偏最小２乗回帰分析について

2003MM012 橋本淳樹
指導教員: 田中豊

1　はじめに

重回帰分析をする際,多重共線性の問題がある場合の手法として,主成分回帰(Principal Components 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偏最小２乗回帰(Partial Least Squares :PLS)が知られている. 今回は,偏最小２乗回帰を,通常の重回帰(Ordinary Least Squares :OLS) や主成分回帰,リッジ回帰などと比較し検討することを目的とする.

2　偏最小２乗回帰分析

2.1　導入

偏最小２乗回帰は,計量化学(Chemometrics) の分野で最もよく用いられ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数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.PLSはデータをじかに使わずにスコアを計算し,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,スコアと応答変数の相関が最も高くなるようにし,スコアが互いに無相関となるように決定する.そのスコアに対して,部分的に最小２乗法で係数を推定していく手法である.

なお,次節以降で用いる変数$\mathbf{X}, \mathbf{y}$はすべて平均0,分散１に基準化されているものとする.

2.2　NIPALSアルゴリズム(Wold)

Woldの提案したNIPALSアルゴリズムは以下のようになる.

step 0　$\mathbf{E}_0 \leftarrow \mathbf{X}$, $\mathbf{f}_0 \leftarrow \mathbf{y}$, $\hat{\mathbf{y}}_0 \leftarrow \mathbf{0}$として計算を開始する.

step 1　$\mathbf{E}_0$と$\mathbf{f}_0$の共分散ベクトルを求め,$\mathbf{w}_1$とする.

$$\mathbf{w}_1 = \mathbf{E}_0^T \mathbf{f}_0 \tag{1}$$

step 2　$\mathbf{E}_0$と$\mathbf{w}_1$の線形結合から$\mathbf{t}_1$を得る.

$$\mathbf{t}_1 = \mathbf{E}_0 \mathbf{w}_1 \tag{2}$$

step 3　$\mathbf{E}_0$と$\mathbf{f}_0$を$\mathbf{w}_1$上へ回帰させて係数$\mathbf{p}_1, q_1$を計算する.

$$\mathbf{p}_1^T = (\mathbf{t}_1^T \mathbf{t}_1)^{-1} \mathbf{t}_1^T \mathbf{E}_0 \tag{3}$$

$$q_1 = (\mathbf{t}_1^T \mathbf{t}_1)^{-1} \mathbf{t}_1^T \mathbf{f}_0 \tag{4}$$

step 4　$\mathbf{E}_0$と$\mathbf{f}_0$を$\mathbf{t}_1$上へ回帰させたときの残差と,PLS回帰式$\hat{\mathbf{y}}_1$ への追加を行う.

$$\begin{aligned}\mathbf{E}_1 &= \mathbf{E}_0 - \mathbf{t}_1 \mathbf{p}_1^T = \mathbf{E}_0 - \mathbf{t}_1 (\mathbf{t}_1^T \mathbf{t}_1)^{-1} \mathbf{t}_1^T \mathbf{E}_0 \\ &= (\mathbf{I} - \mathbf{H}_1) \mathbf{E}_0 \end{aligned} \tag{5}$$

$$\begin{aligned}\mathbf{f}_1 &= \mathbf{f}_0 - \mathbf{t}_1 q_1 = \mathbf{f}_0 - \mathbf{t}_1 (\mathbf{t}_1^T \mathbf{t}_1)^{-1} \mathbf{t}_1^T \mathbf{f}_0 \\ &= (\mathbf{I} - \mathbf{H}_1) \mathbf{f}_0 \end{aligned} \tag{6}$$

$$\hat{\mathbf{y}}_1 = \hat{\mathbf{y}}_0 + \mathbf{t}_1 q_1 \tag{7}$$

step 5　十分な精度が得られるまで,$\mathbf{E}_a$と$\mathbf{f}_a$を用いて step 1 〜 step 4 を繰り返す.添え字はひとつずつ増加する.

2.3　Martensのアルゴリズム

H.Martensによって提案されたアルゴリズムとWoldの提案したNIPALSアルゴリズムの違いは, 係数を直交させるようにして残差を求めていくのでスコアが直交しないため, それぞれのステップで重回帰を用いて推定している点が異なる.ふたつのアルゴリズムに違いはあるが,予測値が等しいことが証明されている.

step 0　$\mathbf{E}_0^* \leftarrow \mathbf{X}$, $\mathbf{f}_0^* \leftarrow \mathbf{y}$, $a = 1$として計算を開始する.

step 1　$\mathbf{E}_{a-1}^*$と$\mathbf{f}_{a-1}^*$の共分散ベクトルを求め,$\mathbf{p}_a^*$とする.

$$\mathbf{p}_a^* = \mathbf{E}_{a-1}^{*T} \mathbf{f}_{a-1}^* \tag{8}$$

step 2　$\mathbf{E}_{a-1}^*$と$\mathbf{p}_a^*$を用いて$\mathbf{t}_a^*$を計算し,スコア行列$\mathbf{T}_a^*$得る.

$$\mathbf{t}_a^* = \mathbf{E}_{a-1}^* \mathbf{p}_a^* / \mathbf{p}_a^{*T} \mathbf{p}_a^* \tag{9}$$

$$\mathbf{T}_a^* = (\mathbf{t}_1^*, \ldots, \mathbf{t}_a^*) \tag{10}$$

step 3　$\mathbf{E}_{a-1}^*$と$\mathbf{f}_{a-1}^*$を$\mathbf{T}_a^*$上へ回帰させてローディング$\mathbf{q}_a^*$を計算し,残差$\mathbf{f}_a^*$を求める.

$$\mathbf{q}_a^* = (\mathbf{T}_a^{*T} \mathbf{T}_a^*)^{-1} \mathbf{T}_a^{*T} \mathbf{f} \tag{11}$$

$$\mathbf{f}_a^* = \mathbf{f} - \mathbf{T}_a^* \mathbf{q}_a^* \tag{12}$$

step 4　$\mathbf{E}_{a-1}^*$を$\mathbf{T}_a^*$上へ回帰させたときの残差$\mathbf{E}_a^*$を求める.

$$\mathbf{E}_a^* = \mathbf{E}_{a-1}^* - \mathbf{t}_a^* \mathbf{p}_a^{*T} \tag{13}$$

step 5　十分な精度が得られるまで,$\mathbf{E}_a$と$\mathbf{f}_a$を用いて step 1 〜 step 4 を繰り返す.

3　クロスバリデーション(Leave-one-out)

$$PRESS = \frac{1}{n} \sum_{i=1}^{n} (y - \hat{y}_{(i)})^2 \tag{14}$$

$\hat{y}_{(i)}$はi番目のサンプルを除いて係数を推定する.

4　PLS,PCRの比較

重みを求める際,PCRでは説明変数側の情報を最も多く取り入れるようにするのに対して, PLSは応答変数の変動を説明しつつ説明変数側からも予測精度の高い情報を取り入れようとする点に違いがある. そのため,PLSのほうが応答変数の情報の吸収が早いとされている.

5　データ解析

統計解析ソフトＲを用いて,PLS,PCR,RR,OLSで解析を行い比較する.PLS,PCRの主成分数はPRESSで決定し,RRのパラメーターはGCVで決定する.

RRのPRESSはGCVが最小となるリッジパラメーターを用いてクロスバリデーションを行い求める．解析に使用する関数は,PLS,PCRは自作の関数,RRはlm.ridge(),OLSはlm()である．

5.1 meatspec(パッケージ faraway)

データmeatspecはある食品の脂肪分fatと100個のスペクトルデータからなっていてサンプル数215のうち,172個をトレーニングデータ,残りの43個をテストデータとして用いる.トレーニングデータのVIF値は最小で321404415となる．PLS,PCRは50主成分まで求め，RRのλは0から5e-8まで1e-10間隔で関数を動かした．表1の残差平方和はテストデータへ当てはめたときのものである．

表 1: ４手法のPRESSと残差平方和

	PRESS	残差平方和
14 components PLS	0.03844	2.45888
20 components PCR	0.03986	2.49328
RR(lambda=1.81e-8)	0.05184	4.28110
OLS	0.08258	4.13588

表 2: PLS,PCRのvariance explained

PLS variance explained(%)							
PCs	1	2	3	4	5	$\cdots$	14
X	98.5	99.2	99.8	99.9	100	$\cdots$	100.0
y	22.6	62.9	82.8	89.7	94.1	$\cdots$	97.62
PCR variance explained(%)							
PCs	1	2	3	4	5	$\cdots$	20
X	98.5	99.5	99.8	99.9	100	$\cdots$	100.0
y	22.3	26.1	65.3	88.9	93.5	$\cdots$	97.60

データmeatspecに対しては,PLSがPRESSと残差平方和を最小にする結果となった．PCRもPRESS,残差平方和をPLSほど小さくできていないが,かなり小さい値となっている．表2から,PLSの方がPCRに比べ応答変数の情報を早く取り入れているのがわかる．とくに,2,3,4主成分数のあたりを見ると違いがよくわかる．

5.2 gasoline(パッケージ pls)

データgasolineは,ガソリンの中に含まれるオクタンの成分量と,401種類の波長で計測されたスペクトルのデータである.オクタンの成分量を応答変数,401のスペクトルを説明変数として回帰分析する． サンプル数は60である.PLS,PCRは30主成分まで求め,RRはリッジパラメーターを8から10まで0.01間隔で動かして最適な値を推定した．

表 3: ４手法のPRESS

	PRESS
5 components PLS	0.01956316
6 components PCR	0.01829873
RR(lambda=9.00)	0.01942261
OLS	NA

データgasolineでは,説明変数の数(401)よりサンプル数(60)が少ないため,OLSでは係数の推定をすることができなかった.PLS,PCR,RRでは係数を推定することができ,この３手法が多重共線性の問題があるようなデータに対しては有効であると言える.PRESSが最小となるのはPCRという結果となった．

5.3 longley(パッケージ MASS)

データlongleyは,多重共線性のあるデータセットとしてよく知られている.Employedを応答変数とし，他の６変数を説明変数として回帰分析する．サンプル数は16である.RRでは,リッジパラメーターを0から1e-2まで1e-4間隔で動かして最適な値を推定する．

表 4: ４手法のPRESS

	PRESS
5 components PLS	0.01404761
6 components PCR	0.01462883
RR(lambda=0.0028)	0.01322694
OLS	0.01462883

データlongleyでは,RRがPRESSを最小とする結果となった.PLSもOLSよりもPRESSを小さくすることができたが,PCRではOLSよりも良いモデルの構築ができなかった．

6 おわりに

PLSの基本的な理論を理解するために,いくつかのアルゴリズムについて研究し, 自作の関数を作成することで理解を深めた．

３つのデータセットを用いて多重共線性のあるデータに対して用いられる手法PCR,RRや通常の重回帰であるOLSとの比較をすることで, PLSの有効性も確認することができた.OLSでは係数が推定できない場面,また多重共線性による通常の重回帰では係数が不安定である場面でPLSが有効であると言える.また,スペクトルデータであるデータmeatspecでは, PLSがPRESSを最小にしテストデータへの当てはめに関しても残差平方和を最小にする結果となり, PLSがよく用いられている計量化学の分野でのスペクトルデータに対しては十分な性能を発揮するだろう．

しかし,PRESSを最小にするという点において,PCR,RRの方が有効である場合もあり,多重共線性のあるデータに対してはこれらのいくつかの手法での解析を行い最も精度良くモデルを構築できているものを採用するのがいいのではないだろうか．

参考文献

[1] Frank,I. and J.Friedman：A statistical view of some chemometrics regression tools, Technometrics, Vol. 35 No. 2, pp. 134–135 (1993).

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