06. PLS（部分的最小二乗法）

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/

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

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

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

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

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

	Data “A”			Data “B”			
y	x1	x2	x3	x1	x2	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？

PLSは優れた特質を有する線形回帰手法である.

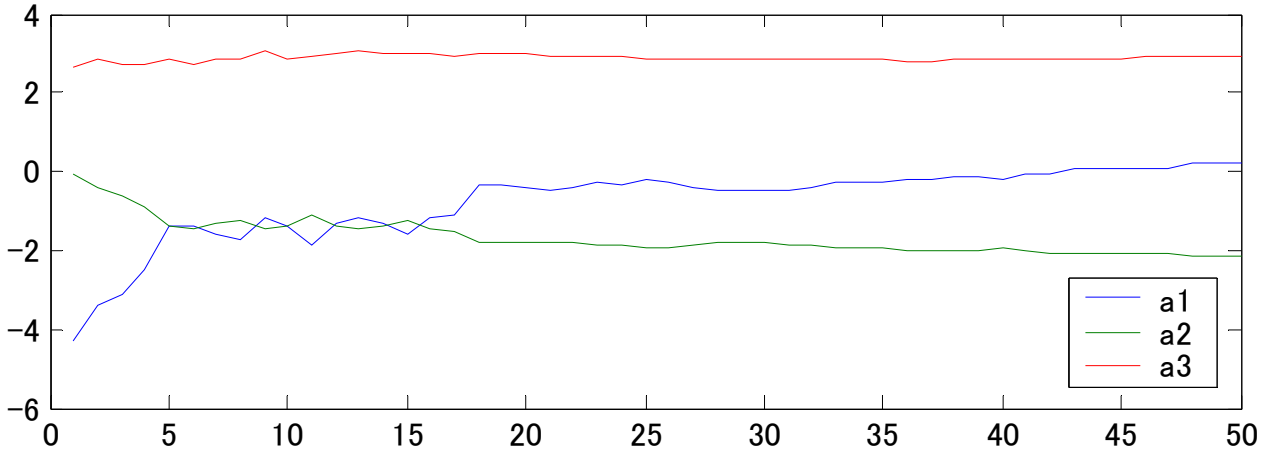
最小二乗法を使うメリットは何もないのだろうか？

＜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);
```

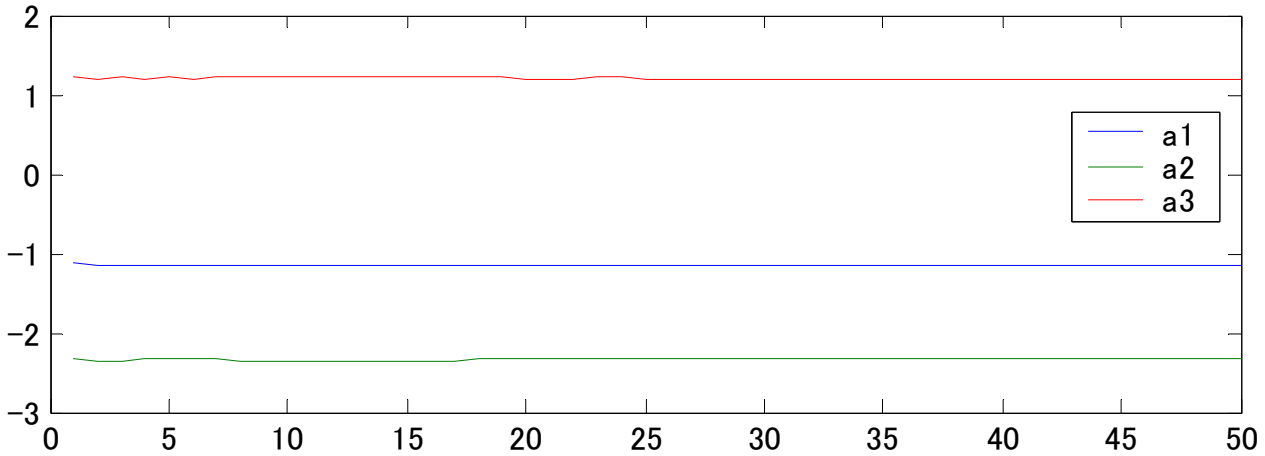
パラメータのサンプル数依存性

OLS



PLS

nLV=1



サンプル数 [x5]

まとめ

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.