1

PLS（部分的最小二乗法）入門

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/

2

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

3

準備

入力変数　　出力変数

測定データ

$$X^* = \begin{bmatrix} x^*_{11} & x^*_{12} & \cdots & x^*_{1M} \\ x^*_{21} & x^*_{22} & \cdots & x^*_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x^*_{N1} & x^*_{N2} & \cdots & x^*_{NM} \end{bmatrix} \quad Y^* = \begin{bmatrix} y^*_{11} \\ y^*_{21} \\ \vdots \\ y^*_{N1} \end{bmatrix}$$

標準化データ

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

4

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？

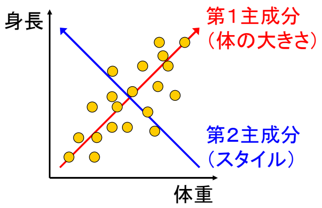
主成分分析を利用すれば，少数の線形独立な変数（主成分）を簡単に作成することができる．

Principal Component Regression (PCR)

5

主成分分析（PCA）

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．

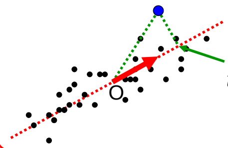


6

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　$x_n = [x_{n1} \quad x_{n2} \quad \cdots \quad x_{nM}]$



主成分得点　$t_{n1} = x_n w_1 = \|x_n\| \cos\theta$

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 $t_1 = Xw_1 \qquad \|w_1\| = 1$

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

Principal Component Regression

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

$\hat{y} = tb = xWb$

主成分を入力変数として，線形回帰式を構築する．

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

$\hat{y} = xa$

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．

PCRで十分か？

PCRを利用することにより，多重共線性の問題を回避し，線形回帰モデルを構築することができる．

主成分は入力データを最もよく表現するように決定されるため，入力変数間の相関関係を捉えることはできる．しかし，主要な主成分が出力変数の推定に寄与するとは限らない．

出力変数との相関が強い潜在変数を入力変数として採用すべきである．

Partial Least Squares (PLS)

Partial Least Squares

出力変数と潜在変数（入力変数の線形結合）との内積が最大となるように，潜在変数を決定する．

OLS $\quad r_{y\hat{y}} = \frac{\sigma_{y\hat{y}}^2}{\sigma_y \sigma_{\hat{y}}} = \frac{y^T \hat{y}}{\|y\| \|\hat{y}\|} = \cos\theta$

PCR $\quad \sigma_z^2 = \frac{1}{N-1} \|z\|^2$

PLS $\quad \langle y, z \rangle = \|y\| \|z\| \cos\theta$

PLSはOLSとPCRの中間的な性質を持つ．出力変数との相関および入力変数間の相関を同時に考慮して，適切な潜在変数を決定する．

PLS：特徴

1. 多重共線性の問題を回避できる．
2. PCRと比較して，より少ない潜在変数を用いて出力変数を推定できる．
3. サンプル数が少なくても，安定したパラメータ推定値が得られる．
4. 真のパラメータが得られるわけではない．

PLS： クロスバリデーション
13

PLSでは，採用する潜在変数の数を決めるために，
クロスバリデーションを利用することが多い．

＜クロスバリデーション＞

サンプルをN個のグループに分割し，N-1個のグループを用いてモデルを構築し，残り1個のデータを用いてモデルの検証を行う．この手順をN回繰り返し，二乗誤差の和が最小となる，最適な潜在変数の数を決定する．

OLS vs. PLS
14

	y	Data "A" x1	Data "A" x2	Data "A" x3	Data "B" x1	Data "B" x2	Data "B" x3	
	241	15.9	34.6	64.8	16.1	34.7	65.1	
	321	37.0	16.1	72.1	36.9	16.3	72.0	
	82	61.1	83.0	28.6	60.6	82.8	28.9	
	156	86.0	65.9	33.9	85.9	65.9	34.2	
OLS		1.36	-0.80	5.01	-4.28	-18.9	-26.0	
PLS (1)		-1.00	-1.59	1.11	-1.00	-1.60	1.12	安定
(2)		0.73	-2.84	1.53	0.74	-2.86	1.54	
(3)		1.36	-0.80	5.01	-4.28	-18.9	-26.0	

OLSはダメなのか？
15

PLSは優れた特質を有する線形回帰手法である．

最小二乗法を使うメリットは何もないのだろうか？

＜例題：相関の強い入力変数＞

x1 = randn(1000,1);

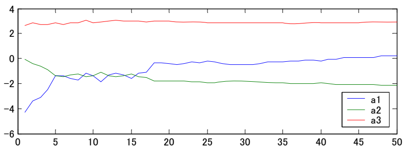
x2 = 2*x1 + 0.1*randn(1000,1);

x3 = -1*x1 + 0.2*randn(1000,1);

Y = -2*x2 + 3*x3 + 0.3*randn(1000,1);

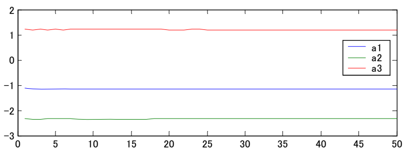
パラメータのサンプル数依存性
16

OLS



PLS

nLV=1



サンプル数 [x5]

まとめ
17

入力変数間に強い相関がある場合，多重共線性の問題を回避するために，PLSが有効である．

PLSは，入力変数間の相関および出力変数との相関を同時に考慮して，潜在変数を決定する．

PLSにより，安定な（ばらつきの少ない）パラメータ推定値が得られるが，潜在変数の数が少ない場合には，推定値は真値に一致しない．

おわり
18