Partial Least Squaresについて
～Biased Regressionとしての統計的性質～

M2008MM009 橋本淳樹
指導教員：田中豊

1 はじめに

重回帰分析の際，多重共線性の問題がある場合の手法として，主成分回帰やリッジ回帰があり，その他に広く応用されているわけではないが，計量化学の分野で開発された偏最小２乗回帰 (Partial Least Squares Regression:PLSR) がある．PLS 回帰をする際には，いくつの成分を用いてモデルを構築すればよいかが重要な問題のひとつとなる．先行研究でいくつかの選択基準が提案 (または回帰分析，主成分分析で用いられている基準が適用) されている．それらの研究 (例えば，Li(2002)) では，従属変数の真の構造がいくつの次元で構成されているかという点で各選択基準を評価しているが，本研究では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という点で比較を行い各選択基準の性能を評価する．また，PLS 回帰係数の振舞いについてもシミュレーションする．

2 Partial Least Squares

偏最小２乗回帰は，計量化学の分野で Wold(1975) によって開発され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，次元縮小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．

以下に最も代表的な NIPALS アルゴリズムを紹介する．

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．

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

3 PLS の代数的な表現

Helland(1988) はクリロフ部分空間法を用いた PLS の代数的な表現を示した．$m \geq 1$ について，クリロフ部分空間は，

$$\mathscr{K}_m = \mathrm{span}\{X^Ty, (X^TX)X^Ty, \cdots, (X^TX)^{m-1}X^Ty\} \tag{1}$$

と定義され，これは PLS で求められる最初の m 個のスコアの張る空間と同じ部分空間となる．つまり，PLS 回帰係数 $\boldsymbol{\beta}$ の推定量は以下の制約付き最適化問題の解となる．

$$\begin{aligned} &\text{minimize} \quad \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2 \\ &\text{subject to} \quad \boldsymbol{\beta} \in \mathscr{K}_m \end{aligned} \tag{2}$$

また，クリロフ部分空間 $\mathscr{K}_m$ の正規直交基底を列に持つ行列を R_m とすると，PLS 回帰係数は以下のように表現される．

$$\hat{\boldsymbol{\beta}}_{PLS} = R_m(R_m^TX^TXR_m)^{-1}R_m^TX^T\boldsymbol{y} \tag{3}$$

4 選択基準

PLS 回帰において最終的なモデルを決定するために，最適な成分数 (次元数) がいくつであるかが問題となる．一般的に，クロスバリデーション (Cross-Validation :CV) を行い，最適な成分数を求める方法がとられている．あらかじめ十分な成分数のスコアを計算しておき，それぞれの成分数のモデルに対して CV を行い，PRESS が最小となる最適なモデルを決定する．

$$PRESS_{(k)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{(i)}^k)^2 \tag{4}$$

$\hat{y}_{(i)}^k$ 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．

4.1 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，つまり，$R_k = PRESS_{(k+1)}/PRESS_{(k)}$ が閾値 1 以上となる成分数 k を選択する基準である (Wold(1978))．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．

4.2 Krzanowski's W criterion，Osten's F criterion

Krzanowski's W criterion は下記の W_k が 1 より大きい成分を選択する基準である．

$$W_k = \Big(PRESS_{(k-1)} - PRESS_{(k)}\Big) \Big/ \frac{PRESS_{(k)}}{n-1-k} \quad (5)$$

(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 成分の情報量が，残りの情報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，Eastment & Krzanowski(1982) 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いる．

また，Osten(1988) は上記の W_k が $F_{(1,n-1-k,0.95)}$ より大きいときにそのモデルを有意とすることを提案している．

5 シミュレーション実験

5.1 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 $\boldsymbol{z}_i \sim \mathrm{N}(\boldsymbol{0}, \mathbf{I}_{10})$ 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 case1,2(その他の状況についてもシミュレーションを行ったが、紙面の都合上割愛する) で作成し，どういった状況で各選択基準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

$$\boldsymbol{y} = \boldsymbol{u}_1 + \boldsymbol{u}_2 + \boldsymbol{u}_3 + \boldsymbol{\varepsilon} \quad (6)$$

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 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 case1 と同様に作成

5.2 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 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

表 1 試行 100 回における選択された成分数 **(case 1)**

	the number of components											
selection criterion	1	2	3	4	5	6	7	8	9	10	mode	mean
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 試行 100 回における選択された成分数 **(case 2)**

	the number of components											
selection criterion	1	2	3	4	5	6	7	8	9	10	mode	mean
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

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．

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切に選択できてないと言える．

case2 では，MSE を見ると 6，7 成分のモデルが適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．表 2 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．

次に，PLS 回帰における係数の傾向についてシミュレーション結果を図 1,2 に示す．説明変数は先の手順における行列 $\boldsymbol{D}, \boldsymbol{V}$ を固定された $\boldsymbol{D}^*, \boldsymbol{V}^*$ を用いて生成し，従属変数については case1 と同様である．図 1 を見ると予測誤差が最小になるのは 3 成分モデルであるがその後の増加はわずかなものであるのに対して，係数の平均二乗誤差は明確に 3 成分モデルが最小であることを示す．つまり，予測誤差がほぼ同程度のモデルにおいても係数の平均二乗誤差の点では明らかに劣るモデルが存在する．図 2 では 1 番目の説明変数に対する回帰係数を成分数ごとにヒストグラムで示した．ヒストグラムの太線は真の値を示しており，3,4 成分モデルにおいては比較的ばらつきも小さく真の値を含んでいる．1,2 成分モデルでは真の値を含んでおらず，5 成分以降のモデルではばらつきが大きく係数の符号も逆転した値が数多くあることがわかる．

6 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．

4 節で取り上げた選択基準は基礎となる統計量として CV の PRESS を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 (Efron & Tibshirani(1993)) が考えられる．gasoline データに対して 4 節で取り上げた選択基準のほか，3 つのブートストラップ予測誤差を利用した基準を含めて成分数の検討を行う．本稿では，3 つのブートストラップ法とクロスバリデーションによる予測誤差の推定のうち最も精度が高いとされている 0.632 ブートストラップ推定法を用いる．

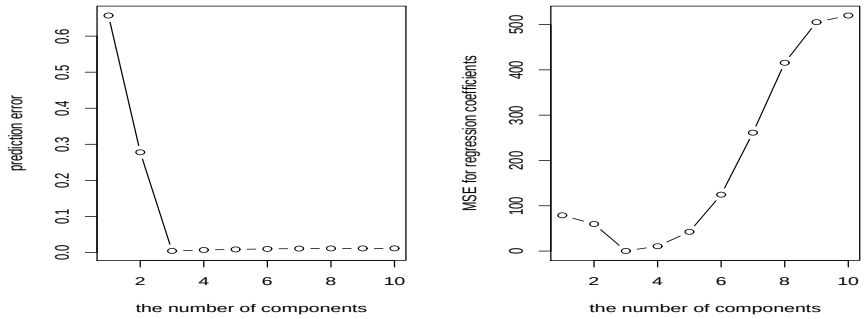

図 1　予測誤差と係数の平均二乗誤差

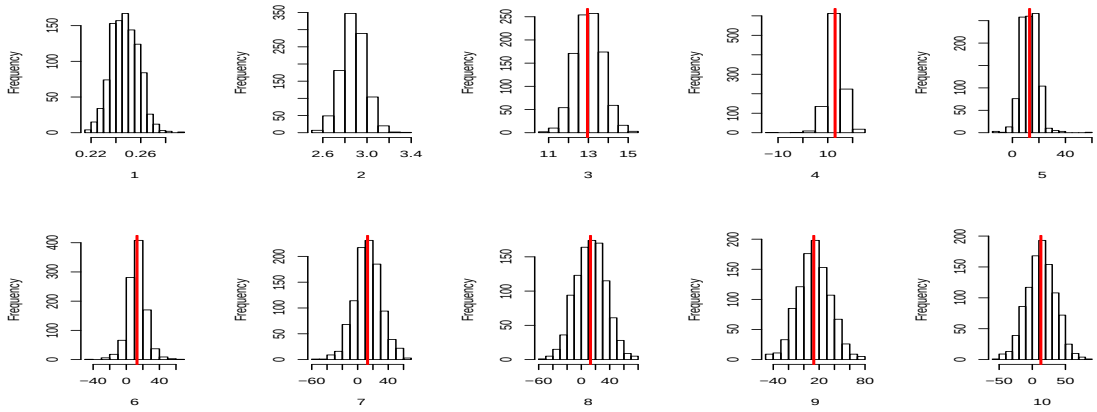

図 2　1 番目の変数に対する回帰係数のヒストグラム (試行 1000 回)

0.632 bootstrap estimate

$$\mathrm{err}^{0.632} = 0.368 \cdot \frac{\mathrm{RSS}}{n} + 0.632 \cdot \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{B_i} \sum_{b \in C_i} \left(y_i - \eta_{X^{*b}}(\mathbf{c}_i) \right)^2 \right] \tag{7}$$

ここに,$(\mathbf{c}_i, y_i)$ 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.

6.1 gasoline データの解析結果

図 3 に PRESS とブートストラップ法による予測誤差 (リサンプリング 100 回) の各成分数の推移を示し，表 3 に各選択基準を適用して求められた成分数を示した.

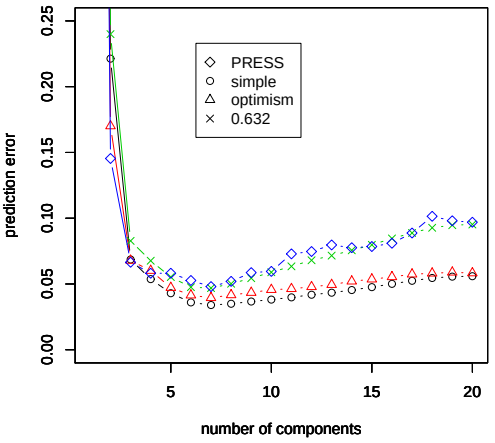

図 3 PRESS とブートストラップ法による予測誤差

表 3 各選択基準による成分数

selection criterion	components
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

図 3 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．各選択基準の結果は，シミュレーション実験と同様の傾向を示しており，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.

7 おわりに

本研究では，PLS 回帰におけるモデル選択に焦点を当て，Osten's F 基準が最も安定した性能を持つことをシミュレーションを通して示した．また，Wold's R 基準についても特定の場合を除いては精度良くモデル選択できると言える．PLS 回帰における係数のシミュレーション実験では係数の平均二乗誤差の方が予測誤差よりもモデル選択に対して敏感に反応することが示され，PRESS を最小とするモデル選択では十分とは言えず本研究で扱ったような基準 (F 基準や R 基準) を用いることが推奨される．また，PLS 回帰の計算では従属変数と共分散の高い成分が順番に求められていくため前進法によるモデル選択が適当であると考えられるが，主成分回帰を例にとると PLS 回帰よりも複雑なモデル選択法が要求され，そういった点も PLS 回帰が利用される理由と言える.

PLS 法の性質を解析的に評価することは困難であるが，その一方で計量化学の分野では広く用いられており，計算機の発展に伴う次元縮小の需要が高まっていく中で PLS 法の利用は今後さらに多くの場面に広がっていくものと期待される.

参考文献

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

[2] Frank, I. E. & Jerome H. Friedman (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics 35*, 109-135.

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

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

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

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

[7] 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.

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