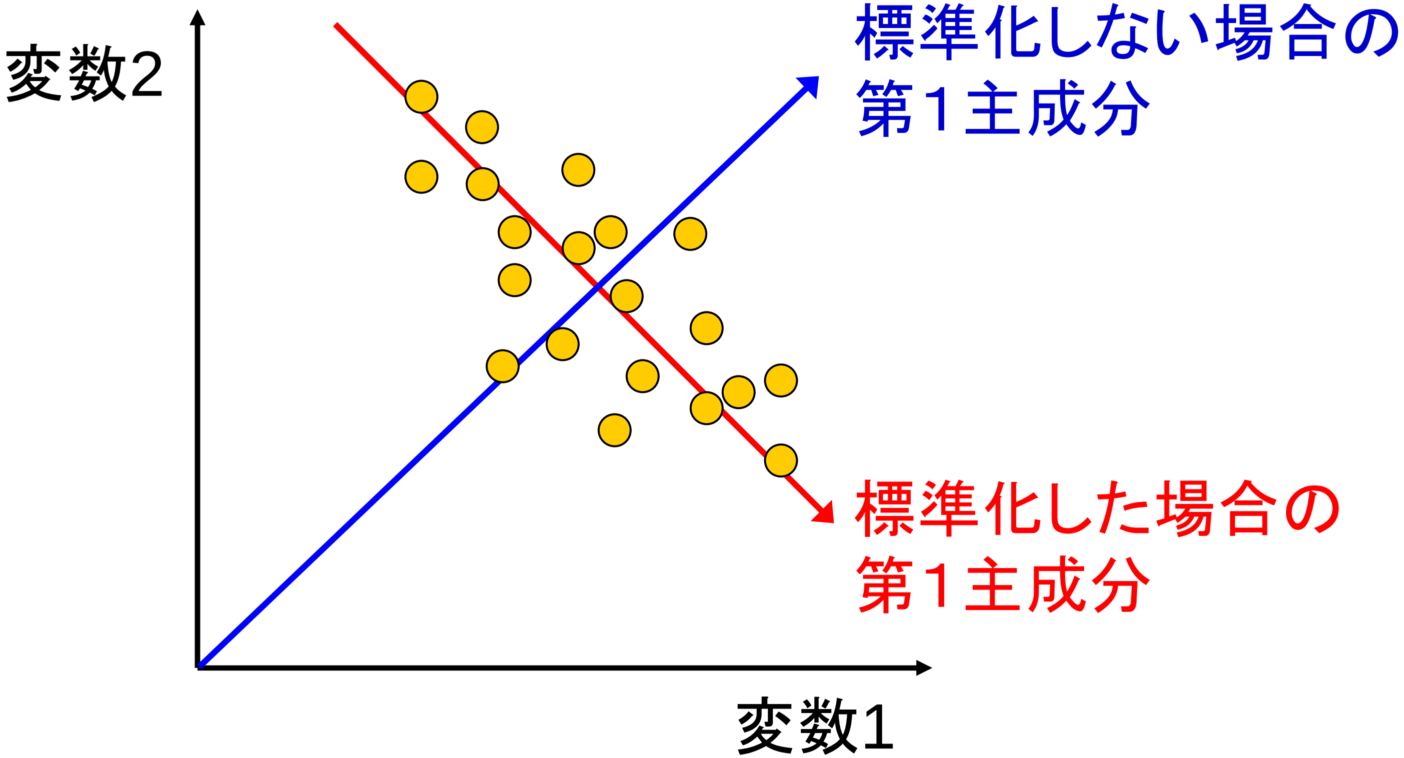
$$C_r = \frac{\sigma_{z_r}^2}{\sum_{i=1}^{R} \sigma_{z_i}^2} = \frac{\lambda_r}{\sum_{i=1}^{R} \lambda_i}$$

累積寄与率

$$P_r = \sum_{i=1}^{r} C_i = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{R} \lambda_i}$$

寄与率が大きな主成分（標準化データを利用する場合 λ>1が目安となる）が重要である.

標準化の意味



必ずしも標準化しなければならないわけではないが,
捉えたい相関関係を捉えられるかどうかの判断は必要.

Outline

- 主成分分析
- 具体例
- 主成分回帰

例：　対象データ

	身長(x1)	胸囲(x2)	体重(x3)
1	167.0	84.0	61.0
2	167.5	87.0	55.5
3	168.4	86.0	57.0
4	172.0	85.0	57.0
5	155.3	82.0	50.0
6	151.4	87.0	50.0
7	163.0	92.0	66.5
8	174.0	94.0	65.0
9	168.0	88.0	60.5
10	160.4	84.9	49.5

例：　主成分分析の結果

Loading Matrix
共分散行列の固有ベクトル

$$
\begin{array}{cccc}
 & \text{体の大きさ} & \text{スタイル} & \\
 & \text{PC1} & \text{PC2} & \text{PC3}
\end{array}
$$

$$
W = \begin{bmatrix} 0.5343 & 0.7465 & 0.3965 \\ 0.5571 & -0.6638 & 0.4990 \\ 0.6357 & -0.0457 & -0.7706 \end{bmatrix}
\begin{array}{l} \text{身長} \\ \text{胸囲} \\ \text{体重} \end{array}
$$

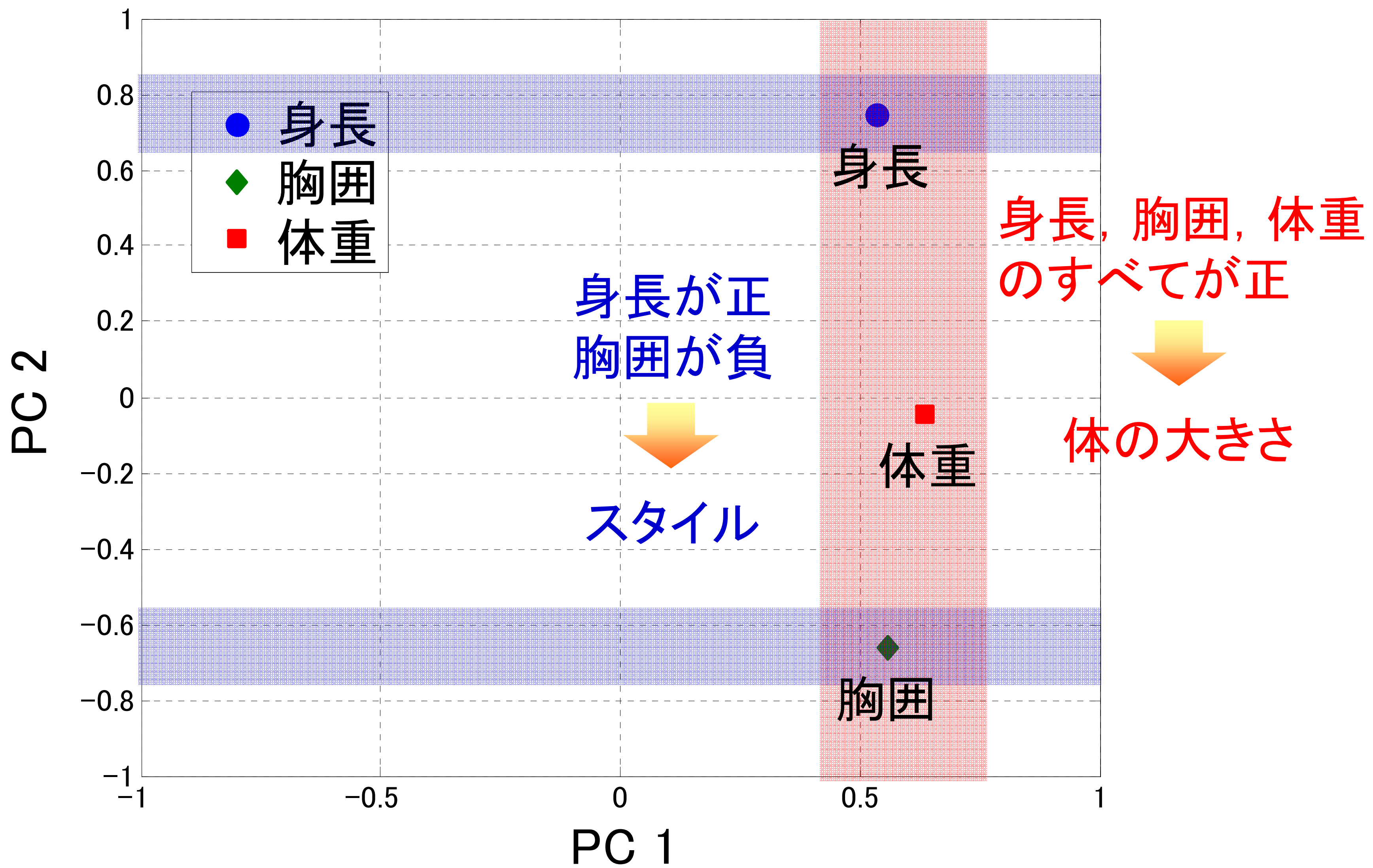
共分散行列の固有値

$$
\Lambda = \begin{bmatrix} 2.2018 & 0 & 0 \\ 0 & 0.6148 & 0 \\ 0 & 0 & 0.1833 \end{bmatrix}
$$

	PC1	PC2	PC3
寄与率	73.4	20.5	6.1
累積寄与率	73.4	93.9	100.0

例： 変量プロット

第2主成分（累積寄与率 94%）までの変量プロットを描く



Outline

- 主成分分析
- 具体例
- 主成分回帰

多重共線性の問題を回避するには

多重共線性の問題を回避するためには，すべての入力変数の中から少数の線形独立な変数を選択すればよい．

言うのは簡単だが，どうやって選ぶのか？

主成分分析を利用すれば，少数の線形独立な変数（主成分）を簡単に作成することができる．

Principal Component Regression (PCR)

主成分回帰

PCR

$$\hat{y} = \sum_{r=1}^{R} b_r t_r = \sum_{r=1}^{R} b_r \sum_{m=1}^{M} x_m w_{mr} = \sum_{m=1}^{M} x_m \sum_{r=1}^{R} b_r w_{mr}$$

$$\hat{y} = tb = xWb$$

主成分を入力変数として，
線形回帰式を構築する．

OLS

$$\hat{y} = \sum_{m=1}^{M} a_m x_m$$

$$\hat{y} = xa$$

すべての主成分を採用する場合，
PCRはMNSに一致する．

主成分数が入力変数の数に等しい場合，PCRはOLSに一致する．