Statistical Software VASPCA

—— Variable Selection in PCA ——

Yuichi MORI
Department of Socio-Information
Faculty of Informatics
Okayama University of Science
Ridai-cho 1-1, Okayama 700-0005, Japan
(Received October 6, 1997)

Abstract

Statistical software VASPCA is developed for variable selection in principal component analysis (PCA). Though this software intends to include a variety of selection procedures proposed by some authors, this version performs variable selection procedures using the idea of the modified PCA proposed by Tanaka and Mori [9]. It has four types of procedure, Backward, Forward, Backward-forward and Forward-backward in which one variable is removed/added at each selection step. Their practical actions are illustrated by applying them to a real data set.

1 Introduction

Principal component analysis (PCA) is a statistical method which reduces the dimensionality of the space using appropriate components. It is stated that in many applications it is desirable not only to reduce the dimension of space, but also to reduce the number of variables that are considered or measured in the future. Suppose we wish to apply PCA to make a small-dimensional rating scale which measures latent traits. From the validity aspect, in order to gather important dimensions well, all the variables should be included. On the other hand, from the aspect of practical application, the number of variables should be as small as possible to avoid waste of time and resources and difficult interpretation of components extracted from too many variables. Hence it is essential to reduce the number of variables as well as possible without disturbing the original features.

The problems of variable selection in PCA have been studied by Jolliffe[1)2)3)], McCabe[4)], Krzanowski[5)6)], Robert and Escoufier[7)], Mori, Tarumi and Tanaka[8)], Tanaka and Mori[9)] and Mori, Tanaka and Tarumi[10)] among others. Jolliffe's methods[1)2)3)] are based on the way to remain the variables related to important principal components (PCs) or to reject those related to unimportant PCs by observing the eigenvalues and the coefficients of the corresponding eigenvectors. McCabe's methods[4)] select variables containing in some sense as much sample information as possible. Krzanowski's method[5)6)] and Mori, Tarumi and Tanaka's method[8)] use the criteria based on Procrustes Analysis and the RV-coefficient, respectively. The aim of their approaches

is to select a subset of variables based on ordinary PCs using the selected variables in such a way that they retain as much information as possible comparing with PCs using all the variables. As against these two studies, Robert and Escoufier[7], Tanaka and Mori[9] and Mori, Tanaka and Tarumi[10] discussed modified PCs which are computed using not only a selected subset of variables but also information of unselected variables to represent all the variables. Robert and Escoufier[7] stated that their idea can be used to select a subset of variables, but no example was shown to illustrate their approach. However, Tanaka and Mori[9] and Mori, Tanaka and Tarumi[10] focused on extracting such PCs and also discussed how to select a reasonable subset of variables with some numerical examples. They called this type of PCA as the modified PCA (M. PCA) to discriminate it from the ordinary PCA.

We have planed to develop a statistical software VASPCA (VAriable Selection in Principal Component Analysis) which selects a subset of variables automatically using the above ideas for variable selection. In the present paper we show the first version of this software in which selection procedures using the idea of M.PCA have been programmed. The general formulation of M.PCA will be shown briefly in the next section. Four stepwise selection procedures and their practical actions in applying them to a real data set will be indicated in section 3 and 4, respectively. Concluding remarks will be summarized in the final section.

2 Modified PCA

Suppose we have obtained a data matrix Y which consists of n observations and p variables. If an original data set has qualitative variables, Y denotes its quantified matrix obtained by appropriate quantification. In M.PCA this Y is represented by r PCs as well as possible, where r is preassigned and the PCs are linear combinations of a subset of variables of Y. To derive such PCs, as discussed by Tanaka and Mori[9], PCA of instrumental variables proposed by Rao[11] is utilized by assigning the subset of variables as instrumental variables. The formulation is as follows.

We wish to make r linear combinations $Z = Y_1A$ which reproduce the original p variables as well as possible, where Y_1 is a subset of Y with q $(1 < q < p)$ variables and A is a $q \times r$ $(1 < r < q)$ coefficient matrix for the q variables. Here $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r)$ is determined in such a way that $\boldsymbol{y}$ can be predicted as well as possible by means of linear functions of $\boldsymbol{z}$. Thus the predictive efficiency is maximized for $\boldsymbol{y}$ by using a linear predictor in terms of $\boldsymbol{z}$.

Let the covariance matrix of $Y = (Y_1, Y_2)$ be $S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$. The residual covariance matrix of $\boldsymbol{y}$ after subtracting the best linear predictor is expressed as

$$S_{res} = S - S_1'A(A'S_{11}A)^{-1}A'S_1 = S - S_{Reg},$$

where $S_1 = (S_{11}, S_{12})$. Then the problem becomes to maximize S_{Reg}. If it is formulated as the maximization problem of $tr(S_{Reg})$ among other possibilities, a generalized eigenvalue problem

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

is obtained. Let the q eigenvalues of (1) be ordered from the largest to the smallest as $\lambda_1, \lambda_2,\ldots,\lambda_q$ and the associated eigenvectors be denoted by $\boldsymbol{a}_1, \boldsymbol{a}_2,\ldots,\boldsymbol{a}_q$. Then, the solution is expressed as $A=(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_r)$ and the maximized value of the criterion $tr(S_{Reg})$ is given by

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

or the proportion of the original variations explained by the r PCs is given by

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

Here we use this P as a criterion for maximization since it is easy to interpret.

When we apply the above method to standardized data rather than raw data, that is, the covariance matrices in the above formulation are replaced by the corresponding correlation matrices, the proportion P indicates the average squared multiple correlation between each of the original variables and the r PCs.

3 Variable Selection Procedures in Modified PCA

Suppose we want to obtain the best subset of variables consisting of q variables. In the meaning of M. PCA we can find the best subset by searching for one which has the largest P value among all possible subsets of size q. It provides the best PCs which represent all the variables very well. Though it is the best way to compute P for all the possible subsets, it is usually impractical due to high computing cost. Therefore, as practical strategies we propose the following two-stage procedures of variable selection, Backward elimination, Forward selection, Backward-forward stepwise selection and Forward-backward stepwise selection. These procedures are to remove or add only one variable at each selection step, but it has been made clear by Mori, Tarumi and Tanaka[12] that P value based on the subset of variables selected by these procedures are not different so much from the largest P which is provided using the best subset of variables among all possible combinations.

Backward elimination

Stage A. Initial fixed-variable stage

A-1 Assign q variables to subset Y_1, usually $q:=p$.

A-2 Solve the eigenvalue problem (EVP) (1).

A-3 Looking carefully at the eigenvalues and the proportions P, determine the number r of PCs to be used.

A-4 Specify kernel variables which are always in Y_1, if necessary. The number of them is less than q.

Stage B. Variable selection stage (backward)

B-1 Removing one of the q variables in Y_1, make a temporary subset of size $q-1$, and obtain the proportion P by solving the EVP (1). Repeat this for each variable in Y_1, then obtain q Ps. Find the best subset of size $q-1$ which provides

the largest P among the q Ps and remove the corresponding variable from the present subset of Y_1. Put $q:=q-1$.

B-2 If both P and q are larger than preassigned values, go back to B-1. Otherwise stop.

Backward-forward stepwise selection

Stage A. Initial fixed-variable stage

A-1 to 4 Same as A-1 to 4 in Backward elimination.

Stage B. Variable selection stage (backward-forward)

B-1 Put $i:=1$.

B-2 Removing one of the q variables in Y_1, make a temporary subset of size $q-1$, and obtain the proportion P by solving the EVP (1). Repeat this for each variable in Y_1, then obtain q Ps. Find the best subset of size $q-1$ which provides the largest P (denoted by P_i) among the q Ps and remove the corresponding variable from the present subset of Y_1. Put $q:=q-1$.

B-3 If both P and q are larger than preassigned values, go to B-4. Otherwise stop.

B-4 Removing one of the q variables in Y_1, make a temporary subset of size $q-1$, and obtain the proportion P by solving the EVP (1). Repeat this for each variable in Y_1, then obtain q Ps. Find the best subset of size $q-1$ which provides the largest P (denoted by P_{i+1}) among the q Ps and remove the corresponding variable from the present subset of Y_1. Put $q:=q-1$.

B-5 Adding one of the $p-q$ variable in Y_2 to Y_1, make a temporary subset of size $q+1$ and obtain the proportion P by solving the EVP (1). Repeat this for each variable in Y_2 except for the variable removed in B-4, then obtain $p-q-1$ Ps. Find the best subset of size $q+1$ which provides the largest P (denoted by P_{temp}) among the $p-q-1$ Ps.

B-6 If $P_i < P_{temp}$, add the variable found in B-5 to Y_1, put $P_i:=P_{temp}$, $q:=q+1$ and $i:=i-1$, and go back to B-5. Otherwise put $i:=i+1$ and go back to B-4.

Forward selection

Stage A. Initial fixed-variable stage

A-1 to 3 Same as A-1 to 3 in Backward elimination.

A-4 Here $q(q \geqq r)$ is redefined as the number of kernel variables. Assign q variables to subset Y_1. If there is no specified subset of variables to be assigned to Y_1, putting $q:=r$, the q variables which provide the largest P among all possible subsets of size q are assigned.

Stage B. Variable selection stage (forward)

B-1 Adding one of the $p-q$ variables in Y_2 to Y_1, make a temporary subset of size $q+1$ and obtain the proportion P by solving the EVP (1). Repeat this for each variable in Y_2, then obtain $p-q$ Ps. Find the best subset of size $q+1$ which provides the largest P among the $p-q$ Ps and add the corresponding variable to the present subset of Y_1. Put $q:=q+1$.

B-2 If both P and q are smaller than preassigned values, go back to B-1. Otherwise stop.

Forward-backward stepwise selection

Stage A. Initial fixed-variable stage

A-1 to 4 Same as A-1 to 4 in Forward selection.

Stage B. Variable selection stage (forward-backward)

B-1 Put $i := 1$.

B-2 Adding one of the $p-q$ variables in Y_2 to Y_1, make a temporary subset of size $q+1$ and obtain the proportion P by solving the EVP (1). Repeat this for each variable in Y_2, then obtain $p-q$ Ps. Find the best subset of size $q+1$ which provides the largest P (denoted by P_i) among the $p-q$ Ps and add the corresponding variable to the present subset of Y_1. Put $q := q+1$.

B-3 If both P and q are smaller than preassigned values, go to B-4. Otherwise stop.

B-4 Adding one of the $p-q$ variables in Y_2 to Y_1, make a temporary subset of size $q+1$ and obtain the proportion P by solving the EVP (1). Repeat this for each variable in Y_2, then obtain $p-q$ Ps. Find the best subset of size $q+1$ which provides the largest P (denoted by P_{i+1}) among the $p-q$ Ps and add the corresponding variable to the present subset of Y_1. Put $q := q+1$.

B-5 Removing one of the q variables in Y_1, make a temporary subset of size $q-1$ and obtain the proportion P by solving the EVP (1). Repeat this for each variable in Y_1 except for the variable added in B-4, then obtain $q-1$ Ps. Find the best subset of size $q-1$ which provides the largest P (denoted by P_{temp}) among the $q-1$ Ps.

B-6 If $P_i < P_{temp}$, remove the variable found in B-5 from Y_1, put $P_i := P_{temp}$, $q := q-1$ and $i := i-1$, and go back to B-5. Otherwise put $i := i+1$ and go back to B-4.

4 Statistical Software VASPCA

Statistical software VASPCA has been developed in Microsoft Visual Basic to perform variable selection in PCA. This version can select a reasonable subset of variables using the idea of M.PCA. The flow is as follows (Fig. 1).

1) Data entry and preliminary analysis

1-1) The users open a data file or input data on a spreadsheet.

1-2) On the users' demand VASPCA can compute basic statistics. If the data has qualitative variables and the users want to quantify them before selection, VASPCA can perform quantification with the alternating least square method (Young et.al[13]). As for M.PCA for qualitative data, see Mori, Tanaka and Tarumi[10]).

2) First stage (Initial fixed-variable stage)

2-1) The users select one of variable selection methods. Here is only Modified PCA.

2-2) The users determine which type of matrix to be used, covariance or correlation.

2-3) If necessary, the users assign q variables to Y_1 (that is, $p-q$ ones to Y_2), and specify kernel variables in Y_1.

2-4) Based on the above conditions VASPCA solves the EVP (1) of $Y = (Y_1, Y_2)$ and

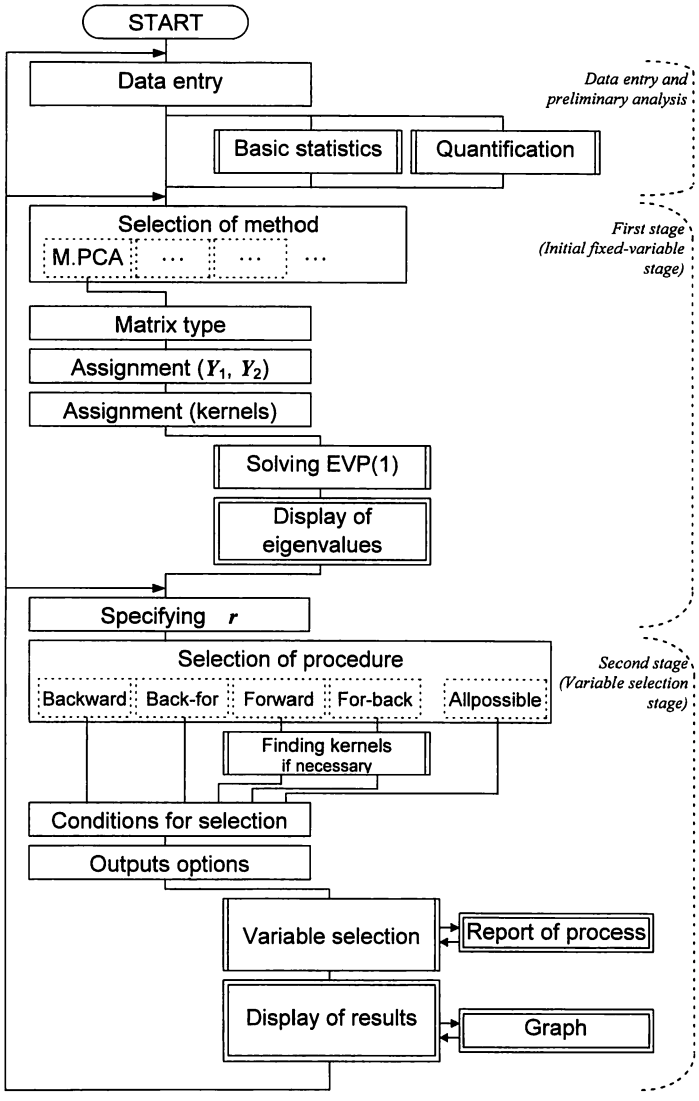

Fig. 1 Flow of variable selection based on M. PCA in VASPCA.

indicates the eigenvalues with their scree graph. Looking at these outputs the users specify the number r of PCs to be used.

3) Second stage (Variable selection stage)

3-1) The users select a variable selection procedure from "Backward", "Backward-forward", "Forward" or "Forward-backward". Though they can select "All possible" procedure optionally, an alert message to high computational costs is indicated when it is selected.

3-2) The users input selection conditions, i.e., preassigned values of q and P just like as a stopping criterion.

3-3) The users specify options on "Basic outputs" and "Additional outputs". As for

"Basic outputs", VASPCA outputs the number of variables in Y_1, P value, selected variable labels and removed ones as default, and P values explained by all PCs as option. As for "Additional outputs", it can output P values, eigenvalues, PC scores and correlation loadings in every computation step. All the outputs can be saved in a file. The users can also determine whether the report of selection process is displayed or not.

3-4) Based on the above conditions VASPCA executes the selection procedure. In computation it reports the selection process on a form when "Report of selection process" is selected in 3-3. Satisfying the stopping criteria, it displays a summary of results and additional outputs. The users can draw an index plot of P (or P based on all PCs) in the summary table. On the graph, the number of variables in Y_1, P values, selected variable labels and removed ones can be indicated at every point interactively. Looking at these outputs, the users specify the number of variables and which subset of variables should be used. After that the user can go back to 2-1) or 3-1) and retry to find a reasonable subset of variables.

Here we illustrate the practical actions of VASPCA by applying it to the data gathered for the purpose of making a rating scale to measure the seriousness of mild disturbance of consciousness (MDOC) due to head injury and other lesions (Sano et al.[14) 15)]). The data set consists of 87 individuals and 25 qualitative variables (test items, four points scale). According to the previous studies[14) 15)] VASPCA is applied to a 23 variables-2 PCs model. All the figures indicating actions of VASPCA are obtained from English version, but we can also use Japanese version.

All the variables of this data are qualitative, then we quantified them by applying 'Quantification' in "Data" menu before proceeding to the first stage.

In the first stage we selected 'Modified PCA' in "Method" menu (Fig. 2; The

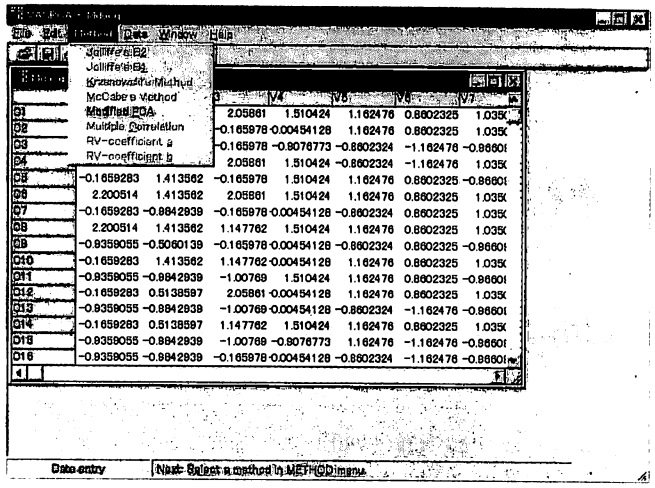

Fig. 2 Data entry and selecting method (All the veriables have been quantified).

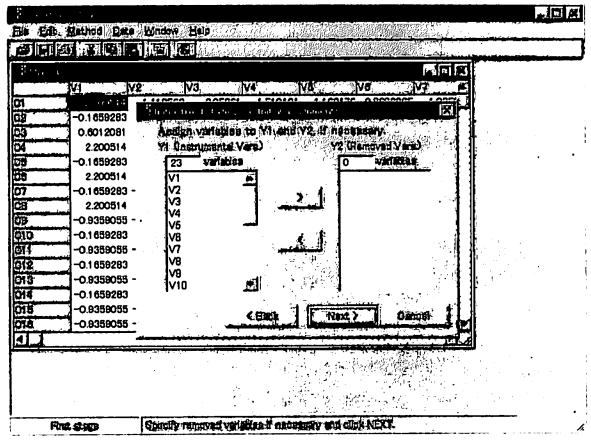

Fig. 3 Assigning variables to Y_1 and Y_2.

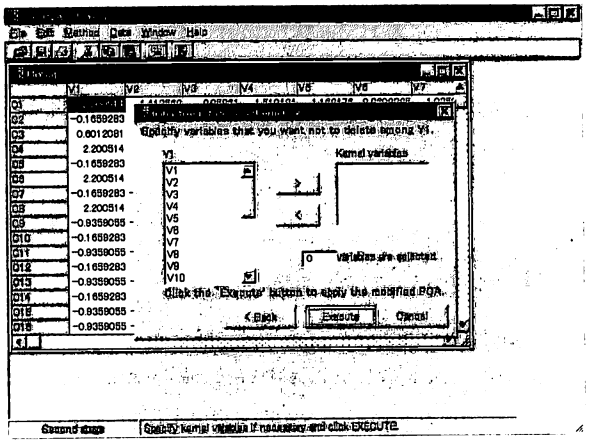

Fig. 4 Specifying kernel variables.

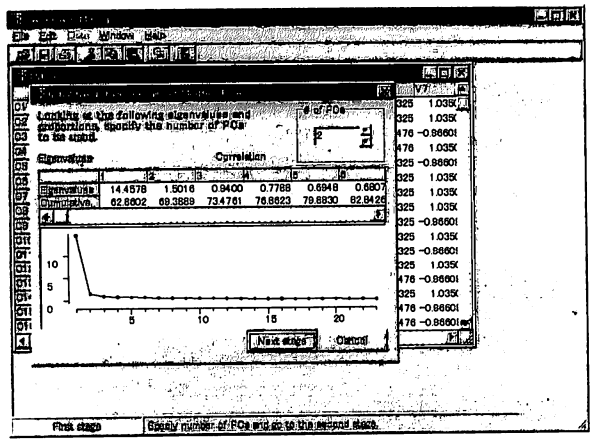

Fig. 5 Results of the initial stage. VASPCA displays eigenvalues with their scree graph obtained by solving EVP (1) with the variable pattern set in Fig. 3 and 4. The users click 'Execute' button after specifying the number of PCs.

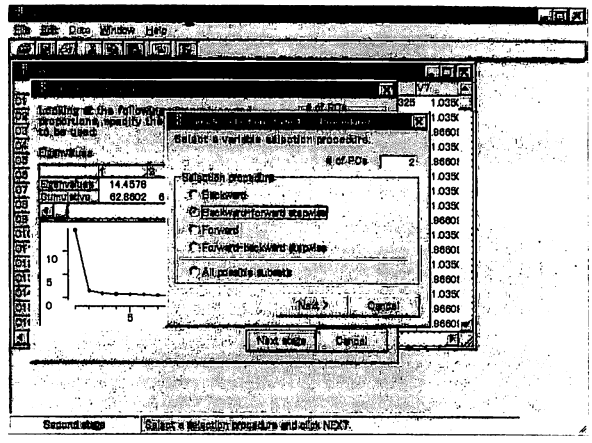

Fig. 6 Selecting one slection procedure among four ones and one option.

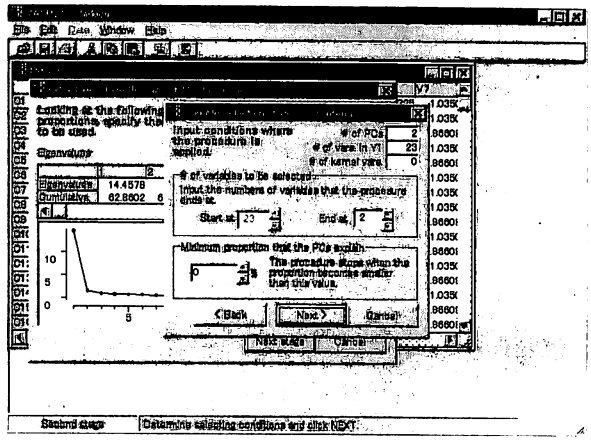

Fig. 7 Inputting conditions for selection.

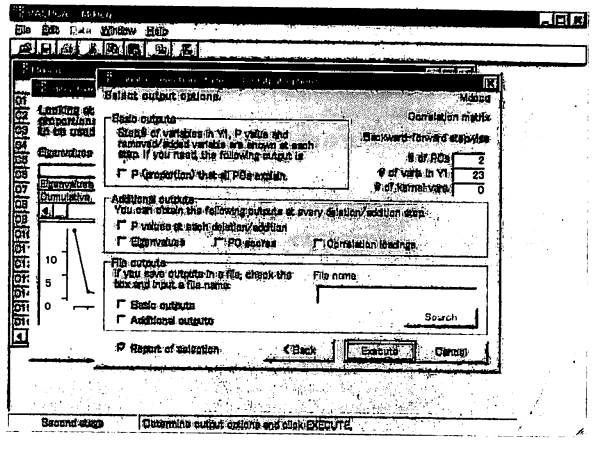

Fig. 8 Specifying output options.

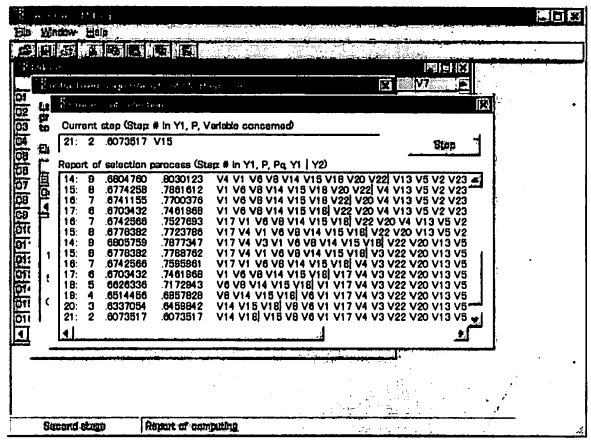

Fig. 9 Report of selecting process. The number of step, q, P and selected/removed variable labels are indicated at every selection.

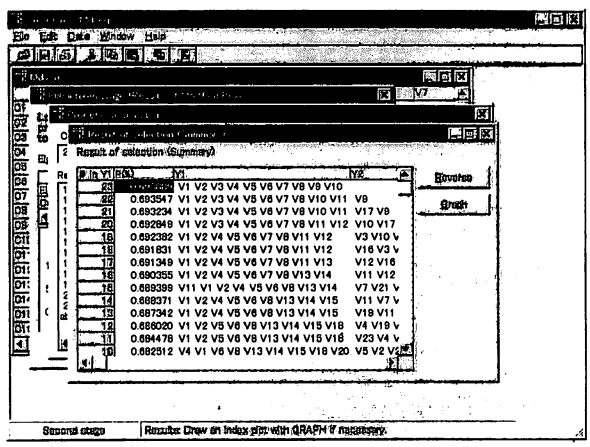

Fig. 10 Results of the variable selection stage. The table illustrates the summary of results which is a list of P, selected subset of variables and unselected one at each q.

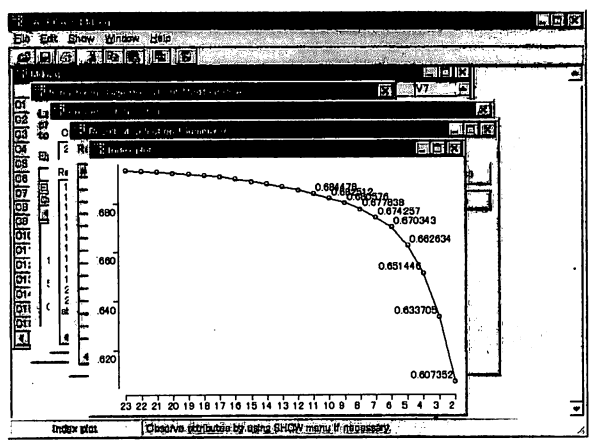

Fig. 11 An index plot of P value. The number q, P value, selected variable labels and unselected ones can be indicated at any point highlighted by users' mouse instruction.

variables have been already quantified). Selecting 'correlation matrix' and assigning no variables to Y_1 and to kernels (Fig. 3 and 4), we clicked 'Execute' button to obtain the results of the first stage (Fig. 5). Looking at Fig. 5 and following the previous studied we specified two PCs and clicked at 'Next stage' button.

In the second staged, we selected 'Backward-forward' procedure (Fig. 6) and inputted no specified condition for stopping criteria (Fig. 7) and outputs (Fig. 8). This means that selection procedure starts when the number of variables is 23 and continues until it becomes 2. Clicking 'Execute' button, VASPCA started stepwise selection and displayed the process of selection (Fig. 9). When the selection was done, VASPCA indicated a summary of selection (Fig. 10). We drew an index plot of P and observed P values at some points highlighted by mouse (Fig. 11). Based on these results we can make the decisions, for example, such that the 11 or 10 selected variables instead of all the 23 ones can be used to extract PCs as a two-dimensional scale in the future investigations because the loss of information is almost negligible by removing 12 or 13 variables among 23.

5 Concluding Remarks

Statistical software VASPCA has been developed to select a reasonable subset of variables using a variety of selection ideas in PCA. By programming these ideas in one package it becomes possible to apply variable selection in PCA easily in the real situation. In the present paper selection procedures based on M.PCA are illustrated with the practical actions and results when they are applied to MDOC data..

Here we shall show some considerations to extend this software as future problems.

(1) Since this version deals with only selection procedures based on the M.PCA, other selection methods mentioned in section 1 should be included in VASPCA.

(2) To improve usefulness of the software, it is desirable to provide some easy and visual functions in VASPCA such as to compare PC scores, correlation loadings and other information computed using the selected subset of variables with those using other subset. Furthermore optional instructions for beginners and rich helps are convenient.

(3) It is necessary to propose a strategy to determine the reasonable number q of variables to be used.

(4) It is also important to suggest to the users which procedure is suitable for each of several problems in the practical situation.

References

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

2) Jolliffe, I. T. (1973): Discarding variables in a principal component analysis. II. Real data. *Appqied Statistics*, **22**, 21-31.

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

4) McCabe, G. P. (1984): Principal Variables. *Technometrics*, **26**, 137-144.

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

6) Krzanowski, W. J. (1987b): Cross—validation in principal component analysis. *Biometrics,* **43**, 575-584.

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

8) Mori, Y., Tarumi T. and Tanaka, Y. (1994): Variable selection with RV-coefficient in principal component analysis. *J. Jap. computational Statistics,* **7**, 47-56 (in Japanese).

9) Tanaka, Y. and Mori, Y. (1997): Principal component analysis based on a subset of variables: Variable selection and sensitivity analysis. *American Journal of Mathematical and Management* Sciences, Special Issue, **17**, 61-89.

10) Mori, Y., Tanaka, Y. and Tarumi, T. (1997): Principal component analysis based on a subset of variables for qualitative data. *Proceedings of IFCS-96,* Springer-Verlag, 547-554.

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

12) Mori, Y., Tarumi, T. and Tanaka, Y. (1996): Principal component analysis based on a subset of variables (2). *The proceedings of the 10th conference of the Japanese Society of Computational Statistics,* 47-56 (in Japanese).

13) Young, F., Yakane, Y. and De Leeuw, J. (1978): The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika,* **43**, 279-281.

14) Sano, K., Manaka, S., Kitamura, K., Kagawa, M., Takeuchi, K., Ogashiwa, M., Kameyama, M., Tohgi, H. and Yamada, H. (1977): Statistical studies on evaluation of mind disturbance of consciousness: Abstraction of characteristic clinical pictures by cross-sectional investigation. *Sinkei Kenkyu no Shinpo,* **21**, 1052-1065 (in Japanese).

15) Sano, K., Manaka, S., Kitamura, K., Kagawa, M., Takeuchi, K., Ogashiwa, M., Kameyama, M., Tohgi, H. and Yamada, H. et al. (1983): Statistical studies on evaluation of mind disturbance of consciousness. *Journal of Neurosurg,* **58**, 223-230.