データ解析における総合的変数選択手法と インタラクティブシステムの開発*

森　裕一
岡山理科大学総合情報学部
<mori@soci.ous.ac.jp>

飯塚誠也
岡山大学法学部
<masa@law.okayama-u.ac.jp>

1. はじめに

外的変量のない多変量手法のうち，対象を主成分分析に絞り，その変数選択手法の研究について報告する。そこでは，次の課題が主となっている。

(1) これまでの主成分分析における変数選択手法の整理
(2) 主成分分析における新しい変数選択規準の検討
(3) 主成分分析における変数選択手法の評価と選択変数数の特定への試み
(4) 主成分分析における変数選択プログラムの開発

以下では，先行研究とわれわれの考案する新しい選択規準を含めた各種の選択手法の整理を行い，主成分分析での変数選択の特徴をまとめ，次に，選択手法の評価についての試みを示す。最後に，開発した変数選択プログラムについて紹介する。

2. 主成分分析における変数選択の特徴

調査や検査において，予備調査では重要な次元をもれなく拾い上げるために多数の項目（変数）を設定して調査するが，本調査の際には実施上の観点から重要な次元は見逃さず，しかもできるだけ少ない項目で調査したいという場合がある。また，調査票が大人や健常者用に完成しているが，それを子どもや健常者でない者を対象に実施するとき，質問や検査項目を減らし，しかも完成している調査票と同様の総合指標を得たい場合がある。このような場合，これまでは項目間の相関分析やクラスター分析などにより主観的な変数の精選を行ってきたが，客観的な規準に基づく選択手法の開発が望まれていた。この場合，外的変量があるデータを扱う重回帰分析や判別分析における変数選択は直接的には利用できない。そこで，外的変量を特定しない変数選択手法の開発の必要性が出てくる。

これまで，主成分分析では，Jolliffe（1972,1973），Robert and Escoufier（1976），McCabe（1984），Bonifas et al.（1984），Krzanowski（1987），Falguerolles and Jmel（1993），森 他（1994），Tanaka and Mori（1997）などの研究が顕著である。これらは，それぞれ独自の選択規準をもっており，大きく分けて次の 3 つに分類できる。(i) 選択された変数群の分散共分散（変動）に注目するもの，(ii) 選択された変数群と元の変数群の主成分得点の空間上の布値の近さを利用するもの，(iii) その他である。未検討であるが可能性のある規準も含めると，選択目的によって多くのバリエーションが存在することになる（表 1）。

このように，主成分分析における変数選択では，

・選択の規準がいくつも存在する（選択の観点が1つとは限らない）
・多くの場合，それぞれの選択規準による選択結果（選択される変数群）が異なる（表 7 参照）

ということが，その特徴となる。外的変量のある多変量手法の変数選択と異なる点である。このような変数選択を，実際の場面で応用する場合には，

・選択を行う目的がはっきりしていれば，その選択規準で選択を行う
・検討が必要な場合はいくつかの手法を試してみて，その結果を比較する

ことになる。

* 本研究の一部は，平成 10-11 年度科学研究費補助金（基盤研究(C)(2)）研究「データ解析における総合的変数選択手法の研究とインタラクティブシステムの開発」(課題番号 10680321)によるものである。

表 1: 主成分分析における変数選択規準とその分類

手 法	規 準 (i)変 動	(ii)空間上の布値	(iii)その他
Jolliffe's B2	-	-	負荷量(小→大)
Jolliffe's B4	-	-	負荷量(大→小)
McCabe	偏分散共分散, 正準相関	-	-
Falguerolles and Jmel	ガウシアンモデル	-	-
Krzanowski	-	プロクラステス回転	-
Robert and Escoufier	-	*RV* 係数	-
Bonifas	-	*RV* 係数	-
Tanaka and Mori	寄与率	*RV* 係数	-
変数の影響分析の利用	-	-	変数の影響分析
予測残差の利用	-	-	*PRESS*
重回帰分析	重相関係数	-	-
クラスター分析	-	-	クラスター (主観的)

3. 主成分分析における変数選択の手法

各選択手法を説明するにあたり, 次の表記を用いる。Y を n 個の個体と p 個の変数をもつデータ行列とする。Y は量的データであるが, 元のデータが質的データの場合はそれを数量化したものとする。この Y を q 個の変数をもつ $n\times q$ 部分行列 Y_1 と残りの $p-q$ 個の変数をもつ $n\times(p-q)$部分行列 Y_2 に分割し, $Y=(Y_1, Y_2)$ と表しておく ($1<q<p$)。これに対応して, $Y=(Y_1, Y_2)$ の分散共分散行列を $S=\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$, $S_1 = (S_{11}, S_{12})$, とする。

また, 評価に用いた共通したデータは, 羽根アリデータ(Alate data: Jeffers, 1967)である(表 2)。

表 2: Alate データ:19 変数 40 個体(Jeffers, 1967)

V1 body length	V2 body width	V3 fore-wing length
V4 hind-wing length	V5 number of spiracles	V6 length of antennal segment I
V7 length of antennal segment II	V8 length of antennal segment III	V9 length of antennal segment IV
V10 length of antennal segment V	V11 number of antennal spines	V12 leg length, tarsus III
V13 leg length, tibia III	V14 leg length, femur III	V15 rostrum
V16 ovipositor	V17 number of ovipositor spines	V18 anal fold
V19 number of hond-wing hooks		

3.1 変数選択の手順

計算コストを考慮し, 通常, Bacward などの簡便法がとられるが, 選択の目的に対応できるように, また精度の検討を行うために, 次の 4 つの選択手順を採用する(フローの詳細は Mori(1997), 森, 垂水, 田中(1998)参照)。ここで, Y_1 の変数の数が q のときの規準値(3.2 節以降の C_i, M^2, P, RV など)を $V(q)$と記す。規準によって, $V(q)$の最大値をとる変数群を選ぶ場合と最小値をとる変数群を選ぶ場合があるが, 以下の手順の説明では, 最大値の場合をあげる。最小値の場合は, 該当部分を読み替えて実行すればよい。

変数減少法 (Backward)

Step A (初期段階):Y_1 を構成する q 変数を決め (通常は $q:=p$), 固有値問題(1)を解き, 主成分数 r を決める。必要なら Y_1 のうち核になる (削除しない) 変数を q より少ない数の範囲で決める。

Step B (変数選択段階):変数の数が q であるとき, この q 個の変数の 1 つを削除して得られる q 個の $V(q-1)$ のうち最大値を与える変数の組合せを $q-1$ 変数の最良の変数群とする。$q:=q-1$ として同様の変数減少手順を繰り返し, 事前に定めた変数の数または規準値を超えたら終了する。

変数減増法 (Backward-forward stepwise)

Step A (初期段階):変数減少法の Step A と同じ。

Step B (変数選択段階):変数の数が q であるとし, $V(q)$を記憶しておく。Backward により 1 つ変数を削除し, $q-1$ 個の変数を得る。このとき, 今削除した変数以外でそれ以前に削除されていた Y_2 の中の $p-q$

－1 個の変数を 1 つずつ現在の Y_1 の q－1 変数に付け加えて，それぞれの規準値 $V'(q)$のうちの最大の $V'_{max}(q)$を見つける（Forward の実行）。ここで先の $V(q)$と比較して，$V(q) \geq V'_{max}(q)$ならば Backward を続行し，$V(q) < V'_{max}(q)$ならば，$V'_{max}(q) >$を与える変数を実際に Y_1 に追加し，続いて残りの $p-q-2$ 個の変数に対して同様の Forward を施す。これを繰り返し，$V(q') \geq V'_{max}(q')$ になったら，そこからあらためて Backward に移る。

変数増加法（Forward）

Step A（初期段階）: 変数減少法の Step A と同様に主成分数 r を決める。この後，Forward を始める核となる変数群 Y_1 を定めるが，特定の変数群がない場合は，$q:=r$ として，すべての q 変数の組合せの中で最大の $V(q)$を与える q 変数を Y_1 とする。

Step B（変数選択段階）: 変数の数が q であるとする。Y_2 に属する $p-q$ 個の変数の 1 つを Y_1 につけ加えて得られる $p-q$ 個の $V(q+1)$ のうち最大値を与える変数の組合せを $q+1$ 変数の最良の変数群とする。$q:=q+1$ として同様の変数増加手順を繰り返し，事前に定めた変数の数または規準値を超えたら終了する。

変数増減法（Forward-backward stepwise）

Step A（初期段階）: 変数増加法の Step A と同じ。

Step B（変数選択段階）: 変数減増法の逆。

いくつかの規準について，各手順の効率を比較した結果，4 つの手順はすべての組合せを調べる変数選択手順と比較して顕著な差は見られないことから計算コストの面で有効であること，4 手順の中では Backward 系より Forward 系の方が，また単純選択より Stepwise 手順の方がよい結果が得られることが明らかになった（森 他，1998）。

3.2 先行研究の変数選択

Jolliffe（1972,1973），Robert and Escoufier（1976），McCabe（1984），Bonifas et al.（1984），Krzanowski（1987）について，概要を述べる。

Jolliffe の手法は，固有ベクトルの係数に注目し，主成分に関する寄与の大きい変数を順に $p-q$ 個削除していく方法である。手法名 B2 は，固有値を小さい方から大きい方に見ていって，各固有値に対する固有ベクトルの中で最も係数の大きい（その主成分に寄与の最も大きい）変数を順に削除するものである。手法名 B4 は，逆に固有値の大きい方から小さい方に見ていって，固有ベクトルの係数の大きい変数を順に q 個採択するものである。

Robert and Escoufier（1976）は，n 個の個体をもつ 2 つの行列 Y と Z 間の空間上の布値の近さを測る指標として，次の RV 係数

$$RV(Y,Z) = tr(\tilde{Y}\tilde{Y}'\tilde{Z}\tilde{Z}') / \{tr(\tilde{Y}\tilde{Y}')^2 \cdot tr(\tilde{Z}\tilde{Z}')^2\}^{1/2}$$

を提唱している。したがって，この Y に全 p 変数を用いたときの主成分得点行列，Z に選択された q 変数を用いたとき主成分得点行列などをあてはめると，その RV 係数が計算でき，基のデータと最も布値の近い q 変数群を見つけることが可能となる（ただし，Robert and Escoufier は具体的な変数選択は行っていない）。

McCabe（1984）は p 変数の中から主変数（Principal Variables）を見つける規準として，次の C_1～C_4 の 4 つを提唱している。

$$C_1 = \min \det(S_{22\cdot 1}), \quad C_2 = \min tr(S_{22\cdot 1}), \quad C_3 = \min \|S_{22\cdot 1}\|, \quad C_4 = \max cancor(Y_1, Y_2)$$

$S_{22\cdot 1}$ は Y_1 によって説明される成分を除いたときの Y_2 の偏分散共分散行列，$concor(Y_1, Y_2)$は Y_1 と Y_2 の正準相関係数である。

Krzanowski（1987）は，2 つの行列の近さを回転を合わせて測るプロクラステス分析の指標

$$M^2 = tr(YY' + ZZ' - 2D)$$

を利用して，Y に全 p 変数を用いたときの主成分得点行列，Z に選択された q 変数を用いたとき主成分得点行列などをあてはめ，M_2 を最大化する q 変数を見つけている。

3.3 拡張主成分分析の規準を利用した変数選択

拡張主成分分析（Modified PCA, 以下 M.PCA と略す）は, Y_1 による r 個の線形結合 $Z=Y_1A$ が元の p 個の変数を最もよく代表するように $A = (\boldsymbol{a}_1, ..., \boldsymbol{a}_r)$ を推定しようというものである（$1<r<q$）。そのために次の 2 つの規準を用いる（Tanaka and Mori, 1997; 森, 1998）。

［規準 1］線形結合 $\boldsymbol{z}$ を用いて $\boldsymbol{y}$ の予測効率を最大にする。

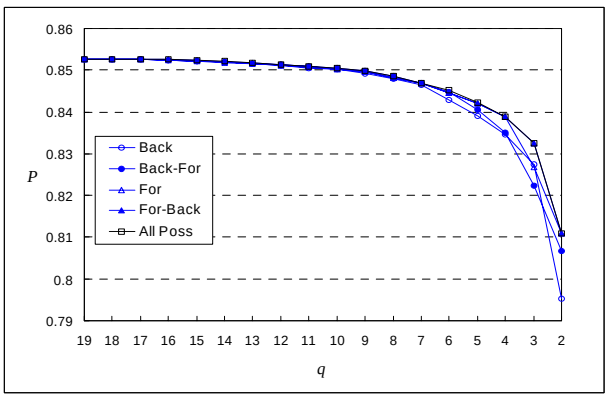
［規準 2］Y と Z の RV 係数 $RV(Y,Z) = tr(\tilde{Y}\tilde{Y}'\tilde{Z}\tilde{Z}') \big/ \{tr(\tilde{Y}\tilde{Y}')^2 \cdot tr(\tilde{Z}\tilde{Z}')^2\}^{1/2}$ を最大にする（$\tilde{Y}, \tilde{Z}$ は Y と Z を中心化した行列）。

$Y = (Y_1, Y_2)$より得られる一般化固有値問題

$$[(S_{11}^2 + S_{12}S_{21}) - \lambda_j S_{11})]\boldsymbol{a}_j = 0 \qquad (1)$$

とその q 個の固有値を大きい順に $\lambda_1, \lambda_2, \ldots, \lambda_q$，対応する固有ベクトルを $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q$ とすると，［規準 1］は，Rao（1964）に従い，一般化固有値問題（1）より得られる r 個の主成分の和によって説明される寄与率 $P = \sum_{j=1}^{r} \lambda_j / tr(S)$，［規準 2］は，Robert and Escoufier（1976）より，RV 係数 $RV = \left\{ \sum_{j=1}^{r} \lambda_j^2 \big/ tr(S^2) \right\}^{1/2}$ が最大化の規準値となる。

M.PCA の規準を利用した変数選択とは，q 個の変数をもつ変数の組み合わせのうち，上記の寄与率 P あるいは RV 係数を最大にする q 変数を見つけていくものである。



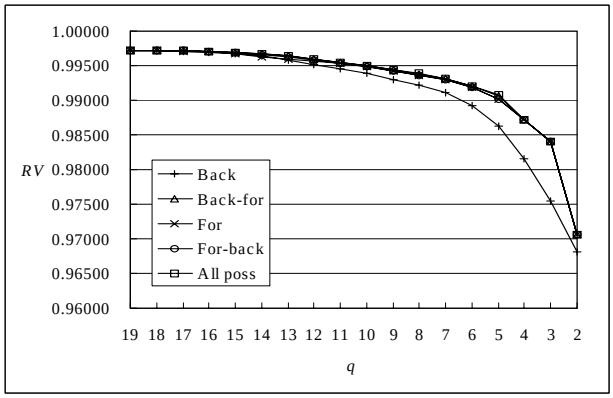


図 1: 規準値 寄与率 P の変化（Alate data, r=2）　　**図 2**: 規準値 RV 係数の変化（Alate data, r=2）

（選択結果については，図 7，図 8 を参照のこと）

3.3 変数の影響分析を利用した変数選択

変数選択の規準として，主成分分析の解析結果あるいは推定するパラメータへの各変数の影響を調べ，それらへの影響が最も小さい変数を削除するという考えを採用する。パラメータとしては，固有値，固有ベクトル，主成分得点などがあり，M.PCA では，さらに規準値である寄与率 P や RV 係数への影響を考えることができる。通常の主成分分析は M.PCA の特殊な場合（以下の $D=I$ とした場合）として扱えるので，変数の影響分析の定式化を M.PCA の場合について述べる（Tanaka and Mori, 1997; 森 他, 2000; Mori et al, 2000）。

変数の影響を考えるために，特定の変数のウエイトを 1 から $1-\varepsilon$ に変化させて結果を評価する。Y_1 の j 番目の変数のウエイトを変化させる場合，J_{jj} を j 番目の要素が 1 で他が 0 である $q \times q$ の対角行列とすると，分散共分散行列 S は，

$$S_{11} \longrightarrow S_{11} - \varepsilon(J_{jj}S_{11} + S_{11}J_{jj}) + O(\varepsilon^2),$$

$$S_{12} \longrightarrow S_{12} - \varepsilon J_{jj}S_{12},$$

$$S_{21} \longrightarrow S_{21} - \varepsilon S_{21}J_{jj}$$

のように変化する($j=1,\ldots,q$)。一般固有値問題(1)を$(C-\lambda D)\boldsymbol{a}=0$と表せば，$C$と$D$は上記変化にともなって，$C+\varepsilon C^{(1)}+O(\varepsilon^2)$，$D+\varepsilon D^{(1)}+O(\varepsilon^2)$へ変化する。ここで，

$$C^{(1)}=-J_{jj}C-CJ_{jj}-2S_{11}J_{jj}S_{11},$$

$$D^{(1)}=-J_{jj}S_{11}-S_{11}J_{jj}$$

である。これらの $C^{(1)}$, $D^{(1)}$を用いると，固有値λ，寄与率 P, RV 係数，固有ベクトルへの影響は次の式で評価される。

(a')固有値への影響　　$\lambda_j^{(1)}=\boldsymbol{a}_j^{'}(C^{(1)}-\lambda_j D^{(1)})\boldsymbol{a}_j$

(b')固有ベクトルへの影響　　$\boldsymbol{a}_j^{(1)}=\sum_{k\neq j}(\lambda_j-\lambda_k)^{-1}\{\boldsymbol{a}_j^{'}(C^{(1)}-\lambda_j D^{(1)})\boldsymbol{a}_k\}\boldsymbol{a}_k-(1/2)(\boldsymbol{a}_j^{'}D^{(1)}\boldsymbol{a}_j)\boldsymbol{a}_j$

(c') 寄与率 P, RV係数への影響

$$P^{(1)}=\left[\sum_{j=1}^{r}\lambda_j\big/tr(S)\right]^{(1)}=\sum_{j=1}^{r}\lambda_j^{(1)}\big/tr(S)-\sum_{j=1}^{r}\lambda_j tr(S^{(1)})\big/(tr(S))^2,$$

$$RV^{(1)}=\left[\left\{\sum_{j=1}^{r}\lambda_j^2\big/tr(S^2)\right\}^{1/2}\right]^{(1)}=\left\{\sum_{j=1}^{r}\lambda_j^2\big/tr(S^2)\right\}^{-1/2}\left\{\sum_{j=1}^{r}\lambda_j\lambda_j^{(1)}\big/tr(S^2)-\sum_{j=1}^{r}\lambda_j^2 tr(SS^{(1)})\big/(tr(S^2))^2\right\}$$

ただし，$S^{(1)}=(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'-S$

このように，肩に(1)をつけて表したものがそのパラメータの影響関数で，この値が大きいほど解析結果に与える影響が大きく，値が小さいと影響があまりないことを示す。1 変数づつ削除する選択では，この SとしてY_1の分散共分散行列をあてはめればよい。

なお，1 変数落としたときの結果を影響関数により近似的に求めることが影響分析にであるので，選択手順としては，Backward のみとなる.

表 3 に Alate データでの結果をあげる.

表 3: 変数の影響関数を利用した変数選択結果

(Alate データ，$r=2$，通常の主成分分析の固有ベクトル，M.PCA の固有値，寄与率 P, RV係数，固有ベクトル)

q	固有ベクトル(PCA)		固有値(M.PCA)		固有ベクトル(M.PCA)		寄与率P(M.PCA)		RV係数(M.PCA)	
	影響測度	削除変数	影響測度	削除変数	影響測度	削除変数	影響測度	削除変数	影響測度	削除変数
18	0.512690	V11	0.069403	V15	0.512690	V11	0.000551	V10	0.000007	V9
17	1.077447	V5	0.060887	V14	1.070646	V5	0.001821	V6	0.000070	V17
16	1.189634	V18	0.045764	V13	1.146930	V18	0.002100	V17	0.000190	V18
15	1.185250	V19	0.042840	V12	1.182362	V19	0.002544	V1	0.000036	V19
14	1.365802	V16	0.042264	V1	1.348856	V16	0.003145	V8	0.000195	V5
13	1.519904	V8	0.065177	V2	1.509261	V17	0.003551	V15	0.001450	V16
12	1.585744	V9	0.070742	V10	1.521576	V8	0.006080	V9	0.001967	V8
11	1.608467	V17	0.081177	V9	1.576243	V9	0.006316	V18	0.002630	V6
10	1.632888	V6	0.145583	V4	1.612934	V6	0.006598	V5	0.002859	V10
9	1.729252	V10	0.234940	V6	1.723279	V10	0.007646	V19	0.003159	V11
8	1.743419	V7	0.269682	V3	1.756584	V1	0.011500	V2	0.003704	V1
7	1.780084	V15	0.265698	V8	1.773749	V15	0.012112	V3	0.003875	V15
6	1.821419	V1	0.306011	V11	1.804726	V7	0.013542	V12	0.004364	V7
5	1.887323	V3	0.349296	V16	1.883830	V2	0.014174	V4	0.008672	V2
4	1.898510	V4	0.403199	V19	1.904059	V4	0.015143	V16	0.008899	V3
3	1.909236	V2	0.482343	V17	1.919654	V3	0.018919	V14	0.010510	V4
2	1.961577	V12	0.717170	V5	1.958717	V14	0.020426	V13	0.013315	V12

3.4 予測残差を利用した変数選択

予測あるいはモデル選択の意味で，主成分分析の各選択ステップを評価するために，クロスバリデーションを用いた予測残差（*PRESS*）が定式化できる（Mori et al, 2000b）。ただし，主成分分析の場合，全データを用いて主成分などを予測するので，ある観測値をクロスバリデーションの通常の方法で予測することには問題が生じる。そこで，次のような工夫をする。$\tilde{Y}_{(i)}$ を Y から i 番目の観測値を抜いた $(n-1)\times p$ の中心化されたデータ行列，$A_{(i)}$，$Z_{(i)}$ をそれぞれ $\tilde{Y}_{(i)}=(\tilde{Y}_{(i)1},\tilde{Y}_{(i)2})$ に基づく固有値問題（1）の係数ベクトルと主成分行列とする。このとき，*PRESS* を次のようにして定義する。

$$PRESS_q=\sum_{i=1}^{n}\sum_{j=1}^{p}(\tilde{y}_{ij}-\hat{y}_{ij})^2 \qquad (2)$$

ただし，$\hat{Y}$ の i 番目の行を $\hat{\boldsymbol{y}}_i=\boldsymbol{z}_iB$ とし，$\boldsymbol{z}_i$ は Y と $A_{(i)}$ に基づく主成分行列の i 番目の行，B は $B=(Z_{(i)}^{'}Z_{(i)})^{-1}Z_{(i)}^{'}\tilde{Y}_{(i)}$ とする。

q 変数の中から j 番目の変数を抜いたときに，この予測残差 $PRESS_q$ を求め，$j=1,\ldots,q$ の中で $PRESS_q$ を最も値を小さくする j 番目の変数が選択または削除の対象となる。

表 4 は，Alate データにおける結果で，図 3 はそれをグラフ化したものである.

表 4: *PRESS* による選択結果（Alate データ，$r=2$）

q	backward	back-for	forward	for-back
18	0.17788	0.17788	0.17794	0.17794
17	0.17837	0.17837	0.17820	0.17820
16	0.17888	0.17888	0.17844	0.17844
15	0.17929	0.17929	0.17859	0.17859
14	0.17984	0.17984	0.17905	0.17905
13	0.18029	0.18029	0.18015	0.18015
12	0.18100	0.18100	0.18127	0.18117
11	0.18190	0.18180	0.18209	0.18209
10	0.18337	0.18257	0.18329	0.18329
9	0.18477	0.18418	0.18430	0.18430
8	0.18649	0.18550	0.18593	0.18593
7	0.18925	0.18763	0.18805	0.18800
6	0.19236	0.19190	0.19003	0.19100
5	0.19568	0.19812	0.19477	0.19404
4	0.20425	0.20414	0.19977	0.19788
3	0.21494	0.21129	0.21095	0.20467
2	0.23096	0.23768	0.22747	0.22747

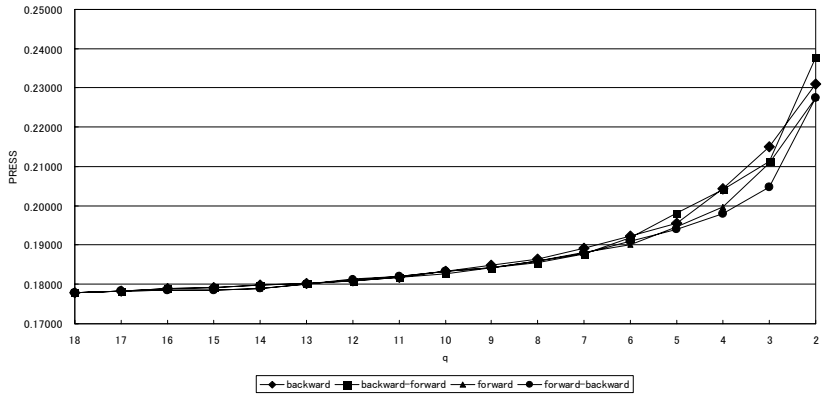


図 3: $PRESS_q$ の値の変化（Alate データ，$r=2$）

4. 手法の評価と選択変数数決定への試み

各選択規準がデータに対してどのような振る舞いをするかについて，2 つのブートストラップタイプのリサンプリングを行う（Mori et al, 1999）。

Bootstrap Type I

Bootstrap サンプリングを B 回行い，B 個の Bootstrap サンプルに対して，変数選択を行う。したがって，1 サンプルごとに，q を p から r まで動かしたときの変数選択が行われる。この結果から，各 q においてできた B 種類の選択変数群とそれぞれの規準値の考察を行い，選択規準の安定性を見る。

Bootstrap Type II

Bootstrap サンプリングを B 回行い，B 個の Bootstrap サンプルに対して，変数選択を行うが，どのサンプルにおいても各 q における選択変数群を，元のデータに対して指定された手法で変数選択したときに得られる変数群に固定して，規準値を計算する。これより，異なるデータが得られたときの規準値の再現性を見る。

ここでは，M.PCA の規準による変数選択を対象として，リサンプリングを行う。Type I では，Alate データに対して，Bootstrap サンプル 200 セットのそれぞれに対して，Forward による変数選択を行ってみると，各 q において得られた寄与率 P の標準誤差は小さく，変数の選ばれ方にかかわらず P の推定値は安定していることがわかった。しかし，各 q で得られる変数の組み合わせは，サンプルごとに異なり，q が小さいほど 200 個の中で飛び抜けて度数の多い変数の組み合わせは見られなくなる（表 5, 6）。用いるデータにより選ばれる変数が大きく変化することがわかる。しかし，規準値は安定しているので，入れ替わっている変数群は，規準値に対して同じような指標を提供することが観察される。一方，Type II では，Alate データ 200 サンプルについて，各 q における寄与率 P を求め，その標準誤差（SE）を求めた（図 4）。P の標準誤差の値およびその変化は大きくなく，Bacward, Forward 等の選択手順間でもその値に大きな差は見られない。事前に指定した変数の並びが指定されているわけであるが，その並びによる規準値の変化は小さく，また選択手順を変えても同等の情報を提供することがわかる。

さらに，選択されるべき変数の数の特定についても試みた（Mori et al, 1999）。これは，クロスバリデーションの手法を用いるもので，(2)式の *PRESS* を求めることになる。すなわち，その値の変化を考察して変数の数を特定しようというものである。図 5 に，選択される変数群を固定して計算した $PRESS_q$ の変化を示す。これと先の図 3（*PRESS* による変数選択の各 q における $PRESS_q$ 値の変化）および図 4（ブートストラップ手法の SE の変化）から，*PRESS* と SE の変化が $q=6$ まではゆるやかであることが見て取れる。これらの結果を利用することは，変数の数を特定するまでには至っていないが，選択されるべき変数数の決定に必要な情報を与えるものといえる。

表 5: $q=17$ のときの選択変数（Y_1）の度数（一部）
（Alate データ，$B=200$，$r=2$，Forward）

																度数	
1	2	3	4	5	6	7	8	9	10	11	12	15	16	18	19	123	62%
1	2	3	4	5	6	7	8	9	10	11	14	15	16	18	19	40	20%
1	2	3	4	5	6	7	8	9	11	12	14	15	16	18	19	3	2%
1	2	3	4	5	6	8	9	10	11	12	14	15	16	18	19	4	2%
1	2	3	5	6	7	8	9	10	11	12	14	15	16	18	19	13	7%
1	2	4	5	6	7	8	9	10	11	12	14	15	16	18	19	6	3%
1	3	4	5	6	7	8	9	10	11	12	14	15	16	18	19	6	3%
2	3	4	5	6	7	8	9	10	11	12	14	15	16	18	19	3	2%

表 6: $q=9$ のときの選択変数（Y_1）の度数（一部）
（Alate データ，$B=200$，$r=2$，Forward）

									度数	
1	4	5	6	10	11	16	18	19	13	7%
3	4	5	10	11	15	16	18	19	5	3%
3	5	6	7	11	15	16	18	19	5	3%
1	5	6	10	11	15	16	18	19	4	2%
2	3	5	7	11	14	16	18	19	4	2%
3	4	5	7	11	15	16	18	19	4	2%
3	4	5	9	10	11	16	18	19	4	2%
3	5	7	11	14	15	16	18	19	4	2%
4	5	6	10	11	15	16	18	19	4	2%
1	3	4	5	7	11	16	18	19	3	2%
2	5	6	8	9	11	16	18	19	3	2%
3	4	5	6	8	11	16	18	19	3	2%
3	4	5	6	10	11	16	18	19	3	2%
3	4	5	6	11	15	16	18	19	3	2%
3	4	5	7	8	11	16	18	19	3	2%
3	5	6	10	11	15	16	18	19	3	2%
3	5	7	10	11	15	16	18	19	3	2%
1	2	5	6	9	11	16	18	19	2	1%
1	4	5	6	8	11	16	18	19	2	1%

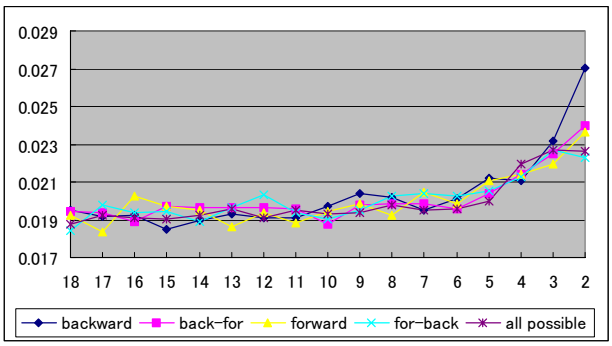


図 4: 寄与率 *P* の標準誤差の変化（Alate data, *r*=2）

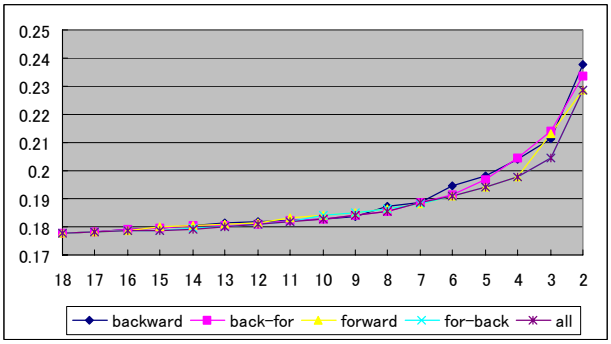


図 5: $PRESS_q$ の変化（Alate data, *r*=2）

5. 主成分分析における変数選択プログラム VASPCA

これまでみてきたような特徴のある主成分分析における変数選択に対しては，柔軟に手法が実行できる環境が必要となる。ソフトウェア“VASPCA（VAriable Selection in Principal Component Analysis）”である（森 他，1997; 飯塚 他，2000; Mori et al., 2000a など）。これには，Windows で動く VASPCA/Win と Web 上で実行が可能な VASPCA/Web の 2 つがある。上にあげた主成分分析における変数選択手法の全てが実行でき，事前に核になる（選択過程で落としたくない）変数や手順を指定できるといったインタラクティブなソフトである。

VASPCA/Web の仕様は以下の通りである。

◆実行可能な選択手法

A) 拡張主成分分析の規準を利用した変数選択
B) 主変数 (Principal variables) の考えを利用した変数選択
C) プロクラステス分析を利用した変数選択
D) *RV* 係数による変数選択
E) 変数の影響分析を利用した変数選択
F) 予測残差を利用した変数選択
G) 主成分負荷量に着目した変数選択

◆実行手順

1. 事前に主成分分析を行い，主成分数 r を決める。
2. データをテキストエディターで作成し保存する。
3. 【ページ 1】データ入力（上の 2 で作成・保存したデータの指定）
4. 【ページ 2】選択のためのパラメータ指定（入力データの確認，主成分数 *r*（手法 G 以外），選択基準，選択手順）
5. 【ページ 3】（選択結果表示）

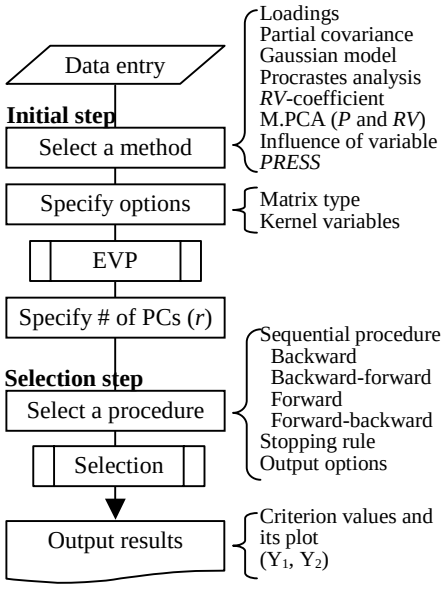


図 6: VASPCA のフロー

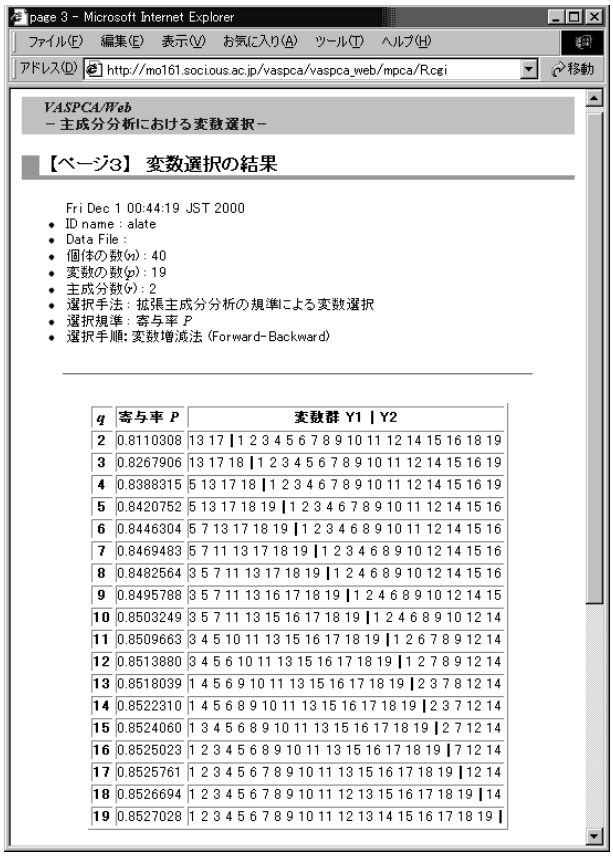
◆VASPCA/Web の仕様

- OS: PC Unix (Linux)
- 記述言語：HTML，CGI（Pearl）
- 統計エンジン：R

◆URL

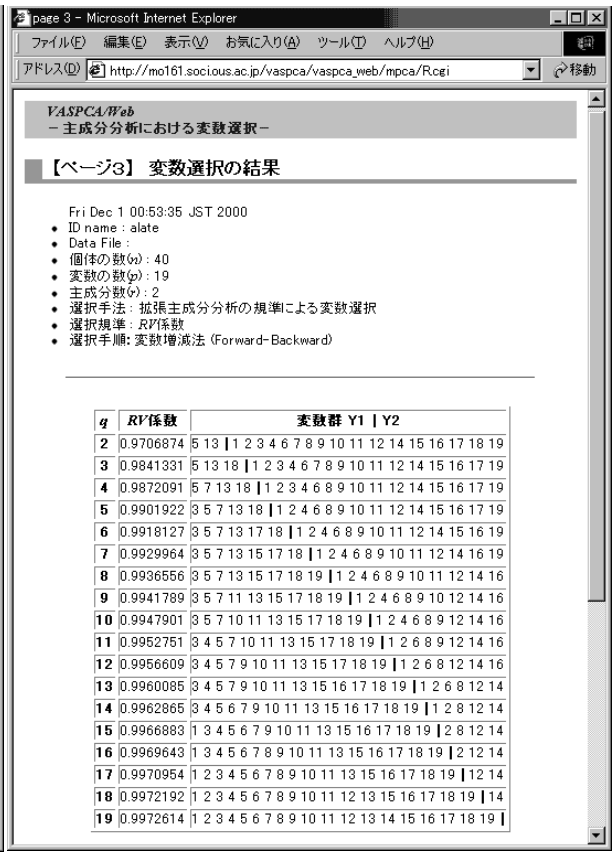
http://face.f7.ems.okayama-u.ac.jp/~masa/vaspca/
http://mo161.soci.ous.ac.jp/vaspca/

この実行結果を図 7 と図 8 に示す。手法が「A）拡張主成分分析の規準を利用した変数選択（M.PCA）」で，選択規準は，図 7 が寄与率 *P*，図 8 が *RV* 係数選択，選択手順はともに Forward-backward である。また，表 7 に，各選択手法で選択された $q=4$ の場合の変数群を示す。



q	寄与率 P	変数群 Y1 | Y2
2	0.8110308	13 17 | 1 2 3 4 5 6 7 8 9 10 11 12 14 15 16 18 19
3	0.8267906	13 17 18 | 1 2 3 4 5 6 7 8 9 10 11 12 14 15 16 19
4	0.8388315	5 13 17 18 | 1 2 3 4 6 7 8 9 10 11 12 14 15 16 19
5	0.8420752	5 13 17 18 19 | 1 2 3 4 6 7 8 9 10 11 12 14 15 16
6	0.8446304	5 7 13 17 18 19 | 1 2 3 4 6 8 9 10 11 12 14 15 16
7	0.8469483	5 7 11 13 17 18 19 | 1 2 3 4 6 8 9 10 12 14 15 16
8	0.8482564	3 5 7 11 13 17 18 19 | 1 2 4 6 8 9 10 12 14 15 16
9	0.8495788	3 5 7 11 13 16 17 18 19 | 1 2 4 6 8 9 10 12 14 15
10	0.8503249	3 5 7 11 13 15 16 17 18 19 | 1 2 4 6 8 9 10 12 14
11	0.8509663	3 4 5 10 11 13 15 16 17 18 19 | 1 2 6 7 8 9 12 14
12	0.8513880	3 4 5 6 10 11 13 15 16 17 18 19 | 1 2 7 8 9 12 14
13	0.8518039	1 4 5 6 9 10 11 13 15 16 17 18 19 | 2 3 7 8 12 14
14	0.8522310	1 4 5 6 8 9 10 11 13 15 16 17 18 19 | 2 3 7 12 14
15	0.8524060	1 3 4 5 6 8 9 10 11 13 15 16 17 18 19 | 2 7 12 14
16	0.8525023	1 2 3 4 5 6 8 9 10 11 13 15 16 17 18 19 | 7 12 14
17	0.8525761	1 2 3 4 5 6 7 8 9 10 11 13 15 16 17 18 19 | 12 14
18	0.8526694	1 2 3 4 5 6 7 8 9 10 11 12 13 15 16 17 18 19 | 14
19	0.8527028	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 |

図 7: VAPCA/Web 選択結果
（Alate data, M.PCA, r=2, 寄与率 P, F-B）



q	RV係数	変数群 Y1 | Y2
2	0.9706874	5 13 | 1 2 3 4 6 7 8 9 10 11 12 14 15 16 17 18 19
3	0.9841331	5 13 18 | 1 2 3 4 6 7 8 9 10 11 12 14 15 16 17 19
4	0.9872091	5 7 13 18 | 1 2 3 4 6 8 9 10 11 12 14 15 16 17 19
5	0.9901922	3 5 7 13 18 | 1 2 4 6 8 9 10 11 12 14 15 16 17 19
6	0.9918127	3 5 7 13 17 18 | 1 2 4 6 8 9 10 11 12 14 15 16 19
7	0.9929964	3 5 7 13 15 17 18 | 1 2 4 6 8 9 10 11 12 14 16 19
8	0.9936556	3 5 7 13 15 17 18 19 | 1 2 4 6 8 9 10 11 12 14 16
9	0.9941789	3 5 7 11 13 15 17 18 19 | 1 2 4 6 8 9 10 12 14 16
10	0.9947901	3 5 7 10 11 13 15 17 18 19 | 1 2 4 6 8 9 12 14 16
11	0.9952751	3 4 5 7 10 11 13 15 17 18 19 | 1 2 6 8 9 12 14 16
12	0.9956609	3 4 5 7 9 10 11 13 15 17 18 19 | 1 2 6 8 12 14 16
13	0.9960085	3 4 5 7 9 10 11 13 15 16 17 18 19 | 1 2 6 8 12 14
14	0.9962865	3 4 5 6 7 9 10 11 13 15 16 17 18 19 | 1 2 8 12 14
15	0.9966883	1 3 4 5 6 7 9 10 11 13 15 16 17 18 19 | 2 8 12 14
16	0.9969643	1 3 4 5 6 7 8 9 10 11 13 15 16 17 18 19 | 2 12 14
17	0.9970954	1 2 3 4 5 6 7 8 9 10 11 13 15 16 17 18 19 | 12 14
18	0.9972192	1 2 3 4 5 6 7 8 9 10 11 12 13 15 16 17 18 19 | 14
19	0.9972614	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 |

図 8: VAPCA/Web 選択結果
（Alate data, M.PCA, r=2, RV係数, F-B）

表 7: 各手法による選択結果（Alate data, 変数の数 q＝4, 主成分数 r＝2）

Method	V1	V2	V3	V4	V5	V7	V8	V9	V11	V12	V13	V14	V16	V17	V18	V19
M.PCA（F-B/P）					X						X			X	X	
M.PCA（AP/P）					X	X					X				X	
M.PCA（F-B/RV）					X	X					X				X	
M.PCA（AP/RV）					X	X					X				X	
SA（B/P）						X			X		X	X				
SA（B/RV）				X						X	X	X				
PRESS (F-B)				X	X						X			X		
Jolliffe B2					X		X		X			X				
Jolliffe B4					X				X		X			X		
McCabe（C2）								X	X					X		X
Krzanowski					X					X		X			X	
Robert & Escoufier			X			X				X	X					
M.Reg					X								X	X		X
Cluster	X	X						X		X						

6. 今後の課題

今後の課題としては，主成分分析における変数選択としては，質的データおよび質的量的混在データの扱いなどの検討，変数の数を決めるにあたっての情報量規準の導入などが考えられる。また，ソフトウェア VASPCA に関しては，その充実と共に明らかになった点を随時に取り込んでいくことになる。さらに，ここで得られたノウハウを利用して，外的変量をもたない多変量手法全体に対する変数選択の整理とソフトウェアへの統合化を行うことが必要と考える。

参考文献

Bonifas, I., Escoufier, Y., Gonzalez, P.L. et Sabatier, R.（1984）: Choix de variables en analyse en composantes principales. *Rev. Statist. Appl.*, **23**, 5-15.

Falguerolles, A. De et Jmel, S.（1993）: Un critere de choix de variables en analyse en composantes principales fonde sur des modeles graphiques gaussiens particuliers. *Rev. Canadienne Statist.*, **21**（3）, 239-256.

Jeffers, J. N. R.（1967）. Two case studies in the application of principal component analysis. *Appl. Statist.*, **16**, 225-236.

Jolliffe, I. T.（1972）. Discarding variables in a principal component analysis. I. Artificial data. *Appl. Statist.*, **21**, 160-173.

Jolliffe, I. T.（1973）. Discarding variables in a principal component analysis. II. Real data. *Appl. Statist.*, **22**, 21-31.

Krzanowski, W. J.（1987）. Selection of variables to preserve multivariate data structure, using principal components. *Appl. Statist.*, **36**, 22--33.

McCabe, G. P.（1984）. Principal variables. *Technometrics*, **26**, 137-44.

Mori,Y.（1997）. Statistical software VASPCA - Variable selection in PCA -. *Bulletin of Okayama University of Science*, **33**（A）, 329-340.

Mori, Y., Iizuka, M. Tarumi, T. and Tanaka, Y.（1999）. Variable selection in "Principal Component Analysis Based on a Subset of Variables", *Bulletin of the International Statistical Institute（52nd Session Contributed Papers Book2）*, 333-334.

Mori, Y., Iizuka, M. Tarumi, T. and Tanaka, Y.（2000a）. Statistical Software "VASPCA" for Variable Selection in Principal Component Analysis, In: COMPSTAT2000 Proceedings in Computational Statistics (Short Coomications), 73-74.

Mori, Y., Iizuka, M. Tarumi, T. and Tanaka, Y.（2000b）. Study of Variable Selection Criteria in Data Analysis, *Proceedings of the Tenth Japan and Korea Joint Conference of Statistics*.

Mori, Y., Tarumi, T. and Tanaka, Y.（1994）. Variable selection with RV-coefficient in principal component analysis. In: *Short Communication in COMPSTAT 1994*（Edited by Dutter, R. and Grossman, W.）, 169-170, Heidelberg: Physica-Verlag.

Rao, C. R.（1964）. The use and interpretation of principal component analysis in applied research. *Sankhya*, **A**, **26**, 329-358.

Robert, P. and Escoufier, Y.（1976）. A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Appl. Statist.*, **25**, 257-65.

Tanaka, Y and Mori, Y.（1997）. Principal component analysis based on a subset of variables: Variable selection and sensitivity analysis. *American Journal of Mathematics and Management Sciences*, **17**（1&2）, 61-89.

飯塚誠也, 森　裕一, 垂水共之 (2000). Development of On-line Program for Statistical Computing. 日本計算機統計学会第 14 回大会論文集, 128-131.

森　裕一（1998）. 変数の一部に基づく主成分分析－RV 係数規準による数値的検討－. 岡山理科大学紀要, 34（A）, 383-396.

森　裕一, 飯塚誠也, 垂水共之, 田中　豊 (2000). 変数の影響分析を利用した変数選択, 日本行動計量学会第 68 回大会抄録.

森　裕一, 垂水共之, 田中　豊（1998）. 変数の一部に基づく主成分分析－変数選択手法の数値的検討－, 計算機統計学, **11**(1), 1-12.