PLS 回帰におけるモデル選択

橋本淳樹* 田中豊†

概要

PLS 法は最初に開発された計量化学の分野では勿論のこと，他の分野においても応用面でその性能が高く評価され広く用いられている．一方で，PLS 法の統計的な性質についてはまだまだ理論的に整理されたとは言えない段階であり，そのモデル選択についても解析者の判断に委ねられることも多い．本研究ではモデル選択の問題に焦点を当てシミュレーション実験により評価した．シミュレーションの結果は，一般に良く用いられるクロスバリデーションの PRESS 最小の基準はやや高次元のモデルに導く傾向があり，本研究で扱った選択基準 (Osten の F 基準や Wold の R 基準) の方がより良いモデルを選択できることを示した．

1 はじめに

重回帰分析をする際，説明変数の中に互いに相関が高い変数が含まれる場合，通常の最小 2 乗法 (Ordinary Least Squares :OLS) では回帰係数の推定精度悪くなることがあり，多重共線性の問題がある．このような問題がある場合の手法として，主成分回帰 (Principal Component Regression :PCR)，リッジ回帰 (Ridge Rgression :RR) とともに計量化学の分野で開発された Partial Least Squares Regression(PLS 回帰) が知られている． PLS 法は，計量化学 (chemometrics) の分野で開発され，その分野では最も良く用いられている回帰分析手法である．応用面からその性能が高く評価されており近年では解析的な面も含めて研究が盛んに行われている．

本研究では，PLS 回帰におけるモデル選択，すなわち，いくつの成分までを用いてモデルを構築すれば将来の観測値に対しても予測誤差を最小にできるか，また，最も良く現象を説明できるかということに焦点を当てる．クロスバリデーションを基にした選択基準をシミュレーション実験で評価し，データ解析においてはブートストラップ法を用いた予測誤差の推定も行う．また，PLS 回帰における成分数ごとの回帰係数の傾向についてもシミュレーション実験を行う．

2 Partial Least Squares

PLS 法は, 計量化学の分野で Wold(1975) によって開発され，その分野でよく用いられている回帰分析手法である．計量化学では，スペクトルの検量などサンプルサイズに比べて圧倒的に波長数 (変量) が多い場合や変数間の共線性が高い場合に有用とされている．また近年では, 回帰分析の精度を高める目的だけでなく, 次元削減あるいは関連因子の抽出といった用法としても注目を集めている．

PLS 回帰はデータをそのまま使わずにスコア (潜在変数, 成分とも呼ばれる) を計算し，そのスコアへの回帰を行う点が通常の重回帰と異なる．スコアを計算する際の重みは, スコアと従属変数の共分散が最も高くなる

* 南山大学数理情報研究科
† 南山大学情報理工学部

ようにし，かつ，スコアが互いに無相関となるように逐次求めていく．そして得られたスコアの一部に対して最小２乗法で係数を推定していく手法である．PLS 回帰は予測性能という点ではリッジ回帰にわずかに劣るものの (Frank & Friedman(1993))，高次元データを従属変数と関連の強い低次元データへ変換するという特徴を持つ．次元縮小という点で主成分回帰と類似の手法と言えるが，PLS 回帰の方が低次元で予測精度の高いモデルを構築できる．また，変数の数が個体数より大きくなるような場合に PLS 法が適用できることも計量化学で広く用いられている理由のひとつと言える．

PLS 回帰のアルゴリズムはいくつか提案されており (SIMPLS アルゴリズム (De jong(1993))，KERNEL アルゴリズム (Rannar, et al.(1994)) など)，本研究では最も代表的な NIPALS アルゴリズム (Wold(1975)) を用いる．

step0　説明変数 $\boldsymbol{X}$と従属変数$\boldsymbol{y}$ を中心化 (または標準化) して $\boldsymbol{X}_0, \boldsymbol{y}_0$ とし,$\hat{\boldsymbol{y}}_0$=$\boldsymbol{0}$ とする．

step1　$\boldsymbol{X}_0$と$\boldsymbol{y}_0$ の共分散として重み $\boldsymbol{w}_1 = \boldsymbol{X}_0^T\boldsymbol{y}_0$ を計算し，スコア $\boldsymbol{t}_1 = \boldsymbol{X}_0\boldsymbol{w}_1$ を求める．

step2　$\boldsymbol{y}_0$ を $\boldsymbol{t}_1$ 上へ回帰して，回帰モデルを $\hat{\boldsymbol{y}}_1 = \hat{\boldsymbol{y}}_0 + \boldsymbol{t}_1(\boldsymbol{t}_1^T\boldsymbol{t}_1)^{-1}\boldsymbol{t}_1^T\boldsymbol{y}_0$ と更新する．

step3　回帰モデルの精度が十分でなければ，スコア上へ回帰した時の残差 $\boldsymbol{X}_1, \boldsymbol{y}_1$ を計算し，添え字を一つずつ増加させて step1〜3 を十分な精度が得られるまで繰り返す．

$$\boldsymbol{X}_1 = (\boldsymbol{I} - \boldsymbol{t}_1(\boldsymbol{t}_1^T\boldsymbol{t}_1)^{-1}\boldsymbol{t}_1^T)\boldsymbol{X}_0$$
$$\boldsymbol{y}_1 = (\boldsymbol{I} - \boldsymbol{t}_1(\boldsymbol{t}_1^T\boldsymbol{t}_1)^{-1}\boldsymbol{t}_1^T)\boldsymbol{y}_0$$

また，k 成分モデルにおける PLS 回帰係数 $\boldsymbol{\beta}_{PLS}^k$ は，重み行列 $\boldsymbol{W}$，係数行列 $\boldsymbol{P}, \boldsymbol{q}$ を用いて以下のように求めることができる (Helland(1988))．

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{PLS}^k \tag{1}$$
$$= \boldsymbol{X}\boldsymbol{W}_k(\boldsymbol{P}_k^T\boldsymbol{W}_k)^{-1}\boldsymbol{q}_k \tag{2}$$

ここに, $\boldsymbol{W}_k = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_k)$ は NIPALS アルゴリズムで得られる重みベクトルを列に持つ行列, それぞれのスコアを列に持つスコア行列 $\boldsymbol{T}_k$ へ $\boldsymbol{X}, \boldsymbol{y}$ を射影したときの係数 $\boldsymbol{P}_k = \boldsymbol{X}^T\boldsymbol{T}_k(\boldsymbol{T}_k^T\boldsymbol{T}_k)^{-1}$, $\boldsymbol{q}_k = (\boldsymbol{T}_k^T\boldsymbol{T}_k)^{-1}\boldsymbol{T}_k^T\boldsymbol{y}$ である．

3　成分数の選択

実際に PLS 回帰をする際，最終的な回帰モデルに含まれるスコアの数を決定する必要があり，一般に良く用いられる方法としてクロスバリデーションがある．クロスバリデーション (Cross-Validation :CV) とはモデルの安定性を調べる手法のひとつである．データセットを G 個のグループに分割して，$G-1$ 個のグループを用いてモデルの構築を行い，残った 1 個のグループを用いてモデルの評価を行う．この操作をすべてのグループが 1 個ずつ評価データとして用いられるように繰り返す手法である．本論文では Leave-one-out でクロスバリデーションを行う．Leave-one-out とはデータセットのすべてのサンプルについて，そのサンプルを取り除いてモデル構築を行い評価する方法である．そのためサンプルサイズ N のデータセットでは，N 回の繰り返しを行うことになる．

$$PRESS_{(k)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{(i)}^k)^2 \tag{3}$$

$\hat{y}_{(i)}^k$ は i 番目の個体を除いて推定された k 成分モデルにおける i 番目の個体の予測値.

一般的に用いられる基準は，(3) 式の PRESS が最小となるように成分数 k を決定することである．クロスバリデーションは，大ざっぱに言えばバイアスはないが，ばらつきが大きいとされており (Efron & Tibshirani(1993))，しばしば必要以上に高次元のモデルを選択する．そのため，クロスバリデーションを利用したいくつかの選択基準が提案されており，本研究では Wold's R criterion，Krzanowski's W criterion，Osten's F criterion を用いる．

3.1 Wold's R criterion

Wold's R criterion は PRESS が局所的に最小値をとる最小の成分数 k を選択する基準である (Wold(1978))．

$$R_k = \mathrm{PRESS}_{(k+1)} / \mathrm{PRESS}_{(k)} \tag{4}$$

$R_k > 1$ となる最初の k が最適な成分数となる．また閾値を 0.95,0.9 とする adjusted Wold's R criterion が Krzanowski(1987) によって提案されている．

3.2 Krzanowski's W criterion

Krzanowski's W criterion は以下の W が 1 より大きい成分を選択する基準である (Eastment & Krzanowski(1982))．

$$W_k = \left(\mathrm{PRESS}_{(k-1)} - \mathrm{PRESS}_{(k)}\right) \div \frac{\mathrm{PRESS}_{(k)}}{n-1-k} \tag{5}$$

(5) 式で，右辺の被除数は k 成分を追加した時の予測誤差平方和の減少量であり，除数は自由度 1 あたりの予測誤差平方和の平均である．k 成分の予測誤差平方和の減少量が，残りの成分の減少の平均より大きければその成分を有意とするということである．また，Krzanowski はサンプリングによるばらつきを許容するために，W が 0.9 以上となる最大の成分数 k を最適な成分数とすることを提案しており (Krzanowski(1987))，我々をこの基準を adjusted Krzanowski's W criterion と呼ぶことにする．

3.3 Osten's F criterion

Osten F criterion は以下の F 統計量が自由度 $(1, n-1-k)$ の F 分布の 0.95 点よりも大きい成分を選択する基準である (Osten(1988))．

$$F_k = \left(\mathrm{PRESS}_{(k-1)} - \mathrm{PRESS}_{(k)}\right) \div \frac{\mathrm{PRESS}_{(k)}}{n-1-k} \tag{6}$$

これは上記の W 統計量に等しく，Krzanowski's W criterion を F 検定することで成分数を決定する基準といえる．

4 予測誤差のシミュレーション実験とデータ解析

4.1 データ作成方法

説明変数 $\boldsymbol{X}$ は特異値分解 $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ を元に作成し，100 サンプル 10 変数とする．

手順 1 $\boldsymbol{z}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$ を多変量正規乱数で 100 個生成し行列 $\boldsymbol{Z}$ とする．$\boldsymbol{Z}$ の固有値分解により，固有値 $\boldsymbol{\Lambda}$ と固有ベクトル $\boldsymbol{W}$ を求め，$\boldsymbol{U} = \boldsymbol{Z}\boldsymbol{W}\boldsymbol{\Lambda}^{-1/2}$ として正規直交行列 $\boldsymbol{U}$ を作成する．

手順 2　特異値行列 $\boldsymbol{D}$ の i 番目の対角要素 d_{ii} を $d_{ii} = 1/i^3, i = 1, \ldots, 10$ とする．

手順 3　[-1,1] の一様乱数を要素とする 10 次元ベクトル $\boldsymbol{v}_i$ を 10 個生成し，最初のベクトルをノルム 1 に基準化した後，逐次グラムシュミットの直交化を用いて正規直交行列 $\boldsymbol{V}$ とする．

手順 4　$\boldsymbol{U}, \boldsymbol{D}, \boldsymbol{V}$ を用いて $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ とする．

従属変数は説明変数の一部の情報と相関が高くなるように作成する．情報量の大きさ (特異値の大きさ) の違う case1〜3 で作成し，どういった状況で PLS 回帰が有効であるのかを検討する．

case 1　1,2,3 番目に大きい特異値に対応する左特異ベクトルを用いて以下のように作成

$$\boldsymbol{y} = \boldsymbol{u}_1 + \boldsymbol{u}_2 + \boldsymbol{u}_3 + \boldsymbol{\varepsilon} \tag{7}$$

ここに,$\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \mathrm{var}(\boldsymbol{u}_1 + \boldsymbol{u}_2 + \boldsymbol{u}_3)/10)$ である．

case 2　3,4,5 番目に大きい特異値に対応する左特異ベクトルを用いて case 1 と同様に作成

case 3　1,2,3,5,7 番目に大きい特異値に対応する左特異ベクトルを用いて case 1 と同様に作成

以下の選択基準でモデルを決定し，選択基準の性能をシミュレーション実験により比較する．

(基準 1)　absolute minimum PRESS

(基準 2)　Wold's R criterion (threshold=1)

(基準 3)　adjusted Wold's R criterion (threshold=0.95)

(基準 4)　adjusted Wold's R criterion (threshold=0.90)

(基準 5)　Krzanowski's W criterion

(基準 6)　adjusted Krzanowski's W criterion (threshold=0.90)

(基準 7)　Osten's F criterion

4.2　シミュレーション実験 結果

各 case におけるシミュレーション実験 100 回の結果を表 1，2，3 に示す. 今回取り上げた 6 つの選択基準と一般的に用いられることが多い CV をして得られた PRESS が最小となる成分数を選択する基準を含めた計 7 つの選択基準を用いた．また，PLS 回帰をするときの変数の基準化については中心化のみを行った．計算は，統計ソフト R の Package *pls* の関数 plsr で行った．

表 1 の case1 実験結果では，従属変数の真の構造は 3 次元で構成されており MSE においても 100 回の実験においてすべて 3 成分が選択されている．表 1 より，すべての基準で最頻値は 3 成分である．しかし，R 基準と Osten's F 基準が 80% あるいは 90% 以上が 3 成分を選択しているのに対して，PRESS を最小とする基準では成分数を若干多めに見積もる傾向が見られ，W 基準については平均値が 5 成分を超えており適当に選択できてないと言える．

case2 では，R 基準 (とくに，adjusted R 基準) が精度良く成分数を選択できていることがわかる．この case においては，W 基準を除いて他のすべての基準がうまく機能したと言える．

case3 では，従属変数が 5 次元で構成されているが MSE[*1] による最適な成分数は 7 成分 (あるいは 6 成分) となっている．これは PLS 回帰によって得られるスコアは従属変数との共分散が最大になるように求めているため，説明変数のごく一部の小さい情報は従属変数の相関が高い場合でも早い段階でスコアとして抽出されないということが考えられる．そのため,5 成分では十分な精度のモデルが構築できず，MSE を見ると 6,7

表 1　実験 100 回における選択された成分数 **(case 1)**

	number of components											
selection criterion	1	2	3	4	5	6	7	8	9	10	mode	mean
absolute minimum $PRESS$	0	0	67	13	2	4	4	3	3	4	3	4.06
Wold's R criterion	0	0	80	16	2	2	0	0	0	0	3	3.26
adjusted R crterion(0.95)	0	0	88	10	1	1	0	0	0	0	3	3.15
adjusted R crterion(0.90)	0	0	91	7	2	0	0	0	0	0	3	3.11
Krzanowski's W criterion	0	0	38	7	8	14	12	13	8	0	3	5.26
adjusted W crterion(0.90)	0	0	37	8	8	14	12	11	10	0	3	5.29
Osten's F criterion	0	0	91	2	3	2	2	0	0	0	3	3.22
MSE[*1]	0	0	100	0	0	0	0	0	0	0	3	3.00

表 2　実験 100 回における選択された成分数 **(case 2)**

	number of components											
selection criterion	1	2	3	4	5	6	7	8	9	10	mode	mean
absolute minimum $PRESS$	0	0	0	0	91	2	1	2	2	2	5	5.28
Wold's R criterion	0	0	0	0	96	2	0	2	0	0	5	5.08
adjusted R crterion(0.95)	0	0	0	0	100	0	0	0	0	0	5	5.00
adjusted R crterion(0.90)	0	0	0	0	100	0	0	0	0	0	5	5.00
Krzanowski's W criterion	0	0	0	0	57	0	12	13	18	0	5	6.35
adjusted W crterion(0.90)	0	0	0		55	0	13	13	19	0	5	6.41
Osten's F criterion	0	0	0	0	92	0	5	3	0	0	5	5.19
MSE[*1]	0	0	0	0	100	0	0	0	0	0	5	5.00

成分のモデルが妥当であると考えられる．各選択基準の結果は最頻値がすべて 7 成分となった．図 1 より，Osten's F 基準が最も精度良く MSE と近い分布を示していることがわかる．一方で，R 基準に関しては 1,2,3 成分といった予測精度の十分でない成分数を選択している．

4.3　gasoline データの解析

gasoline データ (R の Package *pls* のサンプルデータ) は，ガソリンに含まれる成分オクタンと，近赤外線スペクトルのデータである．オクタンの成分量を従属変数，401 波長で測定それたスペクトルを説明変数として PLS 回帰分析する．サンプルサイズは 60 である．変数の数が観測数よりも多く，かつ，互いに変数間の相関が高い，計量化学における典型的なデータと言える．また，説明変数の最大特異値と最小特異値の比で多重共線性の強さを示す condition number は 22233.57 で強い共線性がある．

前節で取り上げた選択基準は基礎となる統計量として PRESS を用いている予測誤差としている．これに対して，ブートストラップ法を用いて予測誤差を推定する方法が考えられる．gasoline データに対して前節のシミュレーションで取り上げた選択基準のほか，以下の 3 つのブートストラップ予測誤差 (Efron &

表 3　実験 100 回における選択された成分数 **(case 3)**

	number of components											
selection criterion	1	2	3	4	5	6	7	8	9	10	mode	mean
absolute minimum $PRESS$	0	0	0	0	0	0	62	23	6	9	7	7.62
Wold's R criterion	1	1	10	0	0	0	57	21	5	5	7	6.95
adjusted R crterion(0.95)	2	3	15	0	0	1	58	16	4	1	7	6.39
adjusted R crterion(0.90)	8	4	19	0	0	3	52	12	2	0	7	5.69
Krzanowski's W criterion	0	0	0	0	0	1	75	16	6	2	7	7.33
adjusted W crterion(0.90)	0	0	0	0	0	1	73	16	7	3	7	7.38
Osten's F criterion	0	0	0	0	0	16	82	2	0	0	7	6.86
MSE[*1]	0	0	0	0	0	27	73	0	0	0	7	6.73

[*1] Mean Squared Error: 従属変数の真の値と PLS 回帰による予測値の差の 2 乗和を最小にする成分数

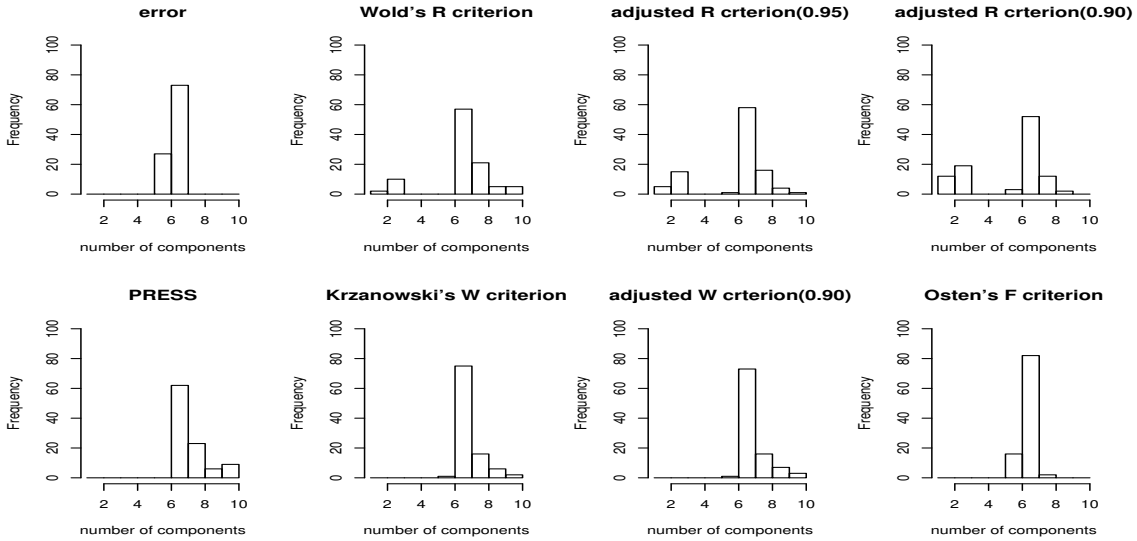

図 1　選択された成分数のヒストグラム (case 3)

Tibshirani(1993)) を利用したモデル選択を含めて成分数の検討を行う.

- **simple bootstrap estimate**

  ブートストラップ標本から推定された回帰関数を，元の標本に応用して予測誤差を推定する.

$$\mathrm{err}^{\mathrm{simple}} = \frac{1}{B}\sum_{b=1}^{B}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \eta_{X^{*b}}(\boldsymbol{x}_i)\right)^2\right] \tag{8}$$

  ここに，$(\boldsymbol{x}_i, y_i)$ は i 番目の標本，X^* はブートストラップ標本，$\eta_{X^{*b}}$ は b 番目のブートストラップ標本から推定された回帰関数である.

- **optimism bootstrap estimate**
  推定に用いたデータそのものに当てはめを行うことによって予測誤差が小さく見積もられる (optimism) 量を推定して，それを残差平方和に加えて求めた予測誤差.

$$\mathrm{err}^{\mathrm{opt}} = \frac{RSS}{n} + \hat{\omega}(\hat{F}) \tag{9}$$

$$\hat{\omega}(\hat{F}) = \frac{1}{B}\sum_{b=1}^{B}\left[\frac{1}{n}\sum_{i=1}^{n}\Big\{\Big(y_i - \eta_{X^{*b}}(\boldsymbol{x}_i)\Big)^2 - \Big(y_i^{*b} - \eta_{X^{*b}}(\boldsymbol{x}_i^{*b})\Big)^2\Big\}\right] \tag{10}$$

  ここに，RSS は元の標本における残差平方和である.

- **0.632 bootstrap estimate**
  ブートストラップ標本に i 番目の個体 $(\boldsymbol{x}_i, y_i)$ が含まれる確率は，$n \to \infty$ のとき，

$$\begin{aligned}\mathrm{Prob}\Big((\boldsymbol{x}_i, y_i) \in (\boldsymbol{x}^{*b}, y^{*b})\Big) &= 1 - \left(1 - \frac{1}{n}\right)^n \\ &\xrightarrow{n\to\infty} 1 - e^{-1} \cong 0.632\end{aligned} \tag{11}$$

  となる．つまり，平均的に各ブートストラップ標本に元のデータの約 37% が含まれないことになる．(11) 式をもとに，予測誤差の 0.632 ブートストラップ推定量は以下のように定義される.

$$\begin{aligned}\mathrm{err}^{0.632} =& 0.368 \cdot \frac{RSS}{n} \\ &+ 0.632 \cdot \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{B_i}\sum_{b\in C_i}\Big(y_i - \eta_{X^{*b}}(\boldsymbol{c}_i)\Big)^2\right]\end{aligned} \tag{12}$$

  ここに，$(\boldsymbol{c}_i, y_i)$ は b 番目のブートストラップ標本に含まれない i 番目の標本，C_i は i 番目の標本が含まれないブートストラップ標本の番号集合，B_i は i 番目のデータが含まれないブートストラップ標本の総数である.

4.3.1 gasoline データの解析結果

図 2 に PRESS とブートストラップ法による予測誤差 (試行 100 回) の各成分数の推移を示し，表 4 に各選択基準を適用して求められた成分数を示した.

図 2 の PRESS の推移を見ると 3 成分のところで PRESS の減少が緩やかになっているのがわかり，最小値は 7 成分になった．また，0.632 ブートストラップ推定による予測誤差は, バイアスがないもののばらつきは大きいとされる PRESS の曲線に近いところで滑らかな曲線を描いており，精度よく推定できていると考えられる．予測誤差のブートストラップ推定量では，3 手法とも 7 成分モデルが最小となった．各選択基準の結果は，シミュレーション実験と同様に R 基準は少なめ，W 基準は多めに成分数を選択しており，Osten's F 基準は予測精度の高いモデルを選択できていると考えられる.

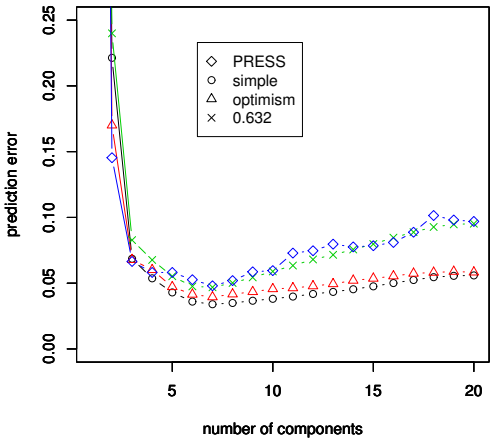

図 2　PRESS とブートストラップ法による予測誤差

表 4　各選択基準による成分数

selection criterion	components
absolute minimum $PRESS$	7
Wold's R criterion	4
adjusted R crterion(0.95)	4
adjusted R crterion(0.90)	4
Krzanowski's W criterion	19
adjusted W crterion(0.90)	19
Osten's F criterion	7

5　PLS 回帰係数のシミュレーション実験

5.1　シミュレーションデータ作成方法

説明変数については前節と同様 (特異値分解の行列 $\boldsymbol{D}, \boldsymbol{V}$ については固定) に作成し，従属変数については先の実験の case1，つまり，$\{\alpha, \beta, \gamma\} = \{1, 2, 3\}$ をシミュレーションする．また，真の係数については以下のように固定されている．

説明変数が以下のような構造で生成されている．

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}^*\boldsymbol{V}^{*T} \tag{13}$$

ここに，$\boldsymbol{D}^*, \boldsymbol{V}^*$ はある値に固定されており，$\boldsymbol{U}$ のみを乱数で生成する．このとき，従属変数 $\boldsymbol{y}$ は，

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{u}_1 + \boldsymbol{u}_2 + \boldsymbol{u}_3 + \boldsymbol{e} \\ &= \boldsymbol{X}(\frac{1}{d_1^*}\boldsymbol{v}_1^* + \frac{1}{d_2^*}\boldsymbol{v}_2^* + \frac{1}{d_3^*}\boldsymbol{v}_3^*) + \boldsymbol{e} \\ &= \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{e} \end{aligned} \tag{14}$$

ここに，$\boldsymbol{\beta}^*$ は真の係数である．

100 個体 10 変数の説明変数行列でコンディションナンバーを 1000 に設定したシミュレーション実験を 1000 回繰り返した (すべての試行において真の係数は固定されている)．

5.2 シミュレーション実験 結果

図 3 に予測値と係数の平均二乗誤差 (MSE) を成分数ごとに図示した．

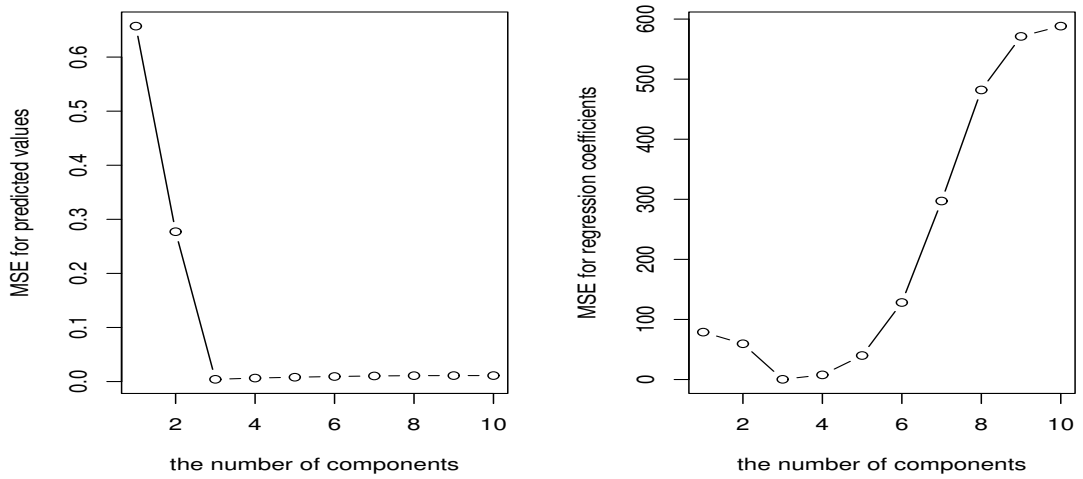

図 3　予測値と係数の平均二乗誤差

図 3 の左の図を見ると，3 成分のところで最小値をとり，そこからわずかにながら徐々に MSE は大きくなっていく．右の図では，3 成分のところで明らかな最小値をとっていることがわかる．次に，変数 1 に対する係数のヒストグラムを示す．

図 4 では，真の値を太い線で示してある．ヒストグラムを見ると，1,2 成分では真の値を含まず，3 成分以降では真の値を区間に含んでいることが分かる．また，3 成分 (また，4 成分も考えられる) においては比較的に係数の分散も小さく真の値を含み良いモデルと言えるが，成分数が多くなるにつればらつきが大きくなり真の係数は正の値にも関わらず負の係数もとりうることが分かる．

このシミュレーションから，予測を目的にした回帰分析では PRESS 最小の基準でもうまくモデル選択できるといえるが，要因分析のような目的の場合には係数の安定性が高く信頼度の高いモデルを選択するためにより慎重なモデル選択を迫られる．よって，予測誤差を最小とすることを通してモデル選択した場合に，係数の安定性を考慮すると Osten の F 基準などより安定したモデル選択基準を用いる必要がある．

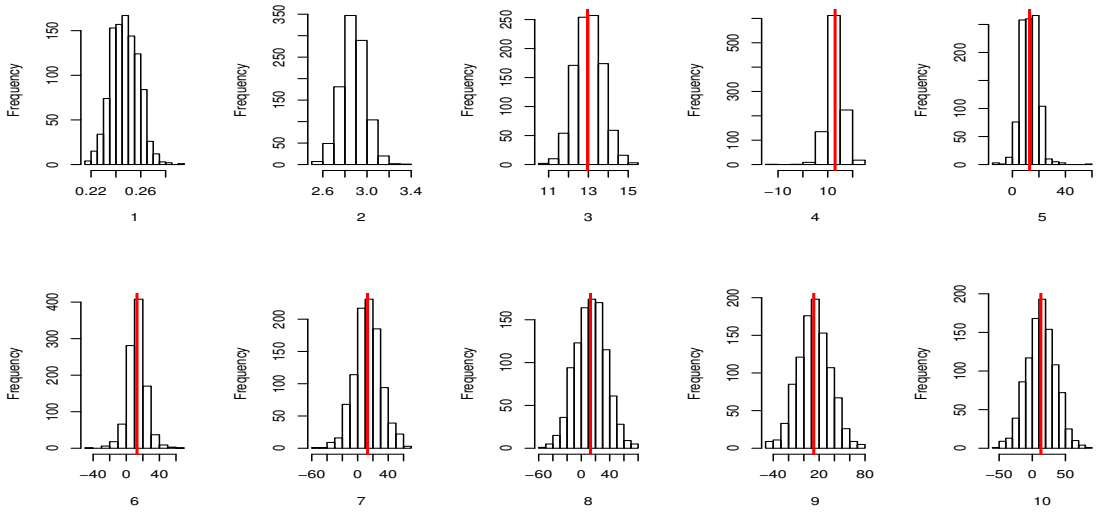

図 4　変数 1 に対する PLS 回帰係数のヒストグラム

6　おわりに

本研究では，PLS 法について文献を調査し，その性能をいくつかの視点から研究した．主に PLS 回帰における成分数の選択というテーマに焦点を当て，クロスバリデーションを基にした選択基準についてシミュレーション実験等でその性能を考察した．その結果，いくつか提案されている選択基準のうち，Osten の F 基準が最も安定した性能を示した．また，Wold の R 基準についても特定の場合を除いて，うまくモデル選択できることが示されており，PRESS 最小というシンプルな基準よりも本研究で扱ったような基準 (Osten の F 基準や Wold の R 基準) を利用したモデル選択が推奨される．

PLS 回帰の係数のシミュレーション実験では，予測値の MSE よりも回帰係数の MSE の方がモデル選択による影響を受けやすいことを示した．予測値の MSE では，ある成分数において最小値をとりその後緩やかに MSE が増加していくがその増加はそれほど大きくない．一方で，係数の MSE においてはある成分数のモデルで明らかな最小値 (最小となる成分数は予測値の MSE と同じ) をとる．つまり，予測誤差を最小にするモデル選択を通して係数についても安定した良い推定値を得ようとすると，より慎重なモデル選択を要求されることになる．PRESS 最小の基準ではやや多めの成分数を選択する傾向があり，予測誤差に関しては最適なモデルとほぼ同程度のモデルを選択できているかもしれないが，係数の MSE については大きく劣るモデルを選択する可能性があると言える．

PLS 法は多くの多変量解析の場面に応用され成果を上げており，利用価値の高い手法であると言える．一方で，まだまだ理論的な解釈に関して未解決な部分もあり，統計学の分野において，とくに日本においてはそれほど良く知られた手法ではないが，理論面での研究が進めばより多くの人を関心を引くだろう．

参考文献

[1] De Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression, *Chemometrics and Intelligent Laboratory Systems 18*, 251-263.

[2] Eastment, H. T., Krzanowski, W. J. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics 24*, 73-77.

[3] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the bootstrap*. Chapman & Hall.

[4] Frank,I. and Friedman, F. (1993). A statistical view of some chemometrics regression tools, *Technometrics 35*, 134-135.

[5] Golub, G. & Loan, C.V. (1996). Matrix Computations, 3rd, Johns Hopkins Univ Press.

[6] Helland, I. S. (1988). On the Structure of Partial Least Squares Regression. *Communications in Statistics - Simulation and Computation 17*, 581-607.

[7] Helland, I.S. (2000). Some theoretical aspects of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems 58*, p.97-107.

[8] Hoskludsson, A. (1988). PLS regression methods, *Journal of chemometrics 2*, 211-228.

[9] Krzanowski, W. J. (1987). Cross-validation in principal component analysis. *Biometrics 43*, 575-584.

[10] Li, B., Morris, J., & Martin, E. B. (2002). Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems 64*, 79-89.

[11] Lindgren, F., Geladi, F., Wold, S. (1993). The Kernel algorithm for PLS, *Journal of Chemometrics 7*, 45 - 59

[12] Lingjarde, O. & Christophersen, N. (2000). Shrinkage Structure of Partial Least Squares. *Scandinavian Journal of Statistics 27*, p.459-473.

[13] Osten, D. W. (1988). Selection of optimal regression models via cross-validation. *Journal of Chemometrics 2*, 39-48.

[14] Rannar, S., Lindgren, F., Geladi, P., Wold, S. (1994). A PLS Kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. *Journal of Chemometrics 8*, 111-125.

[15] Rosipal, R., & Kramer, N. (2006). Overview and Recent Advances in Partial Least Squares. Springer.

[16] Wold, H. (1975). Soft Modeling by Latent Variables: the Nonlinear Iterative Partial Least Squares Approach, in *Perspective in Probability and Statistics*, Paper in Honour of M. S. Bartlett, 520-540, Academic Press.

[17] Wold, S. (1978). Cross-validation estimation of the number of components in factor and principal component analysis. *Technometrics 24*, 397-405.