統計数理 (2001)
第 49 巻 第 2 号 277–292

特集「統計環境におけるインターネットの利用」
［統計ソフトウェア］

主成分分析における変数選択プログラムの WWW への実装

飯塚 誠也[1]・森 裕一[2]・垂水 共之[3]・田中 豊[3]

（受付 2001 年 3 月 30 日；改訂 2001 年 5 月 24 日）

要　　旨

主成分分析における変数選択プログラム VASPCA (VAriable Selection in Principal Component Analysis) の Web バーションである VASPCA/Web について報告する．VAPCA/Web は，Ruby と Perl を用いた CGI によって動作し，R を統計エンジンとするサーバーサイド型のオンラインプログラムである．Web 上に実装することにより，広く統計解析（ここでは主成分分析における変数選択）が実行できる環境を提供し，新しい手法などを即時的に公開し体験できるプラットホームの構築を目指したものである．このプログラムでは，先行研究を含めた 7 つの変数選択手法が実行でき，ユーザーの要求に合わせた選択が可能である．プログラムのフローは，(1) データの入力 (2) 変数選択手法の選択 (3) 通常の主成分分析の実行 (4) 主成分数・選択規準・選択手順の指定 (5) 変数選択の実行 (6) 結果（選択変数群と規準値）の出力となっている．これにより，主成分分析における変数選択が手軽に行えるようになり，複数ある選択規準のうち，目的の選択規準ですぐに選択を実行したり，いくつかの選択規準による結果を比較したりできるなど，実際の場面に対応できる環境を構築することができた．

キーワード：統計解析プログラム，オンラインプログラム，VASPCA/Web.

1. はじめに

検査項目の数を減らしたい場合やある解析の事前分析や事後分析として変数を減らしたい場合など，主成分分析における変数選択が必要になるときがある．このような場面で行われる主成分分析における変数選択には，回帰分析のような外的変量がある手法での変数選択と異なり，次のような特徴がある．

・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

このような特徴の下，主成分分析における変数選択を実際に行う場面を考えるならば，変数を選択する目的がはっきりしていれば，その選択規準で選択を行いたいし，目的がはっきりしていない場合や複数の変数候補を検討してみたいという場合は，いくつかの手法を試してみて

[1]岡山大学 法学部：〒700–8530 岡山市津島中 3–1–1
[2]岡山理科大学 総合情報学部：〒700–0005 岡山市理大町 1–1
[3]岡山大学 環境理工学部：〒700–8530 岡山市津島中 3–1–1

その結果を比較したくなる．しかしながら，これまで，いずれの場合においても，複数存在している選択規準で変数選択をすぐに実行できるような環境はなかった．

そこで，すでに公表・研究されてきた主成分分析における変数選択手法や新しい選択手法を 1 つのパッケージに取り込んだ変数選択プログラム VASPCA (VAriable Selection in Principal Component Analysis) を開発することにした．これは Windows 上で動く VASPCA/Win から開発が始まった (Mori (1997)) が，今回 Web 上で動くオンラインプログラム VASPCA/Web の初期バージョンの実装が完了したので，その詳細について報告する．

以下の節では，次の 2 点に焦点をあて主成分分析における変数選択プログラムについて議論する．

(1) これまでの主成分分析における変数選択手法の整理と特徴
(2) WWW 上のオンラインプログラムと VASPCA/Web の仕様

具体的には，2 節で主成分分析における変数選択の選択規準と VASPCA/Web に実装されている手法および選択手順について述べ，3 節で Web 上の統計解析プログラムの動向を概観し，VASPCA/Web の位置付けを示す．そして，4 節で VASPCA/Web の仕様と実際の動作を紹介し，最後に今後の課題等についてまとめる．

2. 主成分分析における変数選択

2.1 主成分分析における変数選択の規準

主成分分析における変数選択は，Jolliffe (1972, 1973)，Robert and Escoufier (1976)，McCabe (1984)，Bonifas et al. (1984)，Krzanowski (1987)，Falguerolles and Jmel (1993)，森 他 (1994) など多くの研究がなされてきている．また，一部の変数から全体の変数を最もよく再現するような主成分を抽出しようという拡張主成分分析 (Tanaka and Mori (1997)) の観点を利用した変数選択（森 他（1998); 森（1998））や変数の影響分析を利用した変数選択（森 他（2000））などのアイデアが提唱されている．

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 つに分類することができる．

・選択された変数群の分散共分散（変動）に注目するもの
・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置の近さを利用するもの
・その他

これらをまとめると表 1 のようになる．いずれの手法も変数を選択する規準としては可能なものであるので，先に述べた「選択の規準がいくつも存在する」という特徴が出てくる．また，規準が異なるので，選択される変数群も異なるものとなり，2 つ目の特徴「多くの場合，それぞれの選択規準による選択結果が異なる」が出てくるわけである．したがって，元の解析の目的や変数を選択する目的に照らし合わせて規準を採択して変数選択を行えばよいことになる．

2.2 VASPCA/Web に実装された変数選択規準

具体的には，どのような規準であるかを，実際に VASPCA/Web に実装されている選択手法を例に詳細を述べる．

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

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		
	変 動	空間上の布置	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）

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 $(1 < q < p)$．これに対応する $Y = (Y_1, Y_2)$ の分散共分散行列を $S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$, $S = (S_{11}, S_{12})$ とする．

(1) 先行研究の変数選択規準

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

Jolliffe (1972, 1973) 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，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 間の空間上の布置の近さを測る指標として，次の RV 係数を提唱している．

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

ただし，$\tilde{Y}$, $\tilde{Z}$ は Y と Z を中心化した行列である．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 変数を用いたときの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，元のデータと主成分得点の空間的な布置が最も近い q 変数群を見つけることが可能となる．

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 $C_1 \sim C_4$ の 4 つを提唱している．

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

$S_{22\cdot 1}$ 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，$\mathit{cancor}(Y_1, Y_2)$ は Y_1 と Y_2 の正準相関係数である．したがって，$C_1 \sim C_4$ のいずれかの規準を採用したとき，その規準を満たす Y_1 と Y_2 を求めればよいことになる．

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

$$M^2 = \mathrm{tr}(YY' + ZZ' - 2D)$$

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 変数を用いたときの主成分得点行列などをあてはめ，M^2 を最大化する q 変数を見つける手法を提案している．

(2) 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 $Z = Y_1 A$ 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 z を用いて y の予測効率を最大にする.

［規準 2］ Y と Z の RV 係数 $RV(Y,Z) = \mathrm{tr}(\tilde{Y}\tilde{Y}'\tilde{Z}\tilde{Z}')/\{\mathrm{tr}(\tilde{Y}\tilde{Y}')^2 \cdot \mathrm{tr}(\tilde{Z}\tilde{Z}')^2\}^{1/2}$ を最大にする.

$Y = (Y_1, Y_2)$ より得られる一般化固有値問題

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

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 $P = \sum_{j=1}^{r} \lambda_j / \mathrm{tr}(S)$, ［規準 2］は, Robert and Escoufier (1976) より, RV 係数 $RV = \{\sum_{j=1}^{r} \lambda_j^2 / \mathrm{tr}(S^2)\}^{1/2}$ が, それぞれ最大化の規準値となる.

この 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合せのうち, 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.

(3) 変数の影響分析を利用した変数選択

これは, 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である. 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ことを考える. 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,

$$\begin{aligned} S_{11} &\to S_{11} - \varepsilon(J_{jj}S_{11} + S_{11}J_{jj}) + O(\varepsilon^2)\,, \\ S_{12} &\to S_{12} - \varepsilon J_{jj}S_{12}\,, \\ S_{21} &\to S_{21} - \varepsilon S_{21}J_{jj} \end{aligned}$$

のように変化する $(j = 1, \ldots, q)$. 一般化固有値問題を $(C - \lambda D)\boldsymbol{a} = 0$ 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,

$$\begin{aligned} C^{(1)} &= -J_{jj}C - CJ_{jj} - 2S_{11}J_{jj}S_{11}\,, \\ D^{(1)} &= -J_{jj}S_{11} - S_{11}J_{jj} \end{aligned}$$

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 $\qquad \lambda_j^{(1)} = \boldsymbol{a}_j'(C^{(1)} - \lambda_j D^{(1)})\boldsymbol{a}_j$

(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

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

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

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

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 変数ず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．

(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. (2000))．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 $(n-1)\times p$ 行列 $Y_{(i)}$ を標準化したデータ行列，$A_{(i)}$，$Z_{(i)}$ 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．

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

ただし，$\tilde{\tilde{Y}}$ は Y を $Y_{(i)}$ の平均と標準偏差で標準化した行列，$\hat{Y}$ の i 番目の行を $\hat{y}_i = z_i B$ とし，z_i は $\tilde{\tilde{Y}}$ と $A_{(i)}$ に基づく主成分行列の i 番目の行，$B = (Z'_{(i)}Z_{(i)})^{-1}Z'_{(i)}\tilde{Y}_{(i)}$ 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．

なお，この手法は，計算コストが非常に高いことが問題点であり，現在，VASPCA/Web では，計算エンジンである R のソースを公開するにとどめている．

2.3 変数選択の手順

実際の変数選択の場面では，計算コストを考慮し，通常は変数減少法 (Backward) などの簡便法がとられるが，VASPCA/Web では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している（フローの詳細は Mori (1997)，森 他 (1998) などを参照）．ここで，Y_1 の変数の数が q のときの規準値（2.2 節の 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．ただし，規準の性質上，以下の 4 つの選択手順の一部しかできない規準もある．たとえば，変数の影響分析を利用する変数選択では，1 変数を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目的であるので，必然的に Backward のみとなる．

(1) Backward elimination（変数減少法）

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 Backward 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．

(2) Backward-forward stepwise selection（変数減増法）

Step A（初期段階）. Backward 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．

(3) Forward selection（変数増加法）

Step A（初期段階）. Backward 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 に付け加えて得られる $p-q$ 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 Forward 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．

(4) Forward-backward stepwise selection（変数増減法）

Step A（初期段階）. Forward の Step A と同じ．

Step B（変数選択段階）. Backward-Forward 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 系の方がよい結果が得られること，また単純選択と Stepwise 系の結果を比較してみると，Stepwise 系の手順の方がよい結果が得られることが明らかになっている（森 他（1998））.

また，この他に各 q におけるすべての変数の組合せについて計算する All possible (Forward の Step A) があるが，計算コストが非常に高くなるため VASPCA/Web には実装していない.

3. Web における統計解析と VASPCA/Web

2 節で見てきたように，主成分分析における変数選択規準は多種多様である．主成分分析における変数選択を用い解析することを目的とするユーザーが，これらの変数選択手法を自分でプログラミングすることに時間と労力を使うのは本来の目的とは異なる．また，新しい手法を広めることが目的であれば，その手法を簡単に実行できる環境がユーザーの手近に構築されると非常に便利である．特に，Web 上のものはすぐに実行でき，インストール作業もいらないため，簡単にインターネットに接続できる環境となった今では，強力なツールになり得る．そこ

表 2. 統計解析サイトの分類.

分類の観点	分　類	
開発目的からの分類	教　育	（統計教育のために開発されたサイト） （教育に利用できる統計ツール）
	データ解析	（アイデアや手法の試用や普及） （完全な解析サイト）
計算環境とプログラムテクニックからの分類	サーバーサイド型	CGI (＋スクリプト言語，＋統計エンジン)
	クライアントサイド型	Java アプレット　(インタフェースとして) (クライアントマシン上のプログラム)
		Java スクリプト
		その他 (VRML, XML，オリジナルのインターフェースなど)

で，われわれは Web 上で手軽に体験できる環境を開発することにした．

開発にあたって，先行する統計解析サイトの特徴を簡単に概観してみると，表 2 のように分類される（飯塚 他 (2000)）．すなわち，開発目的から分類すると，統計教育の観点とデータ解析そのものを目的にしたサイトに分かれる．また，計算環境とプログラムテクニックの観点で分類してみると，サーバーサイド型とクライアントサイド型，およびその両方の型が見られ，それぞれの型にあったプログラミングテクニックが用いられている．

開発にあたっては，表 2 のうちのどのスタイルをとるかという問題や，また，実際の運用にあたっての問題点などが出てくる．これらをまとめると，

- 計算はサーバーサイドでするのがよいか，クライアントサイドでするのがよいか．
- プログラムをすべて作り上げていくか，既存の統計パッケージなどを用いながら作り上げていくか．
- もし，統計パッケージを使うならば，ライセンスに制限はないか．
- もし，サーバーサイドのプログラムにするならば，同時に何人までのアクセスを許可するか．
- どのようにデータを受け取り，どのように結果をユーザーに与えるか（データの秘匿の問題や不特定多数に結果が漏れる可能性への対処も含む）．
- 解析に対するデータサイズや計算容量の制限はどうするか．

などである．

これらに対して，われわれの目的は，変数選択を手軽に利用できる環境を提供すること，および，主成分分析における新しい手法を広く公開することである．したがって，「データ解析」をサイトの開発目的とし，計算環境とプログラミングテクニックとしては「サーバーサイド型」を採用することにした．サーバーサイド型にしたのは，比較的容易に作ることができメンテナンスも簡単である CGI を用いることができることと，クライアントの計算スピードに依存したり，ユーザーの CPU を占有したりすることがないからである．また，ライセンスの問題も考慮し，計算部分では，既存の統計パッケージである R を統計エンジンとして用いることにした．R を選んだ理由は次の通りである．

- 多くの統計に関する関数をもっており，プログラムが比較的簡単である．
- フリーパッケージである．
- グラフ機能が優れている．

これにより，変数選択の規準が増えた場合でも，R の関数を作るだけで VASPCA/Web の更新が柔軟に素早くできるようになる．

4. オンラインプログラム VASPCA/Web

4.1 サイトの構成

VASPCA の URL は,

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

である．このサイトは，日本語と英語の両方が用意されている．VASPCA のサイトの構成は図 1 にあるように，主に 2 つの構成でできている．1 つは，主成分分析における変数選択の解説ページ，もう 1 つは実行ページ (VASPCA/Web と VASPCA/Win) である．

解説ページでは，手法を広めることも目的の 1 つであるので，初めて主成分分析における変数選択を体験するユーザーにも配慮し，具体的な例や図などを用いた解説を多く用意した (図 2, 図 3)．また，専門的な解説がほしいユーザーのために, 各規準を詳説するページも用意した（図 4)．さらに，参考資料として，われわれのプレゼンテーション資料なども公開している．

VASPCA/Web の実行ページは図 5 のような構成である．実際の計算部分は R，グラフ変換は convert コマンド，CGI スクリプトは原則として Ruby で作成した．Ruby（バージョン 1.6.1）を採用した理由は，日本語処理に非常に優れており CGI 作成も容易であるという点である．以前のバージョンでは Perl で作成していたが，一部を除きすべて Ruby に書き換えた．

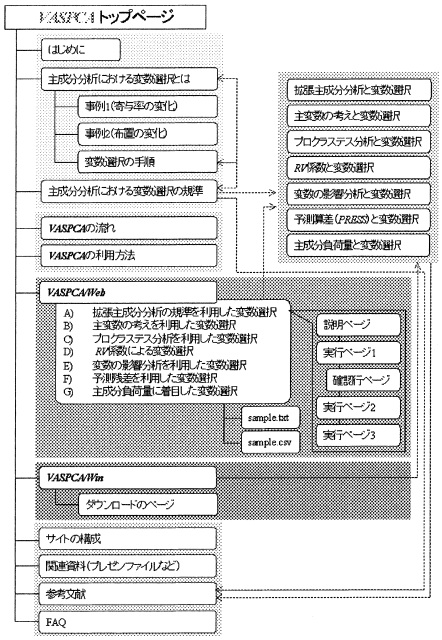

図 1. VASPCA のサイト構成.

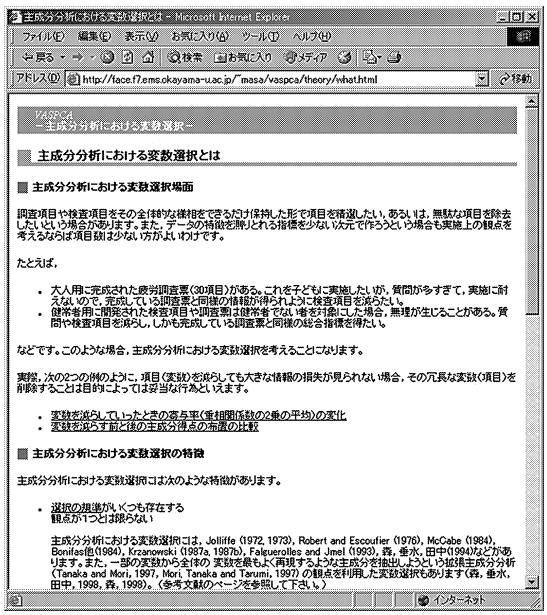

図 2. 解説ページ（主成分分析における変数選択の紹介).

図 3. 解説ページ（選択規準）.

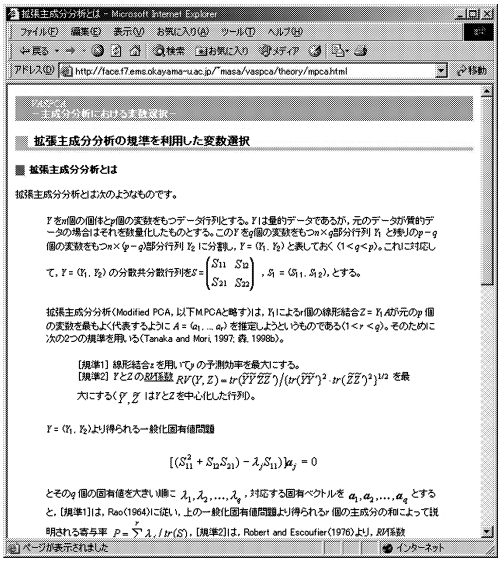

図 4. 解説ページ（選択規準の詳細：M.PCA の場合）.

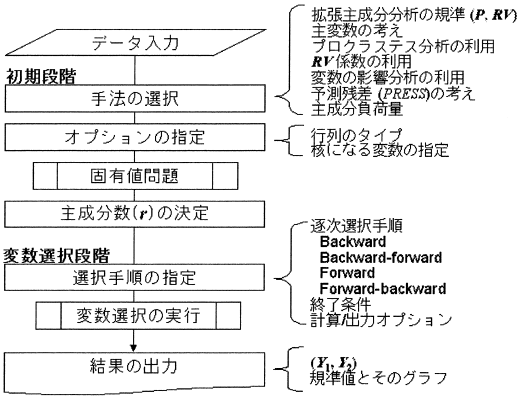

図 5. VASPCA/Web のフロー.

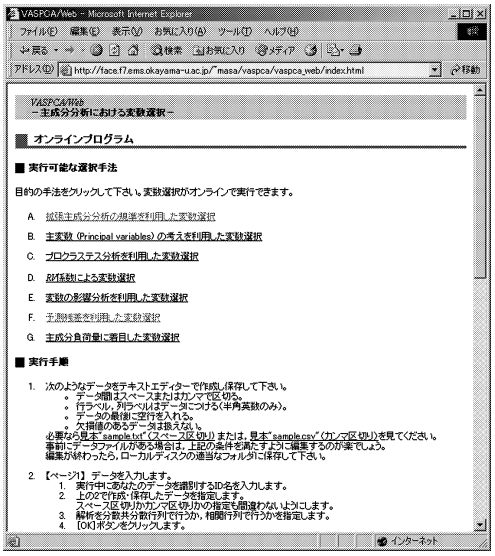

図 6. 実行ページ（選択手法の選択）.

4.2 実行可能な手法

実行可能な選択手法としては，2 節で述べた以下の 7 つを用意した.

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

このうち, B) の主変数の考えを利用した変数選択の規準 C4 の手順は, 規準の性質上, Backward および Forward のみ実行可能となっている．また，E) の変数の影響分析を利用した変数選択も Backward のみである．また，F) の予測残差を利用した変数選択では，計算コストが高いため，実装を控えている．これら以外の手法では，提案されているいずれの規準に対しても，4 つの選択手順すべてが適用できる．

4.3 実行例および解説

(1) VASPCA/Web への接続

VASPCA/Web に接続すると，図 6 のページが現れる．そこには上の 7 つの選択手法が選択できるようになっている．ここから，ユーザーは手法を選択し実行に移ることになる．このページには，実行手順とデータの作成法，および選択全体の流れが示されている．データ解析に慣れていないユーザーでも戸惑うことなく実行できるように配慮した．

(2) データ入力

手法を選択すると簡単な手法の説明ページが現れ(図 7)，そこで[実行ボタン]をクリックすると，図 8 のような【ページ 1】が現れる．ここで，ユーザーはデータを入力（選択）する．データには，計算部分で用いる R が効率よく処理できるように以下の条件をつけている．

・スペース区切りまたはカンマ区切り
・行，列ラベルをつける（ただし日本語は不可）

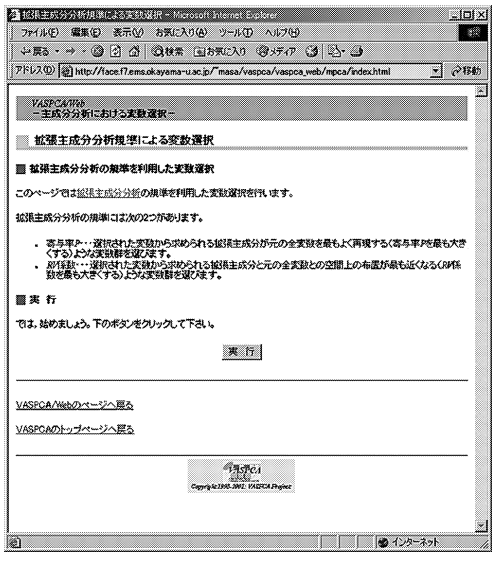

図 7. 実行ページ（簡単な説明).

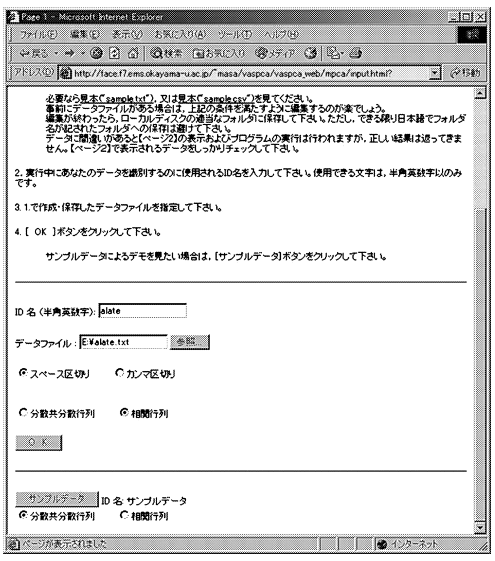

図 8. 実行ページ（【ページ 1】データ選択).

・欠損値は使用できない

上の条件に注意しながらテキストエディタや，MS-Excel などでデータを作成すればよい．参考として，sample.txt（スペース区切り），sample.csv（カンマ区切り）も用意してある．これらを参考にデータを作成し，ユーザーのローカルディスクに保存する．

データを作成したら，【ページ 1】の下にある各項目を指定する．具体的には，ID 名を入力し，作成したデータファイルを選択する．ID 名はデータを識別するために用いるので，ユーザーが自由に設定することができる．これは，データ名の一部として，一時的にサーバー側に保存される．なお，日本語は入力できないようになっている（日本語を使った場合は，正しく計算されない）．次に，「分散共分散行列」か「相関行列」のどちらを使用するか選択する．最後に，作成したデータ形式，すなわちデータがスペース区切りかカンマ区切りかを指定する．

一方，簡単に変数選択を体験してもらうために，サンプルによるデモをできるようにしてある．最下行の[サンプルデータ]ボタンをクリックすることでサンプルデータに対する変数選択が実際に行える．いずれの手法でもサンプルデータを利用することができるため，各規準および手法による結果の比較をすることも可能である．

(3)　実行前確認

(2) の指定を終え，[OK]ボタンをクリックすると，図 9 のデータの確認画面が現れる．この画面で，ID 名，ファイルのパス，受付番号，区切り文字，行列が表示されるので，それらを確認する．VASPCA/Web は，他のデータと重複しないように，「ID 名」＋「受付番号」のファイル名でデータを一時保存する．受付番号は，サーバー側で自動的に割り振った数字である．

すべて確認がとれたら，[計算開始]ボタンをクリックし，次のページに進む．

ただし，【ページ 1】でサンプルデータを選択した場合は，データ確認の必要がないので，このページは表示されない．

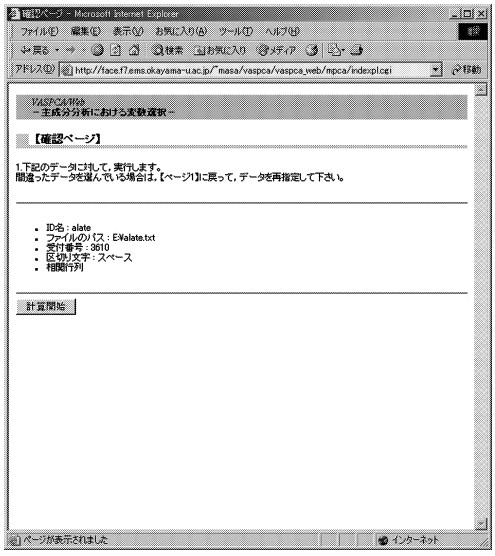

図 9.　実行ページ（【確認ページ】）.

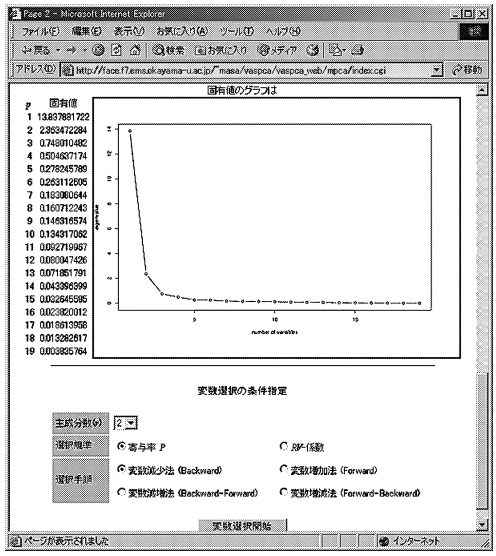

図 10.　実行ページ（【ページ 2】変数選択の条件指定）.

(4) 固有値の表示，選択条件の指定

続いて，【ページ 2】（図 10）が表示されるが，それまでにサーバー内では次のような処理が行われている．

1. 「ID 名」+「シリアル番号」というファイル名でデータを格納する．同時にファイルサイズのチェックを行う．現在，許容しているのは 4MB までである．
2. 通常の主成分分析を行い，固有値およびそのグラフを表示する R のソースを Ruby により作成する．
3. 格納されたデータを R に渡し，固有値の計算を行う．計算結果として，固有値を表示する．同時に，固有値のグラフを EPS ファイルとして一時的に保存する．
4. EPS ファイルを，convert コマンドを用いて JPEG ファイルで出力する．

以上より，次の内容を含んだ【ページ 2】が表示される．

・データの要約情報
・データの内容表示
・固有値の表示
・固有値のグラフの表示
・変数選択の条件指定のフォーム

データが表形式で表示されるので，ここでその内容について確認する．間違っている場合には，【ページ 1】に戻る．

データが間違っていなければ，ページの下の条件指定のフォームに必要事項を指定する．まず，固有値とそのグラフを参考にして，主成分数 r を決定する．次に，選択規準が複数ある場合はそのうちの 1 つを選び，さらに，選択手順を Backward, Forward, Backward-Forward, Forward-Backward の中から 1 つ選んで，[変数選択開始]をクリックする．

このページでは，「ID 名」＋「シリアル番号」という名前でデータを格納することにより，不特定多数のユーザーからデータを守る工夫をしている．すなわち，「ID 名」+「シリアル番号」は，【ページ 3】で出力する結果のテキストファイル名の一部として用いているからである．さらに，ここで保存されたデータと結果は，解析終了後ユーザーによって削除することができるようになっている．

また，【ページ 2】で選択した手法やデータのサイズによって，ユーザーが結果出力まで待ちきれないような状況を考え，結果を後から参照できるように，結果が保存されている URL を表示するようにしている．たとえば，ID 名が「alate」，受付番号が「14」，手法が M.PCA の場合，

http://face.f7.ems.okayama-u.ac.jp/~masa/vaspca/vaspca_web/mpca/data/alate14.txt

のように表示される．[変数選択開始] ボタンをクリックした後，別のページに移るなどしても，しばらくしてから上記 URL にアクセスすれば，結果を見ることができる．

計算部分を実行しているときには，他の計算は行えないようにしている．【ページ 2】では固有値の計算時にはロックを行う．これにより，ユーザーは別の計算を同時にすることはできないが，複数ユーザーが Forward などの計算コストが高い手法を同時に選んだためにサーバーに大きな負荷がかかることを避けるため，このような仕様としている．

(5) 結果表示

【ページ 2】で [変数選択開始] をクリックすると，結果を出力する【ページ 3】（図 11）が表示される．ここでは，データの要約情報（出力時刻，ID 名，データファイルのローカルパス，区切り文字，個体の数 (n)，変数の数 (p)，主成分数 (r)，解析に用いた行列の種類，選択規準，

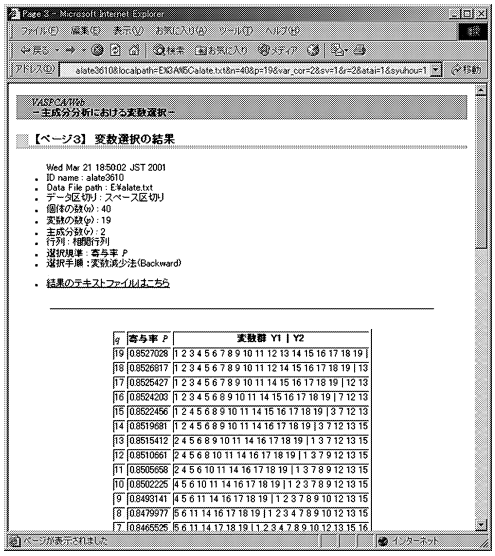

図 11. 実行ページ（【ページ 3】結果表示).

選択手順）と結果のテキストファイルへのリンクに続き，変数選択の結果が表示される．結果は，各 q における規準の値と，Y_1, Y_2 の変数群が表の形式で出力される．また，規準の値のグラフも出力され，視覚的に結果を眺めることができる．

以上が変数選択の 1 つのサイクルである．この結果を見て，ユーザーは，採用する変数の数を決めたり，採用する変数群を検討したりすることになる．

ユーザーによっては，現在の結果をサーバーに残しておきたくない場合もあるであろうし，あるいは，同じデータで主成分数を変えたり，別の選択手順を試したい場合もあろう．これに対して，まず，データと結果の削除については，[データと結果の削除]ボタンを用意した．これをクリックすると，格納された元のデータと計算された結果がすべてサーバーから削除される．一方，さらに続いて解析を行いたい場合については，ページの最下部に【ページ 2】と同様の選択条件の指定フォームを表示しているので，そこであらためて指定をし，解析を続けるようになっている．

また，別のデータあるいは別の手法で選択を行いたい場合は，(2) や (1) のページに戻って，変数選択を行うことになる．

このページでの処理は，CGI が【ページ 2】からパラメータを受け取り，それを用いて計算している．【ページ 2】から渡されるパラメータとは，先のデータの要約情報の各項目である．具体的な処理は次の通りである．

1. Ruby により，ユーザーから指定された変数選択規準，変数選択手順に対応する R のソースを作成する．
2. 格納されたデータを R に渡し，規準値の計算を行う．そのとき，R により変数群も出力している．計算結果を出力し，計算終了と同時に各 q における規準値のグラフを EPS ファイルとして一時的に保存する．
3. EPS ファイルを，convert コマンドを用いて JPEG ファイルで出力する．
4. 結果の HTML ファイルを表示し，同時にそのテキストも出力する．

5. EPS ファイル，R ソース，R で出力されたファイルを削除する．

計算部分で重要となる変数選択規準，変数選択手順の関数は，.Rdata としてサーバー側に用意した．上で作っているソースにあたるものは，その関数で使えるように作成している．.Rdata は Windows などプラットフォームが異なっても使うことができるので，要望があれば公開する予定である．

5. まとめ

以上のように，VASPCA/Web により，主成分分析における変数選択が Web 上で手軽に行えるようになり，複数の選択手法から自由に規準を選び，採択する変数群を選んだり，いくつかの規準による結果を比較・検討するなどの実行環境を構築することができた．

今後の課題としては，核となる変数を選択前に指定できるようにしたり，ストッピングルールを追加したりする機能面の充実がある．また，現在は公開を控えている *PRESS* についてもアルゴリズムなどを改良し実用に耐えるものにすることや，All possible の追加などが上げられる．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も必要と考えている．

参 考 文 献

Bonifas, I., Escoufier, Y., Gonzalez, P. L. and Sabatier, R. (1984). Choix de variables en analyse en composantes principales, *Revue de Statistique Appliqée*, **23**, 5–15.

Falguerolles, A. De and Jmel, S. (1993). Un critère de choix de variables en analyse en composantes principales fondé sur des modèles graphiques gaussiens particuliers, *Revue Canadienne de Statistique*, **21**(3), 239–256.

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.

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I. Artificial data, *Applied Statistics*, **21**, 160–173.

Jolliffe, I. T. (1973). Discarding variables in a principal component analysis. II. Real data, *Applied Statistics*, **22**, 21–31.

Krzanowski, W. J. (1987). Selection of variables to preserve multivariate data structure, using principal components, *Applied Statistics*, **36**, 22–33.

McCabe, G. P. (1984). Principal variables, *Technometrics*, **26**, 137–144.

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

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，垂水共之，田中 豊 (1994). 主成分分析における *RV* 係数を利用した変数選択, 計算機統計学, **7**, 47–56.

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.

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 28 回大会発表論文抄録集, 301–302.

Mori, Y., Iizuka, M., Tarumi, T. and Tanaka, Y. (2000). Study of variable selection criteria in data analysis, *Proceedings of the Tenth Japan and Korea Joint Conference of Statistics*, 119–124.

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

Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The *RV*-coefficient, *Applied Statistics*, **25**, 257–265.

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

Implementation of Variable Selection Program for Principal Component Analysis to WWW

Masaya Iizuka

(Faculty of Law, Okayama University)

Yuichi Mori

(Faculty of Informatics, Okayama University of Science)

Tomoyuki Tarumi and Yutaka Tanaka

(Faculty of Environmental Science and Technology, Okayama University)

This paper discusses VASPCA/Web, which is a web version of the statistical software VASPCA (VAriable Selection in Principal Component Analysis) for variable selection in principal component analysis. VASPCA/Web is an on-line CGI program written in Ruby and Perl and constructed in our server with the statistical package R as a computation engine. This program is implemented in WWW to provide an environment for performing statistical analysis (variable selection in principal component analysis in this program) and rapidly spreading some new statistical methods so that many users can access them. Users can perform on demand any one of seven selection methods implemented in VASPCA/Web. Some of the methods have been proposed in previous studies and others use new possible selection criteria. The flow of VASPCA/Web is as follows: (1) Entering data (2) Choosing a method (3) Applying ordinary principal component analysis to the data (4) Specifying the number of principal components, a selection criterion and a selection procedure (5) Selecting a subset of variables sequentially (6) Outputting the result (subsets of selected variables and criterion values). This system makes it easy for users in various application fields to perform variable selection in principal component analysis.

Key words: Statistical program, on-line program, VASPCA/Web.