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——

森　　裕　一

岡山理科大学総合情報学部社会情報学科

(1998年10月5日　受理)

1．はじめに

主成分分析や因子分析を用いて，少ない次元でデータの隠された特徴までを測りとれるような指標を作ることを考える。妥当性の高い指標を得ようとすると，調査にはできるだけ多くの項目（変数）を用いたいが，調査の実施上の観点からは，項目数はできるだけ少ない方がよい。このような場合，主成分分析や因子分析における変数選択に関する問題を考えることになる。

これに対して，森，垂水，田中 (1994)[1]，Tanaka and Mori (1997)[2]は，元の変数全体を最もよく代表する総合指標を一部の変数を用いて抽出する変数の一部に基づく主成分分析（他と区別するために拡張主成分分析（Modified PCA，M. PCA）と呼ぶことにする）を提案している。M. PCA では，一部の変数を基にしながらも残りの変数の情報も取り込んだ主成分（総合指標）を抽出するために次の2つの規準を用意している。

1）Rao (1964)[3]の操作変数の主成分分析を利用した規準

2）Robert and Escoufier (1976)[4]の RV 係数による規準

M. PCA に関しては，さらにその主成分の抽出にある個体や変数がどのように影響したかを調べる感度分析（Tanaka and Mori. 1997[2]）や変数が質的データの場合でも扱えるデータの尺度によらない手法の構築（Mori. Tanaka and Tarumi, 1997[5]）を試みてきている。また，主成分分析における変数選択手法の1つとして変数選択プログラム VASPCA（森，1997[6]）に組み込んだり，Backward，Forward，Stepwise などの選択手順を数値的に比較検討(森，垂水，田中，1998[7])している。ただし，これらの研究はすべて規準1）に基づいたものであり，規準2）の方は，1つの事例を Tanaka and Mori (1997)[2]で扱っているだけで，詳細な検討はなされていない。

そこで，本稿では RV 係数のアイデアを利用する規準2）による主成分の抽出について，Jolliffe (1972[8]，1973[9]，1986[10])，Robert and Escoufier (1976)[4]，McCabe (1984)[11]，Krzanowski (1987 a[12]，1987 b[13]) などの主成分分析における変数選択の先行研究や規準1）と比較しながら考察を行うことにする。以下，2節で M. PCA の2つの規準と変数選

択手順を概説し，3節で数値例を示し，4節で考察を行う。

2．変数の一部に基づく主成分分析（M. PCA）

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。この Y の一部の変数 Y_1 を用いて元の全変数 Y をできるだけよく予測しよう，すなわち，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 $A=(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r)$ を推定しようというのが M. PCA である $(1<r<q)$。このような主成分の抽出に，Rao (1964)[3]の操作変数の主成分分析と Robert and Escoufier (1976)[4]の RV 係数のアイデアを利用する。

2．1 Rao の操作変数の主成分分析による定式化

Rao(1964)[3]の操作変数の主成分分析のアイデアに従い，次の規準により，$Z=Y_1A$ が元の p 個の変数を最もよく代表するような $A=(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r)$ を推定する。

（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。

最良の線形予測が得られたときの残差分散共分散行列は，$S_{res}=S-S_1'A(A'S_{11}A)^{-1}A'S_1=S-S_{Reg}$ と表される。したがって，規準1の問題はこの S_{Reg} を最大化する問題に帰着される。最大化の方法はいくつか考えられるが，Rao(1964)[3]に従い，ここでは $tr(S_{Reg})$ の最大化を用いる。すると，次の一般化固有値問題

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

が得られる。この(1)式の q 個の固有値を大きい順に $\lambda_1, \lambda_2, \ldots, \lambda_q$ とし，対応する固有ベクトルを $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q$ とすれば，問題の解，すなわち A は，$A=(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r)$ であり，基準値である $tr(S_{Reg})$ の最大値は，

$$\max tr(S_{Reg})=\sum_{i=1}^{r}\lambda_i \tag{2}$$

で求められる。なお，(2)式の値を全分散で割った

$$P=\sum_{i=1}^{r}\lambda_i/tr(S) \tag{3}$$

は，r 個の主成分によって説明される元の分散の割合（寄与率）を表すことになり，解釈に便利なので，この P を最大化の規準値，すなわち最適な変数を選択する規準値として用いる。なお，この規準は，$\boldsymbol{y}$ が標準化されている場合，$\boldsymbol{z}$ によって $\boldsymbol{y}$ の各要素を予測する

ときの重相関係数を最大化することにあたる。

2．2 Robert and Escoufier の RV 係数による定式化

Robert and Escoufier (1976)[4]に従って次の規準を用いることにより，$Z = Y_1A$ が元の p 個の変数を最もよく代表するように $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r)$ の推定が行える。すなわち，

Y と Z の configuration が最も近くなるように

$$\| \tilde{Y}\tilde{Y}' / [tr(\tilde{Y}\tilde{Y}')^2]^{1/2} - \tilde{Z}\tilde{Z}' / [tr(\tilde{Z}\tilde{Z}')^2]^{1/2} \|$$

を最小化する $Z = Y_1A$ を推定する。ただし，$\tilde{Y}$，$\tilde{Z}$ は Y，Z を中心化した行列，$\|\cdot\|$ はユークリッドノルムを表す。

この規準は次の RV 係数を最大化する規準と同値である。

（規準2）Y と Z の RV 係数

$$RV(Y, Z) = tr(\tilde{Y}\tilde{Y}'\tilde{Z}\tilde{Z}') / \{tr(\tilde{Y}\tilde{Y}')^2 \cdot tr(\tilde{Z}\tilde{Z}')^2\}^{1/2} \tag{4}$$

を最大化する。

(4)式は，$RV(Y, Y_1A) = tr(S_1'AA'S_1) / \{trS^2 \cdot tr(A'S_{11}A)\}^{1/2}$ となるので，（規準1）と同じ固有値問題(1)を解くことになる。したがって，$\boldsymbol{a}_i$ を(1)式の i 番目に大きい固有値 λ_i に対応する固有ベクトルとすれば，正規化 $\boldsymbol{a}_i'S_{11}\boldsymbol{a}_i = \delta_{ij}\lambda_i$（$\delta_{ij}$ はクロネッカーのデルタ）の下で得られる $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r)$ が問題の解 A となる（Robert and Escoufier, 1976[4]）。このときの規準値 $RV(Y, Z)$ の最大値は，

$$RV = \left\{ \sum_{i=1}^{r} \lambda_i^2 / tr(S^2) \right\}^{1/2} \tag{5}$$

で求められる。

2．3 変数選択の手順

M. PCA の規準による変数選択手順としては，q 個の変数のすべての組合せのうちで，規準値 P または RV の値を最も大きくする変数の組を見つけることが最良である。しかし，計算コストの関係から何らかの簡便な方法をとる必要があり，M. PCA では，変数減少法（Backward），変数増加法（Forward），および Backward と Forward を1変数に関して交互に繰り返していく変数減増法（Backward-forward stepwise）と変数増減法（Forward-backward stepwise）の計4つの手順を採用している。各手順の概略は次の通りである（フローの詳細は Mori (1997)[6]，森，垂水，田中 (1998)[7]を見よ）。なお，Y_1 の変数の数が q のときの基準値 p または RV の値を $V(q)$ と記す。

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（削除対象としない）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）：今，変数の数が q であるとする。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基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)：今，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，それぞれの基準値 $V'(q)$ 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 を続行する。$V(q) < V'_{\max}(q)$ 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 変数の組合わせの中で最大の $V(q)$ を与える q 変数を Y_1 とする。

Step B（変数選択段階）：今，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基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。

以下の説明のため，4手順を選択の詳しさで分けて，Backward と Forward を「単純選択系」，Backward-forward と Forward-backward を「Stepwise 系」と記し，また，変数の増減の向きで分けて，Backward と Backward-forward を「Backward 系」，Forward と Forward-backward を「Forward 系」と記すことにする。

規準 P に関しては，すべての組合せの選択結果と比較した結果，4つの手順の選択結果はすべての組合せによる結果と大差がないこと，Backward 系より Forward 系の方がより高い P の値を得ることがわかっている（森，垂水，田中，1998[7])。

3．数　値　例

M. PCA の（規準 2）による総合指標の抽出を行う。適用するデータは，主成分分析における変数選択問題の先行研究にも用いられ，一連の M. PCA の評価に用いてきた「羽根アリデータ（Jeffers, 1967[14]）」である。羽根アリデータは40個体×19変数で，このデータの相関行列に M. PCA の各手順を適用する。先行研究と合わせ主成分数 r は 2 とする（固有値は，$\lambda_1=13.838(72.83\%)>\lambda_2=2.363(85.27\%)>\lambda_3=0.748(89.21\%)>\lambda_4=0.505(91.86\%)>\cdots\cdots$）。

表Ⅰから表Ⅳに各手順の選択過程を示す。Forward 系では Y_1 として $q=r=2$ としたときの最良の組合せ {V5, V13} を用いて選択を開始した。表が示す通り，Stepwise 系ではいくつかの変数が入れ替わりながら選択が行われている。

表Ⅴは各手順の選択過程の要約である。各 q において各手順が選択した最良の変数群と RV の値である（Forward 系は Backward 系に合わせて得られた結果を逆順に示してある）。また，参考として，すべての組合せを調べた結果（All possible）を併記し，これと各手順との RV の差を表右に示した。図 1 は，表Ⅴの RV の値をグラフ化したものである。これらより，Stepwise 系では，RV の値のより高い変数群を選択できること，「羽根アリデータ」では Backward 以外の 3 手順に大きな差はないこと，また，各手順とも All possible の結果と比較して大きな差が見られないことがわかる。さらに，RV の変化が示すように，このデータでは変数の数を 7 ～ 6 に減らしても19変数の RV の値と大きな差はない。All possibleの場合，$q=7$ の RV（0.99311）と $q=p=19$ の RV（0.99726）との差は0.00415で，元の変数全体を再現する主成分を構成するとき19変数のうち12変数を減らしても RV 係数をわずかしか下げない。すなわち，選択された変数群から求められる拡張主成分（固有値問題(1)を解いて得られる主成分）の相対的な布置（configuration）は元の

表Ⅰ　変数の選択過程（Backward）

¥は Y_1 と Y_2 の境界，＋はその変数の添加，－はその変数の削除を示す（表Ⅰ～Ⅳ共通）

q	RV	variable	Y_1 ¥ Y_2																	
19	0.997256		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19
18	0.997233	−V13	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V14	V15	V16	V17	V18	V19¥V13	
17	0.997072	−V7	V1	V2	V3	V4	V5	V6	V8	V9	V10	V11	V12	V14	V15	V16	V17	V18	V19¥V7	V13	
16	0.996923	−V12	V1	V2	V3	V4	V5	V6	V8	V9	V10	V11	V14	V15	V16	V17	V18	V19¥V12	V7	V13	
15	0.996698	−V3	V1	V2	V4	V5	V6	V8	V9	V10	V11	V14	V15	V16	V17	V18	V19¥V3	V12	V7	V13	
14	0.996335	−V15	V1	V2	V4	V5	V6	V8	V9	V10	V11	V14	V16	V17	V18	V19¥V15	V3	V12	V7	V13	
13	0.995829	−V18	V1	V2	V4	V5	V6	V8	V9	V10	V11	V14	V16	V17	V19¥V18	V15	V3	V12	V7	V13	
12	0.995211	−V1	V2	V4	V5	V6	V8	V9	V10	V11	V14	V16	V17	V19¥V1	V18	V15	V3	V12	V7	V13	
11	0.994522	−V4	V2	V5	V6	V8	V9	V10	V11	V14	V16	V17	V19¥V4	V1	V18	V15	V3	V12	V7	V13	
10	0.993876	−V16	V2	V5	V6	V8	V9	V10	V11	V14	V17	V19¥V16	V4	V1	V18	V15	V3	V12	V7	V13	
9	0.993001	−V9	V2	V5	V6	V8	V10	V11	V14	V17	V19¥V9	V16	V4	V1	V18	V15	V3	V12	V7	V13	
8	0.992185	−V8	V2	V5	V6	V10	V11	V14	V17	V19¥V8	V9	V16	V4	V1	V18	V15	V3	V12	V7	V13	
7	0.991071	−V2	V5	V6	V10	V11	V14	V17	V19¥V2	V8	V9	V16	V4	V1	V18	V15	V3	V12	V7	V13	
6	0.989252	−V10	V5	V6	V11	V14	V17	V19¥V10	V2	V8	V9	V16	V4	V1	V18	V15	V3	V12	V7	V13	
5	0.986223	−V17	V5	V6	V11	V14	V19¥V17	V10	V2	V8	V9	V16	V4	V1	V18	V15	V3	V12	V7	V13	
4	0.981634	−V11	V5	V6	V14	V19¥V11	V17	V10	V2	V8	V9	V16	V4	V1	V18	V15	V3	V12	V7	V13	
3	0.975543	−V6	V5	V14	V19¥V6	V11	V17	V10	V2	V8	V9	V16	V4	V1	V18	V15	V3	V12	V7	V13	
2	0.968131	−V19	V5	V14¥V19	V6	V11	V17	V10	V2	V8	V9	V16	V4	V1	V18	V15	V3	V12	V7	V13	

表II 変数の選択過程 (Backward-forward)

q	RV	variable	Y_1 ¥ Y_2																	
19	0.997256		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19
18	0.997233	−V13	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V14	V15	V16	V17	V18	V19¥V13	
17	0.997072	−V7	V1	V2	V3	V4	V5	V6	V8	V9	V10	V11	V12	V14	V15	V16	V17	V18	V19¥V7	V13	
16	0.996923	−V12	V1	V2	V3	V4	V5	V6	V8	V9	V10	V11	V14	V15	V16	V17	V18	V19¥V12	V7	V13	
17	0.997073	+V13	V1	V2	V3	V4	V5	V6	V8	V9	V10	V11	V13	V14	V15	V16	V17	V18	V19¥V12	V7	
16	0.997038	−V14	V1	V2	V3	V4	V5	V6	V8	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V14	V12	V7	
17	0.997124	+V12	V1	V2	V3	V4	V5	V6	V8	V9	V10	V11	V12	V13	V15	V16	V17	V18	V19¥V14	V7	
16	0.997038	−V12	V1	V2	V3	V4	V5	V6	V8	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V12	V14	V7	
15	0.996928	−V2	V1	V3	V4	V5	V6	V8	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V2	V12	V14	V7	
14	0.996707	−V3	V1	V4	V5	V6	V8	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V3	V2	V12	V14	V7	
13	0.996335	−V8	V1	V4	V5	V6	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V8	V3	V2	V12	V14	V7	
12	0.995927	−V16	V1	V4	V5	V6	V9	V10	V11	V13	V15	V17	V18	V19¥V16	V8	V3	V2	V12	V14	V7	
11	0.995396	−V4	V1	V5	V6	V9	V10	V11	V13	V15	V17	V18	V19¥V4	V16	V8	V3	V2	V12	V14	V7	
10	0.994610	−V1	V5	V6	V9	V10	V11	V13	V15	V17	V18	V19¥V1	V4	V16	V8	V3	V2	V12	V14	V7	
9	0.993811	−V9	V5	V6	V10	V11	V13	V15	V17	V18	V19¥V9	V1	V4	V16	V8	V3	V2	V12	V14	V7	
10	0.994751	+V3	V3	V5	V6	V10	V11	V13	V15	V17	V18	V19¥V9	V1	V4	V16	V8	V2	V12	V14	V7	
9	0.994009	−V19	V3	V5	V6	V10	V11	V13	V15	V17	V18¥V19	V9	V1	V4	V16	V8	V2	V12	V14	V7	
10	0.994866	+V4	V3	V4	V5	V6	V10	V11	V13	V15	V17	V18¥V19	V9	V1	V16	V8	V2	V12	V14	V7	
9	0.994303	−V6	V3	V4	V5	V10	V11	V13	V15	V17	V18¥V6	V19	V9	V1	V16	V8	V2	V12	V14	V7	
10	0.994957	+V16	V3	V4	V5	V10	V11	V13	V15	V16	V17	V18¥V6	V19	V9	V1	V8	V2	V12	V14	V7	
9	0.994347	−V11	V3	V4	V5	V10	V13	V15	V16	V17	V18¥V11	V6	V19	V9	V1	V8	V2	V12	V14	V7	
8	0.993690	−V16	V3	V4	V5	V10	V13	V15	V17	V18¥V16	V11	V6	V19	V9	V1	V8	V2	V12	V14	V7	
7	0.992562	−V4	V3	V5	V10	V13	V15	V17	V18¥V4	V16	V11	V6	V19	V9	V1	V8	V2	V12	V14	V7	
6	0.990942	−V17	V3	V5	V10	V13	V15	V18¥V17	V4	V16	V11	V6	V19	V9	V1	V8	V2	V12	V14	V7	
5	0.988806	−V10	V3	V5	V13	V15	V18¥V10	V17	V4	V16	V11	V6	V19	V9	V1	V8	V2	V12	V14	V7	
6	0.991365	+V7	V3	V5	V7	V13	V15	V18¥V10	V17	V4	V16	V11	V6	V19	V9	V1	V8	V2	V12	V14	
7	0.992996	+V17	V3	V5	V7	V13	V15	V17	V18¥V10	V4	V16	V11	V6	V19	V9	V1	V8	V2	V12	V14	
6	0.991901	−V5	V3	V7	V13	V15	V17	V18¥V5	V10	V4	V16	V11	V6	V19	V9	V1	V8	V2	V12	V14	
7	0.993108	+V16	V3	V7	V13	V15	V16	V17	V18¥V5	V10	V4	V11	V6	V19	V9	V1	V8	V2	V12	V14	
8	0.993714	+V4	V3	V4	V7	V13	V15	V16	V17	V18¥V5	V10	V11	V6	V19	V9	V1	V8	V2	V12	V14	
7	0.993108	−V4	V3	V7	V13	V15	V16	V17	V18¥V4	V5	V10	V11	V6	V19	V9	V1	V8	V2	V12	V14	
6	0.991983	−V15	V3	V7	V13	V16	V17	V18¥V15	V4	V5	V10	V11	V6	V19	V9	V1	V8	V2	V12	V14	
5	0.990663	−V16	V3	V7	V13	V17	V18¥V16	V15	V4	V5	V10	V11	V6	V19	V9	V1	V8	V2	V12	V14	
4	0.986957	−V3	V7	V13	V17	V18¥V3	V16	V15	V4	V5	V10	V11	V6	V19	V9	V1	V8	V2	V12	V14	
3	0.981027	−V17	V7	V13	V18¥V17	V3	V16	V15	V4	V5	V10	V11	V6	V19	V9	V1	V8	V2	V12	V14	
4	0.987209	+V5	V5	V7	V13	V18¥V17	V3	V16	V15	V4	V10	V11	V6	V19	V9	V1	V8	V2	V12	V14	
3	0.984133	−V7	V5	V13	V18¥V7	V17	V3	V16	V15	V4	V10	V11	V6	V19	V9	V1	V8	V2	V12	V14	
2	0.970687	−V18	V5	V13¥V18	V7	V17	V3	V16	V15	V4	V10	V11	V6	V19	V9	V1	V8	V2	V12	V14	

表III 変数の選択過程 (Forward)

q	RV	variable	Y_1 ¥ Y_2																	
2	0.970687		V5	V13¥V1	V2	V3	V4	V6	V7	V8	V9	V10	V11	V12	V14	V15	V16	V17	V18	V19
3	0.984133	+V18	V5	V13	V18¥V1	V2	V3	V4	V6	V7	V8	V9	V10	V11	V12	V14	V15	V16	V17	V19
4	0.987209	+V7	V5	V7	V13	V18¥V1	V2	V3	V4	V6	V8	V9	V10	V11	V12	V14	V15	V16	V17	V19
5	0.990192	+V3	V3	V5	V7	V13	V18¥V1	V2	V4	V6	V8	V9	V10	V11	V12	V14	V15	V16	V17	V19
6	0.991812	+V17	V3	V5	V7	V13	V17	V18¥V1	V2	V4	V6	V8	V9	V10	V11	V12	V14	V15	V16	V19
7	0.992996	+V15	V3	V5	V7	V13	V15	V17	V18¥V1	V2	V4	V6	V8	V9	V10	V11	V12	V14	V16	V19
8	0.993655	+V19	V3	V5	V7	V13	V15	V17	V18	V19¥V1	V2	V4	V6	V8	V9	V10	V11	V12	V14	V16
9	0.994179	+V11	V3	V5	V7	V11	V13	V15	V17	V18	V19¥V1	V2	V4	V6	V8	V9	V10	V12	V14	V16
10	0.994789	+V10	V3	V5	V7	V10	V11	V13	V15	V17	V18	V19¥V1	V2	V4	V6	V8	V9	V12	V14	V16
11	0.995274	+V4	V3	V4	V5	V7	V10	V11	V13	V15	V17	V18	V19¥V1	V2	V6	V8	V9	V12	V14	V16
12	0.995661	+V9	V3	V4	V5	V7	V9	V10	V11	V13	V15	V17	V18	V19¥V1	V2	V6	V8	V12	V14	V16
13	0.996009	+V16	V3	V4	V5	V7	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V1	V2	V6	V8	V12	V14
14	0.996287	+V6	V3	V4	V5	V6	V7	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V1	V2	V8	V12	V14
15	0.996688	+V1	V1	V3	V4	V5	V6	V7	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V2	V8	V12	V14
16	0.996965	+V8	V1	V3	V4	V5	V6	V7	V8	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V2	V12	V14
17	0.997096	+V2	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V12	V14
18	0.997219	+V12	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V15	V16	V17	V18	V19¥V14
19	0.997256	+V14	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19

表Ⅳ 変数の選択過程 (Forward-backward)

q	RV	variable	Y_1 ¥ Y_2																	
2	0.970687		V5	V13¥V1	V2	V3	V4	V6	V7	V8	V9	V10	V11	V12	V14	V15	V16	V17	V18	V19
3	0.984133	+V18	V5	V13	V18¥V1	V2	V3	V4	V6	V7	V8	V9	V10	V11	V12	V14	V15	V16	V17	V19
4	0.987209	+V7	V5	V7	V13	V18¥V1	V2	V3	V4	V6	V8	V9	V10	V11	V12	V14	V15	V16	V17	V19
5	0.990192	+V3	V3	V5	V7	V13	V18¥V1	V2	V4	V6	V8	V9	V10	V11	V12	V14	V15	V16	V17	V19
6	0.991812	+V17	V3	V5	V7	V13	V17	V18¥V1	V2	V4	V6	V8	V9	V10	V11	V12	V14	V15	V16	V19
7	0.992996	+V15	V3	V5	V7	V13	V15	V17	V18¥V1	V2	V4	V6	V8	V9	V10	V11	V12	V14	V16	V19
8	0.993655	+V19	V3	V5	V7	V13	V15	V17	V18	V19¥V1	V2	V4	V6	V8	V9	V10	V11	V12	V14	V16
9	0.994179	+V11	V3	V5	V7	V11	V13	V15	V17	V18	V19¥V1	V2	V4	V6	V8	V9	V10	V12	V14	V16
10	0.994789	+V10	V3	V5	V7	V10	V11	V13	V15	V17	V18	V19¥V1	V2	V4	V6	V8	V9	V12	V14	V16
11	0.995274	+V4	V3	V4	V5	V7	V10	V11	V13	V15	V17	V18	V19¥V1	V2	V6	V8	V9	V12	V14	V16
10	0.994872	−V7	V3	V4	V5	V10	V11	V13	V15	V17	V18	V19¥V7	V1	V2	V6	V8	V9	V12	V14	V16
9	0.994200	−V11	V3	V4	V5	V10	V13	V15	V17	V18	V19¥V11	V7	V1	V2	V6	V8	V9	V12	V14	V16
10	0.994872	+V11	V3	V4	V5	V10	V11	V13	V15	V17	V18	V19¥V7	V1	V2	V6	V8	V9	V12	V14	V16
11	0.995364	+V16	V3	V4	V5	V10	V11	V13	V15	V16	V17	V18	V19¥V7	V1	V2	V6	V8	V9	V12	V14
12	0.995772	+V6	V3	V4	V5	V6	V10	V11	V13	V15	V16	V17	V18	V19¥V7	V1	V2	V8	V9	V12	V14
13	0.996195	+V9	V3	V4	V5	V6	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V7	V1	V2	V8	V12	V14
12	0.995816	−V16	V3	V4	V5	V6	V9	V10	V11	V13	V15	V17	V18	V19¥V16	V7	V1	V2	V8	V12	V14
13	0.996195	+V16	V3	V4	V5	V6	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V7	V1	V2	V8	V12	V14
14	0.996571	+V1	V1	V3	V4	V5	V6	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V7	V2	V8	V12	V14
15	0.996928	+V8	V1	V3	V4	V5	V6	V8	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V7	V2	V12	V14
14	0.996571	−V8	V1	V3	V4	V5	V6	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V8	V7	V2	V12	V14
15	0.996928	+V8	V1	V3	V4	V5	V6	V8	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V7	V2	V12	V14
16	0.997037	+V2	V1	V2	V3	V4	V5	V6	V8	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V7	V12	V14
17	0.997124	+V12	V1	V2	V3	V4	V5	V6	V8	V9	V10	V11	V12	V13	V15	V16	V17	V18	V19¥V7	V14
16	0.997038	−V12	V1	V2	V3	V4	V5	V6	V8	V9	V10	V11	V13	V15	V16	V17	V18	V19¥V12	V7	V14
17	0.997124	+V12	V1	V2	V3	V4	V5	V6	V8	V9	V10	V11	V12	V13	V15	V16	V17	V18	V19¥V7	V14
18	0.997219	+V7	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V15	V16	V17	V18	V19¥V14
19	0.997256	+V14	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19

表Ⅴ 手順ごとの RV の変化 (羽根アリデータ)

	RV					All possible との差			
q	Back	Back-for	For	For-back	All poss	Back	Back-for	For	For-back
19	0.99726	0.99726	0.99726	0.99726	0.99726	0	0	0	0
18	0.99723	0.99723	0.99722	0.99723	0.99723	0	0	0.00001	0
17	0.99707	0.99712	0.99710	0.99712	0.99712	0.00005	0	0.00003	0
16	0.99692	0.99704	0.99696	0.99704	0.99704	0.00011	0	0.00007	0
15	0.99670	0.99693	0.99669	0.99693	0.99693	0.00023	0	0.00024	0
14	0.99634	0.99671	0.99629	0.99657	0.99671	0.00037	0	0.00042	0.00014
13	0.99583	0.99634	0.99601	0.99620	0.99634	0.00051	0	0.00033	0.00014
12	0.99521	0.99593	0.99566	0.99582	0.99593	0.00072	0	0.00027	0.00011
11	0.99452	0.99540	0.99527	0.99536	0.99540	0.00087	0	0.00012	0.00003
10	0.99388	0.99496	0.99479	0.99487	0.99496	0.00108	0	0.00017	0.00008
9	0.99300	0.99435	0.99418	0.99420	0.99438	0.00138	0.00003	0.00020	0.00018
8	0.99219	0.99371	0.99366	0.99366	0.99385	0.00167	0.00014	0.00020	0.00020
7	0.99107	0.99311	0.99300	0.99300	0.99311	0.00204	0	0.00011	0.00011
6	0.98925	0.99198	0.99181	0.99181	0.99202	0.00277	0.00004	0.00021	0.00021
5	0.98622	0.99066	0.99019	0.99019	0.99078	0.00456	0.00012	0.00059	0.00059
4	0.98163	0.98721	0.98721	0.98721	0.98721	0.00557	0	0	0
3	0.97554	0.98413	0.98413	0.98413	0.98413	0.00859	0	0	0
2	0.96813	0.97069	0.97069	0.97069	0.97069	0.00256	0	0	0

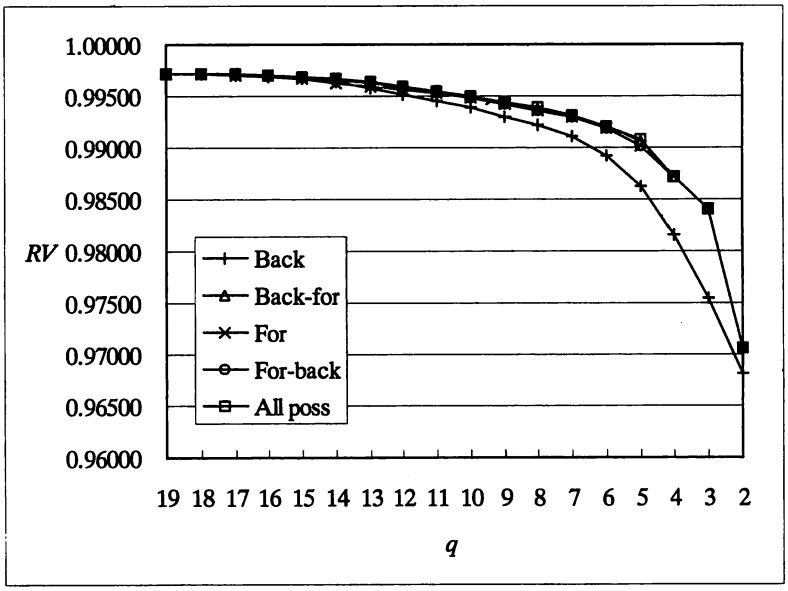

図1 手順ごとの RV の変化（羽根アリデータ）

変数全体が構成する主成分の布置と大きな差はないということである。具体的に $q=7$ のときの選択された変数群（表Ⅵ）と主成分スコアの散布図（図2）を示す。図2の(a)は全19変数を通常の主成分分析に適用して得られた第1主成分と第2主成分の散布図である。元の行列 Y と得られた主成分スコア行列 $Z=YA$ の RV 係数は0.99726である。(b)は RV 規準による手順のうち Backward-forward 手順が選択した7変数 {V3, V7, V13, V15, V16, V17, V18} による拡張主成分スコアの散布図である。Backward-forward は表Vより All possible との差が最も少なく，$q=7$ では All possible と同じ変数を選択している。このときの RV 係数は0.99311である。一方，(c)は同じ7変数を通常の主成分分析に適用して得られた主成分スコアである。7変数だけをデータとしているので得られた主成分には残りの12変数の情報は含んでいないことになる。このときの RV 係数は0.96840である。これより，拡張主成分の方が通常の主成分より元の変数の情報をよく再現していることがわかる。

次に，外的変量を用いない変数選択として考えられる手法や主成分分析における変数選択の先行研究との比較を示す。比較する選択手順は，回帰分析による方法，クラスター分析による方法，Jolliffe (1972[8], 1973[9], 1986[10]) の B2 と B4，これに M. PCA の P（規準1）の Forward-backward を加えたものである。図3にそれぞれの選択結果として RV の変化を示す。M. PCA の RV 規準による選択結果は最も RV の値が高い Backward-forward と最も低い Backward のみ掲載した。公平を期すために，各 q における RV は，それぞれの手法によって選択された変数群を Y_1 として(1)式を解いた固有値を用いて(5)式で算出している。M. PCA の RV 規準による4つの手順では1変数ずつの増減しか

表VI　$q=7$ のときの RV の値と選択された変数（羽根アリデータ）

手順	RV	選択された変数												
Bacward	0.99107		V5	V6		V10	V11		V14			V17		V19
Back-for	0.99311	V3			V7			V13		V15	V16	V17	V18	
Forward	0.99300	V3	V5		V7			V13		V15		V17	V18	
For-back	0.99300	V3	V5		V7			V13		V15		V17	V18	
All possible	0.99311	V3			V7			V13		V15	V16	V17	V18	

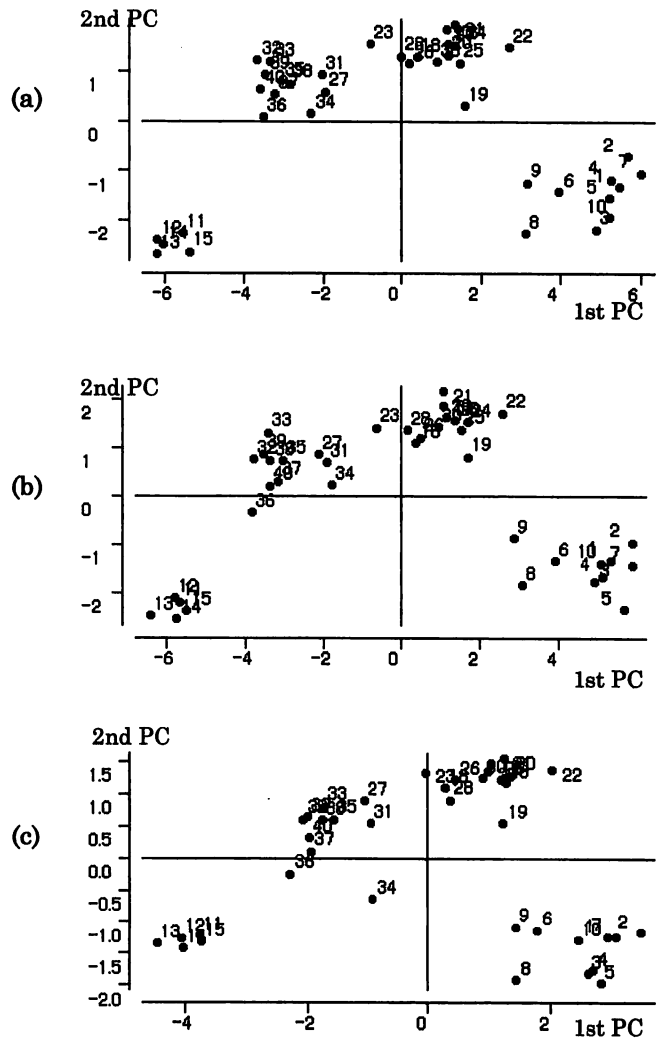

図2　主成分スコアの散布図（羽根アリデータ）
(a)　19変数による通常の主成分スコア
(b)　選択された7変数 {V3, V7, V13, V15, V16, V17, V18} による拡張主成分スコア
(c)　同じ7変数による通常の主成分スコア

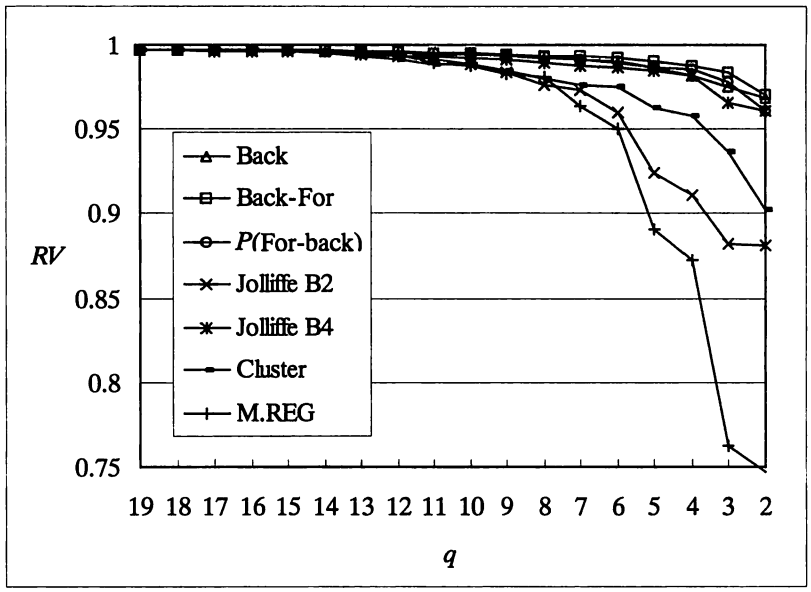

図3 手順ごとの RV の変化（先行研究との比較，羽根アリデータ）

行っていないため，最適な変数を選んでいる保証はなく，他の手法の方がそれらより RV の値を大きくする変数群を選び得る可能性がある。実際，Backward は M. PCA の P 規準より RV の値が小さくなっている部分が観察される。しかし，図3を見る限り，M. PCA による Backward 以外の3つの手順は他の手法より元の主成分の布置をよく再現する変数群を選んでいるといえる。全体的な RV の値の変化を見ると，どの手法においても $q=14$ までは RV の値の変化は微小で，手法間の差もほとんどない。先の図1の観察と合わせて，このデータの5つの変数は主成分の抽出には冗長であることがわかる。また，手法間の差が大きく開き始めるのは $q=9\sim6$ のところである。このあたりまで変数を減らすことが可能であると同時に，変数の選び方（規準）を慎重に決めなければならないことを示唆しているといえる。

手法ごとにどの変数が選ばれているかについては，Krzanowski (1987 a)[12]が，この羽根アリデータに対して，$r=2$, $q=4$ のとき，自ら提唱する選択手法と先行研究である先の Jolliffe や主変数を選択しようとする McCabe(1984)[11]の結果と比較したものがあるので，その結果に，M. PCA の RV 規準の5つの手順と上記先行研究の5つの選択結果（RV の値と選択された変数名）を付け加えてみる（表Ⅶ）。Krzanowski (1987 a)[12]の選択手法は，プロクラステス規準を用いて q 変数と $q-1$ 変数の主成分得点の布置の近さを評価しながら逐次変数を減らしていく Backward である。選択規準としては，M. PCA の RV も Krzanowski の方法もともに主成分の布置の近さを問題にしているが，RV 係数とプロクラス変換の違い以外に，Y_2 の情報も含んだ拡張主成分を利用しているか Y_1 のみから算出される通常の主成分を用いているか，また，常に元の変数全体との近さを比較している

か，step ごとに前後の近さを比較しているか，といった違いがある。実際，元の変数全体から求められる主成分と選ばれた 4 変数から求められる通常の主成分との RV 係数は，Krzanowski の 4 変数では0.98114で，M. PCA の RV 規準が選択した 4 変数では0.98721

表Ⅶ　$q=4$ のときの RV の値と選択された変数（先行研究との比較，羽根アリデータ）

手順	RV	選択された変数
Bacward	0.98163	V5 V6 V14 V19
Back-for	0.98721	V5 V7 V13 V18
Forward	0.98721	V5 V7 V13 V18
For-back	0.98721	V5 V7 V13 V18
All possible	0.98721	V5 V7 V13 V18
M. Reg	0.87245	V5 V16 V17 V19
Cluster	0.95772	V1 V2 V9 V12
Jolliffe'sB2	0.91121	V5 V8 V11 V13 V17
Jolliffe'sB4	0.98152	V5 V11 V13 V17
P(For-back)	0.98565	V5 V13 V17 V18
McCabe	0.89508	V9 V11 V17 V19
Krzanowski	0.98114	V5 V12 V14 V18

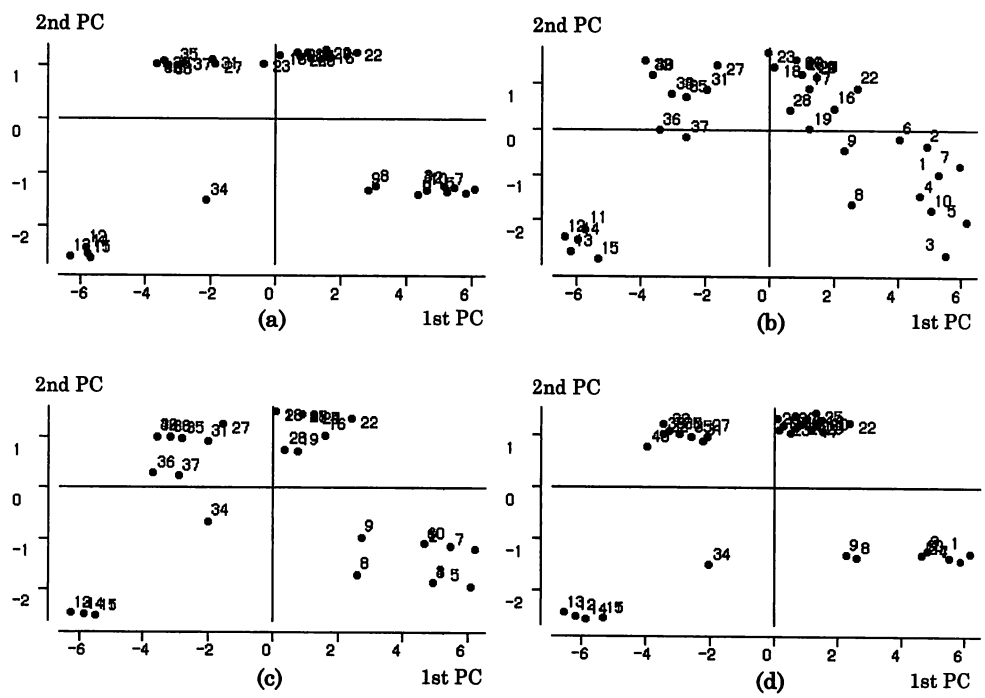

図 4　主成分スコアの散布図（4 変数の場合の拡張主成分スコア，羽根アリデータ）
(a) RV による {V5, V7, V13, V18}
(b) Jolliffe's B4 による {V5, V11, V13, V17}
(c) P（規準 1）の Forward-backward による {V5, V13, V17, V18}
(d) Krzanowski による {V5, V12, V14, V18}

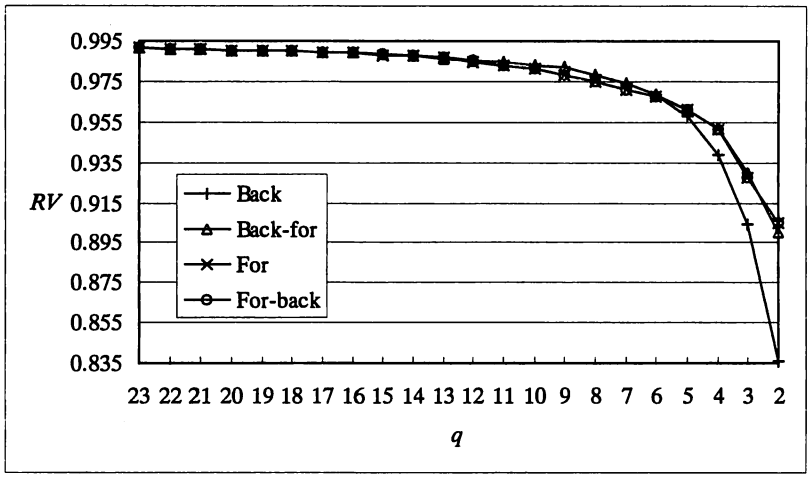

図5 手順ごとの RV の変化（MDOC データ）

である。図4は各手法によって選択された4変数から求められる拡張主成分の第1主成分と第2主成分の散布図で，RV の値が大きい方から4つの手法，(a) RV，(b) Jolliffe's B4，(c) P の Forward-backward，(d) Krzanowski について示した。RV 係数はそれぞれ (a) 0.98721，(b) 0.98152，(c) 0.98565，(d) 0.98114である。

最後に別のデータとして，87個体×23変数の「MDOC データ（軽症意識障害，佐野他，1977[15]）」に適用する。MDOC データの変数はすべて4値または2値のカテゴリカルデータであるので，最小交互二乗法（Young et al.，1976[16]）による数量化を施した上で，各手順を適用した。$r=2$ である。図5に4手順の RV の値の変化を示す。$q=5$ 以下では Backward が他の3手法より RV の値が低くなる変数群を選択しているが，「羽根アリデータ」と同様，4手順に大きな差はなく，このデータも RV 規準の意味で冗長な変数を含んでいることが示されている。

4．考　　察

前節の数値例での検討から，次のような傾向が得られた。

(1) Stepwise の利用によって，単調選択系より高い RV の値を得られる変数群を選択できる。

(2) q が大きい（選択する変数の数が多い）ところでは，4つの手順でほぼ同じ RV の値を得ている。

(3) q が小さくなると Backward による単調選択より他の3手順の方が RV の値が高い変数群を選択し得る。

(4) 4つの手順は All Possible と比較して顕著な差はない。

(5) 先行研究の選択手法との比較では，Backward を除いた M. PCA の3つの手順に

よって選ばれた変数群の方がより高い RV の値を得ることこができる。

いずれもデータの特質に左右されるものではあるが，限られた時間内でよりよい変数群を選択するためには，提唱した4つの手順は実用的であり，当然ながら，単調選択系よりstepwise 系が総じて優秀であることがわかった。

実際，事前に決めた主成分数 r の範囲では，減らしても RV 係数に変化を与えない変数が数個含まれていることがM. PCAによる変数選択の根拠となっている。羽根アリデータや MDOC データでは半数前後の変数を落としても RV の変化が小さい。落とされた変数が異なっても4つの手順で RV の値にあまり差がないような q の部分では，それらの変数を落としても十分に拡張主成分が元の変数の布置を再現できるといえる。

今回，M. PCA の2つの規準のうち未検討であった（基準2）の RV 係数規準について数値的検討を行った。数値例が示す通り，元の全変数の布置が一部の変数に基づく主成分で再現が可能で，図1のようなRVの値の変化を観察することにより利用可能な変数の数やそのときの主成分が取り出せることがわかった。（規準1）のPによる規準と合わせて，目的に合わせて使い分けることで1節に述べたような実際的な場面での活用が可能と考える。

今後は，4手順の使い分けや応用，合理的な q の数の決定方法，調査での再現性など実用場面での考察が課題である。

参考文献

1）森　裕一，垂水共之，田中　豊 (1994)．変数の一部を用いた総合指標の抽出．第8回日本計算機統計学会シンポジウム論文集，16-19.

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

3）Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya,* A, **26**, 329-58.

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

5）Mori, Y., Tanaka, T. and Tarumi, T. (1997). Principal component analysis based on a subset of qualitative variables. *Proceedings of IFCS-96: Data Science, Classification and Related Methods*, Springer-Verlag, 547-554.

6）Mori, Y. (1997). Statistical Software VASPCA — Variable Selection in PCA —. 岡山理科大学紀要，**38**(A)，329-340.

7）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)（受理済）.

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

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

10）Jolliffe, I. T. (1986). *Principal component analysis.* New York: Springer-Verlag.

11) McCabe, G. P. (1984). Principal Variables. *Technometrics*, **26**, 137-44.
12) Krzanowski, W. J. (1987a). Selection of variables to preserve multivariate data structure, using principal components. *Appl. Statist.*, **36**, 22-33.
13) Krzanowski, W. J. (1987b). Cross-validation in principal component analysis. *Biometrics*, **43**, 575-84.
14) Jeffers, J. N. R. (1967). Two case studies in the application of principal component analysis. *Appl. Statist.*, **16**, 225-236.
15) 佐野圭司他(1977). 軽症意識障害の評価方法に関する統計的研究 — 断面調査による特徴的臨床像の抽出. 神経進歩, 1052-65.
16) Young, F. et al. (1978). The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, **43**, 279-281.

Principal Component Analysis Based on a Subset of Variables

—— Numerical Investigation Using RV-coefficient Criterion ——

Yuichi MORI
Department of Socio-Information,
Faculty of Informatics,
Okayama University of Science
Ridai-cho 1-1, Okayama 700-0005, Japan
(Received October 5, 1998)

Principal component analysis based on a subset of variables proposed by Tanaka and Mori (1997) tries to extract reasonable principal components which are computed using a subset of variables but represent all the variables very well. This method has two mathematical tools, i.e., the ideas of Rao (1964)'s principal component analysis of instrumental variables and Robert and Escoufier (1976)'s RV-coefficient. In this paper to evaluate the performance of the RV-coefficient criterion, four variable selection procedures, Backward, Forward, Backward-forward stepwise and Forward-backward stepwise are applied to a couple of real data set. In the numerical examples, the criterion is verified in comparison with the results of all possible selection procedure and previous methods of variable selection in principal component analysis.