VARIABLE SELECTION IN PRINCIPAL COMPONENT ANALYSIS AND ITS APPLICATIONS

MARCH 1995

Yuichi MORI

The Graduate School of Natural Science and Technology
(Doctor Course)
OKAYAMA UNIVERSITY

(1)

VARIABLE SELECTION IN PRINCIPAL COMPONENT ANALYSIS AND ITS APPLICATIONS

MARCH 1995

Yuichi MORI

The Graduate School of Natural Science and Technology
(Doctor Course)
OKAYAMA UNIVERSITY

Contents

1 Introduction

Principal component analysis (PCA) is a statistical method which reduces the dimensionality of the space using appropriate components. In general, each component is a linear combination of all of the original variables, but this is sometimes regarded as a deficiency of this approach. That is, all the original variables are still needed to define new components or variables. It is also 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 (see, e.g., McCabe, 1984).

Actually, we often meet the problem of selecting variables in many practical situations. Suppose we wish to apply PCA or factor analysis (FA) to make a small dimensional rating scale which measures latent traits. From the validity aspect, in order to gather important dimensions well, the items or variables should include all possible ones. On the other hand, from the aspect of practical application, the number of variables should be as small as possible not only because of waste of time and resources but also because of difficult interpretation of components extracted from too many variables. It often happens that investigators measure more variables than strictly necessary on each sample individual. Hence it is essential to reduce the number of variables as well as possible without disturbing the sample features .

In such a case, while analysts have tried to reduce the number of variables subjectively by applying correlation analysis or cluster analysis of variables, it has been desirable to develop an appropriate procedure to select variables automatically. Since procedures for selecting variables in multiple regression or discriminant analysis cannot be used directly under this circumstance, it is necessary to propose variable selection methods in multivariate analysis without response variables, i.e., PCA, FA and so on.

The problem of variable selection in the multivariate analysis without response variables has been studied by some authors. Variable selection methods in PCA have discussed by Jolliffe (1972, 1973), Robert and Escoufier (1976), McCabe (1984) and Krzanowski (1987a, b) among others. Xia and Yang (1988) have derived some criteria and procedures of variable selection in Hayashi's third method of quantification. Works on variable selection in FA have been proposed by Tanaka and Kodake (1981) and Tanaka (1983).

This thesis consists of two main parts. The first part is discussed backward elimination procedures for variable selection using Escoufier's RV-coefficient and the so-called perturbation theory as mathematical tools, comparing with the above authors' methods. The procedures are proposed in PCA and Hayashi's third method of quantification. We focus on the behavior of the principal component (PC) score matrix and the sample score matrix in PCA and Hayashi's third method of quantification, respectively, when a variable is discarded. In the second part, the generalized PCA is proposed as an applied version of variable selection. It extracts the generalized principal components (PCs) which are computed using only a selected subset of variables but represent all the original variables. The selection procedure and such PCs are discussed. In this part, sensitivity analysis of individuals and variables are also applied to observe the influence of them when such PCs are found by discarding variables.

In chapter 2, as a preliminary a brief review is presented on some of studies about variable selection in PCA and some of the mathematical tools and concepts that will be useful for the study of variable selection. It includes a number of variable selection methods in PCA studied by Jolliffe (1972, 1973), McCabe (1984) and Krzanowski (1987a, b). The idea of variable selection presented by Robert and Escoufier (1976) is also summarized. Mathematical tools are "RV-coefficient (Robert and Escoufier, 1976)", the so-called "perturbation theory" which includes influence functions and perturbation theory both of ordinary and generalized eigenvalue problems, and Rao(1964)'s PCs of instrumental variables.

In chapter 3, a backward procedure of variable selection in PCA is proposed in which we discard a variable which has the closest configuration of the PC score matrix among the existing variables successively. This means that the variable selection methods select a set of variables reproducing as closely as possible the general features of the complete data. In our study, RV-coefficient is used to evaluate the closeness between the configuration of PC score matrix before discarding a variable and that after discarding. The perturbation theory of eigenvalue problems as well as the exact method are also utilized in computation. To evaluate our method it is compared with Jolliffe's and McCabe's methods, and with biplot and cluster analysis of variables. As numerical examples, we apply our method to "Crime rates data (Ahamad, 1697)" which was analyzed by both Jolliffe and McCabe, to the artificial data sets generated by Jolliffe (1972), and to "Automobile data (Becker, et al., 1988)". In this numerical study three more procedures are applied to evaluate the goodness of successive way and usage of perturbation in our procedure.

In chapter 4, since Hayashi's third method of quantification can be thought the categorical version of PCA, a similar procedure to variable selection in PCA proposed in chapter 3 is applied to Hayashi's third method of quantification. Backward procedures of variable selection are proposed in which we discard a variable which has the smallest effect on the sample score matrix among the existing variables successively. In the procedures we use the RV-coefficient and the perturbation theory of eigenvalue problems as well as the exact method in computation. The procedures deal with the following two typical problems on categorical data and its variable selection: categorical data has two data forms, free-choice and item-category forms, which have the same information but lead to different results in Hayashi's third method of quantification; there are some cases where we cannot continue to compute because some row sums in the denominator get 0 when a variable is discarded. As solutions for the problems we propose two procedures which treat both two data forms and introduce perturbation to the data matrix instead of discarding variables exactly. We evaluate these methods by analyzing two real data sets, "Spirits data (Arima and Ishimura, 1987)" and "Fatigue data (Maehashi et al., 1993)".

In the last chapter 5, we discuss PCs which are computed using only a selected subset of variables but represent all the variables including those not selected. To find such PCs we borrows the ideas of Rao(1964)'s PCA of instrumental variables and Robert and Escoufier(1976)'s approach based on RV-coefficient. This is called the generalized PCA. In the meaning of variable selection, the method finds specified variables which represent all the original variables as well as possible. Furthermore, when such PCs are found, we propose a method of sensitivity analysis by deriving influence functions related with the generalized PCA. We also discuss the influence of variables to the results of analysis. To evaluate the proposed methods we analyze two data sets, "Alate adelges data (Jeffers, 1967)" and "Mild disturbance of consciousness (MDOC) data".

2 Preliminary Foundations

In this chapter, a brief review is presented on some of preliminary foundations. First the variable selection methods in principal component analysis (PCA) will be shown, focusing those proposed by Jolliffe, McCabe, Krzanowski, Robert and Escoufier among others. Next the mathematical tools and concepts will be presented, which are useful to study variable selection in the later chapters. They contain Escoufier's RV-coefficient, the so-called perturbation theory and Rao's principal components (PCs) of instrumental variables.

2.1 Overview of variable selection in principal component analysis

The problems of variable selection in multivariate analysis without response variables have been studies by some authors. Jolliffe (1972, 1973, 1986), McCabe (1984) and Krzanowski (1987a, 1987b) studied variable selection in PCA. Robert and Escoufier (1976) also discussed the possibility of variables selection in PCA but presented no example. In the other analysis without response variables, variable selection procedures have been proposed by Tanaka and Kodake (1981) and Tanaka (1983) in factor analysis, and Xia and Yang (1988) in Hayashi's third method of quantification. Here, as an overview of these studies, we will review the first three authors' methods in this section, while the possibility of variable selection presented by fourth authors will be summarized briefly in the last section.

Suppose that X is an observation data matrix which has p variables observed on each n individuals. We would now like to select q $(q < p)$ variables among the original p variables.

Jolliffe (1972, 1973) discussed a number of variable selection methods based on multiple correlation coefficients, PCA and cluster analysis of variables. His concept is to select a subset of variables which preserve most of the variation in X. He examined three main types of method using PCs and concluded that the following two methods, which are called B2 and B4, were satisfactory:

B2 Associate one variable with each of the last $p-q$ PCs and delete those $p-q$ variables. The reasoning behind this method is that small eigenvalues correspond to near-constant relationships between a subset of variables. If one of the variables involved in such a relationship is deleted, little information is lost. A sensible choice for deletion is the variable with the highest coefficient in absolute value in the relevant PC. An iterative version can be considered;

B4 Associate one variable with each of the first q PCs, namely the variable not already chosen, with the highest coefficient in absolute value in each successive PC. These q variables are retained, and the remaining $p-q$ are deleted.

Then he applied his proposed methods, including B2 and B4, to simulated data (1972) and various real data sets (1973) to evaluate them. Through these examinations, he found that none of them was informally best, but several of them selected reasonable subsets in most cases.

McCabe (1984) started from the fact that PCs satisfy a number of different optimality criteria. His approach is based on the aim to select a subset of variables that contain, in some sense, as much information as possible. A subset of the original variables which optimizes one of these criteria is called *principal variables.* To find the *principal variables*, he considered 12 criteria which lead to one of four criteria

$$\text{Minimize} \quad \prod_{j=1}^{p-q} \phi_j; \tag{2.1a}$$

$$\text{Minimize} \quad \sum_{j=1}^{p-q} \phi_j; \tag{2.1b}$$

$$\text{Minimize} \quad \sum_{j=1}^{p-q} \phi_j^2; \tag{2.1c}$$

$$\text{Maximize} \quad \sum_{j=1}^{q^-} \rho_j^2; \tag{2.1d}$$

where ϕ_j, $j = 1, 2, \ldots, p-q$ are the eigenvalues of the conditional covariance (or correlation) matrix of the $p-q$ deleted variables, given the value of the q selected variables, and ρ_j, $j = 1, 2, \ldots, q^-$, $q^- = \min(q, p-q)$ are the canonical correlations between the set of $p-q$ deleted variables and the set of q selected variables. Then he argued that the first criterion is computationally feasible to explore all possible subsets and the second one can be used to define a stepwise procedure, although the other two criteria were not explored further in his paper. He also stated that applying the PCs optimality criteria to the variable selection problem dose not lead to a unique solution.

Krzanowski (1987a) proposed another selection method in which a selected subset of variables conveys the main features of the whole samples. As a reason for proposing his method he pointed out that the methods currently available for selecting variables in PCA, namely Jolliffe's and McCabe's methods, may not lead to an appropriate subset. His method, based on Procrustes Analysis, is as follows: Suppose that X is an $n \times p$ data matrix and the essential dimensionality of the data is r. Let Y be the $n \times r$ transformed data matrix of PC scores, yielding the best r-dimensional approximation to the original data configuration X. When we want to select q of the original p variables, they should be hoped recovering the true structure. Denote the $n \times q$ data matrix which retains only q variables selected from and the $n \times r$ matrix of PC scores of these reduced data by $\widetilde{X}$ and $\tilde{Z}$, respectively. $\tilde{Z}$ is therefore the best r-dimensional approximation to the original data configuration $\widetilde{X}$. To measure the discrepancy between Y and $\tilde{Z}$, Procrustes Analysis is conducted. This analysis yields the sum of squared differences between the two configurations as

$$M^2 = tr(YY' + \tilde{Z}\tilde{Z}' - 2D_\alpha) \tag{2.2}$$

where $tr(\cdot)$ denotes a trace of the matrix $(\cdot)$, $D_\alpha = diag(\alpha_1, \ldots, \alpha_r)$, α_j are singular values of $\tilde{Z}'Y$, and both Y and $\tilde{Z}$ are centered. The best subset of q variables will be that subset which yields the smallest value of M^2 among all q-variable subsets. He proposed a backward elimination based on this criterion and found that his method lead to a better subset than the other authors'.

2.2 RV-coefficient

Robert and Escoufier (1976) has derived a measure of similarity of the two configurations, taking into account the possibly distinct metrics to be used on them to measure the distances between points. The measure is called RV-coefficient.

Consider a given sample of n individuals on which two sets of observations, an $n \times p$ data matrix X and an $n \times q$ data matrix Y. Denote the centered matrices corresponding to X and Y by $\widetilde{X}$ and $\widetilde{Y}$, respectively. Let $C(X)$ and $C(Y)$ be the two associated configurations, in $\mathcal{R}^p$ and $\mathcal{R}^q$, respectively. As a measure of the *relative* positions of points in a configuration, say $C(X)$, the matrix $\widetilde{X}\widetilde{X}'/\{tr(\widetilde{X}\widetilde{X}')^2\}^{1/2}$ is used. This matrix is translation and rotation independent and the scalar denominator $\{tr(\widetilde{X}\widetilde{X}')^2\}^{1/2}$ ensures that it is also independent of global changes of scale. The distance between the configurations $C(X)$ and $C(Y)$ is therefore measured by

$$dist\,\{C(X), C(Y)\} \;=\; \left\| \frac{\widetilde{X}\widetilde{X}'}{\left\{tr(\widetilde{X}\widetilde{X}')^2\right\}^{1/2}} - \frac{\widetilde{Y}\widetilde{Y}'}{\left\{tr(\widetilde{Y}\widetilde{Y}')^2\right\}^{1/2}} \right\|$$

$$
\begin{aligned}
&= \left[2 \left\{ 1 - \frac{tr(\widetilde{X}\widetilde{X}'\widetilde{Y}\widetilde{Y}')}{\left\{ tr(\widetilde{X}\widetilde{X}')^2 \cdot tr(\widetilde{Y}\widetilde{Y}')^2 \right\}^{1/2}} \right\} \right]^{1/2} \\
&= \left[2 \left\{ 1 - RV(X,Y) \right\} \right]^{1/2}, \qquad (2.3)
\end{aligned}
$$

where $||\cdot||$ indicates L_2 or Euclidean norm, especially $||\widetilde{X}\widetilde{X}'/\{tr(\widetilde{X}\widetilde{X}')^2\}^{1/2}|| = 1$. Thus

$$
RV(X,Y) = \frac{tr(\widetilde{X}\widetilde{X}'\widetilde{Y}\widetilde{Y}')}{\left\{ tr(\widetilde{X}\widetilde{X}')^2 \cdot tr(\widetilde{Y}\widetilde{Y}')^2 \right\}^{1/2}}. \qquad (2.4)
$$

The coefficient $RV(X,Y)$ can be used as the actual measure of closeness of $C(X)$ and $C(Y)$. The value of $RV(X,Y)$ is in the closed interval $[0,1]$ and the closer to 1 it is, the closer the patterns are. When $p = q = 1$, $RV(X,Y)$ is equal to the squared ordinary correlation coefficient.

2.3 Perturbation theory

2.3.1 Influence functions

As a basic tool or concept to evaluate the influence of individuals or variables in the data matrix $X(n \times p)$, we can make use of the notion of influence function proposed by Hampel (1974). We shall show the case where we observe the influence of individuals. In influence function a perturbation is introduced to the cumulative distribution function (cdf) F in such a way that F is changed to

$$
F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_x \qquad (2.5)
$$

where δ_x is the cdf with a unit point mass at $\boldsymbol{x}$. The theoretical influence function (TIF) is defined for a quantity θ which is expressed as a functional of the cdf as

$$
I(\boldsymbol{x};\theta) = \lim_{\varepsilon \to 0} \frac{\theta((1-\varepsilon)F + \varepsilon\delta_x) - \theta(F)}{\varepsilon}. \qquad (2.6)
$$

Consider the case where $\theta((1-\varepsilon)F + \varepsilon\delta_x) = \theta(\varepsilon)$ is expanded to the Taylor series as

$$
\theta(\varepsilon) = \theta(0) + \varepsilon\theta^{(1)}(0) + (\varepsilon^2/2)\theta^{(2)}(0) + O(\varepsilon^3), \qquad (2.7)
$$

in the neighborhood of $\varepsilon = 0$. Then, the TIF is obtained as the coefficient $\theta^{(1)}$ of the first order term of ε in the power series (2.7) or simply defined as the first order differential coefficient of $\theta(\varepsilon)$ at $\varepsilon = 0$.

The above (2.6) is the definition of influence function based on the population distribution function. As sample versions two kinds are often used. One is the empirical influence function (EIF), which is obtained by replacing the empirical cdf $\widehat{F}$ for F in the definition of the TIF. Of particular interest are the values at $\boldsymbol{x} = \boldsymbol{x}_i$ $(i = 1, \ldots, n)$ given by

$$\widehat{I}(\boldsymbol{x}_i; \widehat{\theta}) = \lim_{\varepsilon \to 0} \frac{\theta((1-\varepsilon)\widehat{F} + \varepsilon\delta_{x_i}) - \theta(\widehat{F})}{\varepsilon}. \tag{2.8}$$

The other is the sample influence function (SIF), which is obtained by omitting "lim" and putting $\varepsilon = -1/(n-1)$ in (2.8), i.e.,

$$\widetilde{I}(\boldsymbol{x}_i; \widehat{\theta}) = -(n-1)(\widehat{\theta}_{(i)} - \widehat{\theta}), \tag{2.9}$$

where the subscript (i) indicates the omission of the i-th individual.

Influence function discussed so far is useful to evaluate the influence of a single observation. To deal with the influence of multiple individuals it is convenient to consider the perturbation from F to $F_\varepsilon = (1-\varepsilon)F + \varepsilon G$, where $G = k^{-1} \sum \delta_{x_i}$, the summation being taken for a subset of k individuals $\{\boldsymbol{x}_i\}$, and define a generalized influence function for this subset of individuals as the differential coefficient of $\theta(F_\varepsilon)$ with respect to ε at $\varepsilon = 0$. Then, it can be verified easily that this generalized influence function is equal to the average of the ordinary influence functions for the individuals belonging to this subset. This property suggests that a subset of individuals whose EIF vectors have similar directions and large lengths may compose an influential subset and that PCA or canonical variate analysis (PCA with metric $[\widehat{cov(\theta)}]^{-1}$) is useful for finding out such individuals. From the above property a general procedure based on EIF has been developed for sensitivity analysis of individuals to evaluate the influence of multiple as well as single individuals (see, Tanaka, Castaño-Tostado and Odaka, 1990; Tanaka, 1992).

The perturbation as (2.5) has the same meaning as the following change of weight on each row of data matrix:

$$w_\alpha = 1 \longrightarrow w_\alpha = \begin{cases} 1-\varepsilon & \alpha \notin S \\ 1+(n-1)\varepsilon & \alpha \in S \end{cases}, \tag{2.10}$$

where S is a specified set of variables.

On the other hand, we can use the above influence functions to observe the influence of variables as sensitivity analysis of variables, but in the meaning of variable selection it is often easier-to-interpret to introduce the perturbation as the weighting (2.10) replacing $(n-1)$ by $(p-1)$. Moreover we can also use the following weighting:

$$w_\alpha = 1 \longrightarrow w_\alpha = \begin{cases} 1 & \alpha \notin S \\ 1-\varepsilon & \alpha \in S \end{cases}. \tag{2.10'}$$

2.3.2 Perturbation theory in ordinary eigenvalue problems

Consider an ordinary eigenvalue problem

$$(H - \lambda_j I)\boldsymbol{v}_j = 0, \tag{2.11}$$

where H is a $p \times p$ real symmetric matrix, λ_j is the j-th eigenvalue and $\boldsymbol{v}_j$ is the associated eigenvector ($j = 1, \ldots, p$). Introducing some small perturbation in this eigenvalue problem as

$$H \longrightarrow H(\varepsilon) = H + \varepsilon H^{(1)} + (\varepsilon^2/2)H(2)(0) + O(\varepsilon^3), \tag{2.12}$$

the eigenvalues and eigenvectors can be expanded as a convergent power series in the neighborhood of $\varepsilon = 0$ as

$$\lambda_j(\varepsilon) = \lambda_j + \varepsilon\lambda_j^{(1)} + (\varepsilon^2/2)\lambda_j^{(2)} + O(\varepsilon^3), \tag{2.13}$$
$$\boldsymbol{v}_j(\varepsilon) = \boldsymbol{v}_j + \varepsilon\boldsymbol{v}_j^{(1)} + (\varepsilon^2/2)\boldsymbol{v}_j^{(2)} + O(\varepsilon^3), \tag{2.14}$$

from the perturbation theory of eigenvalue problems. If the eigenvalue of interest is simple, it is easy to obtain the coefficient of the first order term in the above expansions. Without loss of generality, we can assume that we are interested in the first q ($q < p$) eigenvalues and that they are all simple. Then we have the following formulas of the first differential:

$$\lambda_j^{(1)} = a_{jj}^{(1)}, \tag{2.15}$$
$$\boldsymbol{v}_j^{(1)} = \sum_{j \neq k}(\lambda_j - \lambda_k)^{-1} a_{kj}^{(1)} \boldsymbol{v}_k, \tag{2.16}$$

where

$$a_{kj}^{(1)} = \boldsymbol{v}_k' H^{(1)} \boldsymbol{v}_j. \tag{2.17}$$

Furthermore, the following two matrices, which are functions of eigenvalues and eigenvectors, contribute important roles in the formulation (Tanaka, 1988):

$$P = \sum_{j=1}^{q} \boldsymbol{v}_j \boldsymbol{v}_j', \tag{2.18}$$

$$T = \sum_{j=1}^{q} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j'. \tag{2.19}$$

Considering a small perturbation which corresponds to the perturbation (2.12) on H, these two quantities can be expanded as

$$P = P + \varepsilon P^{(1)} + (\varepsilon^2/2)P^{(2)} + O(\varepsilon^3), \tag{2.20}$$
$$T = T + \varepsilon T^{(1)} + (\varepsilon^2/2)T^{(2)}(0) + O(\varepsilon^3). \tag{2.21}$$

The coefficients $P^{(1)}$ and $T^{(1)}$ are obtained as

$$P^{(1)} = \sum_{j=1}^{q}\sum_{k=q+1}^{p}(\lambda_j-\lambda_k)^{-1}(\boldsymbol{v}_j'H^{(1)}\boldsymbol{v}_k)(\boldsymbol{v}_j\boldsymbol{v}_k'+\boldsymbol{v}_k\boldsymbol{v}_j'), \tag{2.22}$$

$$\begin{aligned} T^{(1)} &= \sum_{j=1}^{q}\sum_{k=1}^{q}(\boldsymbol{v}_j'H^{(1)}\boldsymbol{v}_k)\boldsymbol{v}_j\boldsymbol{v}_k' \\ &+ \sum_{j=1}^{q}\sum_{k=q+1}^{p}\lambda_j(\lambda_j-\lambda_k)^{-1}(\boldsymbol{v}_j'H^{(1)}\boldsymbol{v}_k)(\boldsymbol{v}_j\boldsymbol{v}_k'+\boldsymbol{v}_k\boldsymbol{v}_j'), \end{aligned} \tag{2.23}$$

in spite of the fact whether the eigenvalues of interest are all simple or not (Tanaka, 1988). See Castaño-Tostado and Tanaka (1990) and Tanaka (1992) about the details of the second differential coefficient.

2.3.3 Perturbation theory in generalized eigenvalue problems

Here we consider the following type of eigenvalue problem, namely a generalized eigenvalue problem

$$(A-\theta_j B)\boldsymbol{u}_j = 0, \tag{2.24}$$

where A is a $p\times p$ symmetric matrix, B is a $p\times p$ positive definite symmetric matrix and $\boldsymbol{u}_j$ is the eigenvector associated with the j-th largest eigenvalue θ_j normalized such that $\boldsymbol{u}_j'B\boldsymbol{u}_j = 1$ $(j=1,\ldots,p)$.

To derive influence functions related with eq.(2.24), the following lemma provides a useful tool.

Lemma (Tanaka, 1989) Suppose that A and B in (2.24) are functionals of the cdf and that they are twice continuously differentiable with respect to ε. Then, the influence functions or equivalently the differential coefficients with respect to ε at $\varepsilon=0$ are obtained as

$$I(\boldsymbol{x};\theta_j) = \boldsymbol{u}_j'(A^{(1)}-\theta_j B^{(1)})\boldsymbol{u}_j, \quad j=1,\ldots,p, \tag{2.25}$$

$$\begin{aligned} I(\boldsymbol{x};\boldsymbol{u}_j) &= \sum_{k\neq j}(\theta_j-\theta_k)^{-1}\{\boldsymbol{u}_j'(A^{(1)}-\theta_j B^{(1)})\boldsymbol{u}_k\}\boldsymbol{u}_k \\ &- (1/2)(\boldsymbol{u}_j'B^{(1)}\boldsymbol{u}_j)\boldsymbol{u}_j, \quad j=1,\ldots,p, \end{aligned} \tag{2.26}$$

$$\begin{aligned} I(\boldsymbol{x};\sum_{j\in\mathcal{S}}\boldsymbol{u}_j\boldsymbol{u}_j') = &-\sum\sum_{j,k\in\mathcal{S}}(\boldsymbol{u}_j'B^{(1)}\boldsymbol{u}_k)\boldsymbol{u}_j\boldsymbol{u}_k' \\ &+ \sum_{j\in\mathcal{S}}\sum_{k\notin\mathcal{S}}(\theta_j-\theta_k)^{-1}\{\boldsymbol{u}_j'(A^{(1)}-\theta_j B^{(1)})\boldsymbol{u}_k\}(\boldsymbol{u}_j\boldsymbol{u}_k'+\boldsymbol{u}_k\boldsymbol{u}_j'), \end{aligned} \tag{2.27}$$

where $\mathcal{S}$ indicates the subset of the indices of the eigenvalues of interest. Note that θ_j is assumed to be a simple eigenvalue in (2.26) but not in (2.27). In (2.27) there may be multiple eigenvalues among those of interest ($\mathcal{S}$) and/or among those of no interest ($\bar{\mathcal{S}}$). It is only assumed that the eigenvalues are separated between $\mathcal{S}$ and $\bar{\mathcal{S}}$, namely, eigenvalues which take the same value belong one and only one of $\mathcal{S}$ and $\bar{\mathcal{S}}$.

2.4 Principal components of instrumental variables

When two data matrix X $(n \times p)$ and Z $(n \times q)$ are given on the same n-individual sample, Rao (1964) treated the problem to find the optimal r linear combinations $Y = ZA$ in such a way that the predictive efficiency of Y for X is a maximum. He called a new matrix Y as *principal components of instrumental variables.*

Let again X be an $n \times p$ data matrix and Z an $n \times q$ data matrix. Z may include some or all the variable of X theoretically. Denote the covariance matrices of (X, Z), which indicates an $n \times (p+q)$ matrix such that Z is added to the right side of X, by

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{2.28}$$

Suppose we wish to replace Z by the r linear combinations $Y = ZA$ which jointly predict X as well as possible. The covariance matrix of (X, Y) is

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12}A \\ A'\Sigma_{21} & A'\Sigma_{22}A \end{pmatrix}, \tag{2.29}$$

and the residual covariance matrix of X subtracting its best linear predictor in terms of Y is

$$\Sigma - \Sigma_{12}A(A'\Sigma_{22}A)^{-1}A'\Sigma_{21}. \tag{2.30}$$

We may consider the two measure of predictive efficiency of Y as

$$tr\{\Sigma - \Sigma_{12}A(A'\Sigma_{22}A)^{-1}A'\Sigma_{21}\}, \tag{2.31}$$

or

$$||\Sigma - \Sigma_{12}A(A'\Sigma_{22}A)^{-1}A'\Sigma_{21}||. \tag{2.32}$$

Although the solution is obtained by minimizing either (2.31) or (2.32), it is easier to compute (2.31).

Minimizing (2.31) is the same as maximizing

$$\begin{aligned} tr\{\Sigma_{12}A(A'\Sigma_{22}A)^{-1}A'\Sigma_{21}\} &= tr\{(A'\Sigma_{22}A)^{-1}A'\Sigma_{21}\Sigma_{12}A\} \\ &= \frac{\boldsymbol{a}_1'\Sigma_{21}\Sigma_{12}\boldsymbol{a}_1}{\boldsymbol{a}_1'\Sigma_{22}\boldsymbol{a}_1} + \cdots + \frac{\boldsymbol{a}_r'\Sigma_{21}\Sigma_{12}\boldsymbol{a}_r}{\boldsymbol{a}_r'\Sigma_{22}\boldsymbol{a}_r}, \end{aligned} \tag{2.33}$$

which is the second term of (2.31), assuming that $\boldsymbol{a}_i\Sigma_{22}\boldsymbol{a}_j = 0$, $i \neq j$, without loss of generality. The best choice of A is the set of the r eigenvectors associated with the largest r eigenvalues of the matrix $\Sigma_{21}\Sigma_{12}$ with respect to Σ_{22}, i.e., those of the following eigenvalue problem:

$$(\Sigma_{21}\Sigma_{12} - \lambda_j\Sigma_{22})\boldsymbol{a}_j = 0, \quad j = 1, \ldots, q. \tag{2.34}$$

Then the maximized value of (2.33) is given by

$$\max \quad tr\{\Sigma_{12}A(A'\Sigma_{22}A)^{-1}A'\Sigma_{21}\} = \sum_{i=1}^{r}\lambda_i, \tag{2.35}$$

where λ_i are in order of magnitude, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q$.

Furthermore, this problem can be treated in the sense of maximizing RV-coefficient. Robert and Escoufier (1976) derived the solution in the sense that the geometrical representation of the sample $C(X)$ and $C(Y) = C(ZA)$ will be similar as possible. They call the new variables Y the *principal components of Z with respect to X*, which is the same meaning of Rao(1964)'s PCs of instrumental variables.

With the RV-criterion of optimality,

$$RV(X, ZA) = \frac{tr(A'\Sigma_{21}\Sigma_{12}A)}{\{tr(\Sigma_{11}^2)\cdot tr(A'\Sigma_{22}A)^2\}^{1/2}} \tag{2.36}$$

must be maximized within a multiplicative factor of $1/n$. Then the same eigenvalue problem as (2.34) is solved under the constraint

$$A'\Sigma_{22}A = diag(\sigma_i), \tag{2.37}$$

and the eigenvalues are obtained. The columns of A should be the eigenvectors associated with the first r eigenvalues. The value of the RV-coefficient is then

$$RV(X, ZA) = \left(\sum_{i=1}^{r}\lambda_i\sigma_i\right) \Big/ \left\{tr(\Sigma_{11}^2)\cdot\left(\sum_{i=1}^{r}\sigma_i^2\right)\right\}^{1/2}, \tag{2.38}$$

where σ_i is the variance of the i-th variable. If the values of the variances of the new variables have not been preassigned, an optimal choice of the σ_i's is given by λ_i and global maximum for RV will be attained

$$\max \quad RV(X, ZA) = \left\{\sum_{i=1}^{r}\lambda_i^2/tr(\Sigma_{11}^2)\right\}^{1/2}. \tag{2.39}$$

Robert and Escoufier (1976) also stated about the possibility of variable selection with the RV-coefficient in the above sense. Without loss of generality, assume that Z consists of the first q variables of X. Then we can select a set of variables as Z, which has the largest value of (2.39). This set of variables is the best one among the sets of q variables in the sense of maximizing RV-coefficient.

3 Variable Selection with RV-coefficient in Principal Component Analysis

In principal component analysis, we propose a backward procedure of variable selection in which we discard a variable which has the smallest effect on the principal component (PC) score matrix among the existing variables successively (Mori, Tarumi and Tanaka, 1994a, b). In particular, we focus on the closeness of the relative positions of individuals' PC scores, namely the closeness between the configurations of the PC score matrix before discarding variables and that after discarding. This is to propose variable selection methods in which we select a set of variables which can reproduce as closely as possible the general features of the complete data.

Variable selection methods in PCA have studied by Jolliffe (1972, 1973), McCabe (1984) and Krzanowski (1987a, b) among others. As shown in the overview in section 2.1, Jolliffe's methods are based on the way to remain the variables related to important PCs or to reject those related to unimportant PCs by observing the eigenvalues and the coefficients of the corresponding eigenvectors. McCabe's methods select variables containing (in some sense) as much sample information as possible. Their methods satisfy various optimality criteria derived by themselves, however, they do not necessarily meet the requirement such as to reproduce the general features of the complete data. On the other hand, the aim of Krzanowski's method is to satisfy this requirement with the criterion based on Procrustes Analysis of PC scores.

In our study, the RV-coefficient (Robert and Escoufier, 1976) is used to evaluate the effect on the PC score matrix, and in computation the so-called perturbation theory of eigenvalue problems as well as the exact method are utilized as an approximation of discarding variables.

Since RV-coefficient is a good tool to measure the closeness of the configurations of points associated with two matrices representing the same individuals, it is able to evaluate the closeness between the PC score matrix based on original variables and that based on selected variables. Robert and Escoufier have already discussed the possibility of variable selection with RV-coefficient in their paper (1976), but no examples were given. Then we use RV-coefficient in our methods, although its usage is different from their original idea. (The original idea on variable selection with RV-coefficient will be described in chapter 5.)

The perturbation theory is utilized such as weighting 0 on a variable of interest instead of discarding exactly. This has the following two main purposes: to avoid recomputing to solve an eigenvalue problem every time when a variable is discarded; and to observe the effect of each variable by changing the weight in the future.

3.1 Formulation

3.1.1 Formulation of principal component analysis

Suppose X is an $n \times p$ centered observation matrix with n individuals and p variables. Consider an eigenvalue problem of the matrix X, that is,

$$\frac{1}{p} X X' \boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j, \tag{3.1}$$

where λ_j are the eigenvalues ordered from the largest to the smallest as $\lambda_1, \lambda_2, \ldots, \lambda_p$ and $\boldsymbol{u}_j$ are their associated eigenvectors normalized as $\boldsymbol{u}_j' \boldsymbol{u}_j = 1$, $(j = 1, \ldots, p)$. Let $C \equiv \frac{1}{p} X X'$, the spectral decomposition of C is given by

$$C = U_1 \Lambda_1 U_1' + U_2 \Lambda_2 U_2', \tag{3.2}$$

where $\Lambda_1 = diag(\lambda_1, \ldots, \lambda_r)$ and $\Lambda_2 = diag(\lambda_{r+1}, \ldots, \lambda_p)$ are the r largest eigenvalues and the remaining $p - r$ ones, respectively, and $U_1 = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r)$ and $U_2 = (\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_p)$ are their associated eigenvectors. The PC score matrix A of the r largest eigenvalues is given by

$$A = U_1 \Lambda_1^{1/2}, \tag{3.3}$$

$$T = AA' = U_1 \Lambda_1 U_1'. \tag{3.4}$$

Then the aim of this study is to observe the behavior of this T when a variable is discarded.

3.1.2 Introduction of perturbation

For the sake of convenience for the below formulations, we generalize the eigenvalue problem (3.1) to

$$\frac{1}{p} X W X' \boldsymbol{u}_s = \lambda_s \boldsymbol{u}_s, \tag{3.1$'$}$$

where $W = diag(w_1, \ldots, w_p)$ is a diagonal matrix which has weights on each column, $w_\alpha (\alpha = 1, \ldots, p)$, as diagonal elements. In the case of the original eigenvalue problem $w_\alpha = 1$.

Now let introduce the following perturbation to the weight matrix W:

$$w_\alpha = 1 \longrightarrow w_\alpha = \begin{cases} 1-\varepsilon & \alpha \neq j \\ 1+(p-1)\varepsilon & \alpha = j \end{cases} \quad (1 \leq j \leq p). \tag{3.5}$$

This change of weights with a small perturbation ε is done as the sum of weights keeps p. According to the perturbation in shown (3.5), C is changed to

$$C \longrightarrow \widetilde{C} = C + \varepsilon C^{(1)}. \tag{3.6}$$

Let $c_{ii'}(i, i' = 1, \ldots, n)$ be the elements of C and $x_{ik}(i = 1, \ldots, n;\ k = 1, \ldots, p)$ those of the data matrix X, then

$$\begin{aligned} c_{ii'} &= \frac{1}{p}\sum_{k=1}^{p} x_{ik}x_{i'k}, \\ c_{ii'}^{(1)} &= -\frac{1}{p}\sum_{k=1}^{p} x_{ik}x_{i'k} + x_{ij}x_{i'j}, \end{aligned} \tag{3.7}$$

(see, e.g., Mori and Tarumi, 1993), that is,

$$C^{(1)} = \boldsymbol{x}_j\boldsymbol{x}_j' - C. \tag{3.8}$$

In particular, $\varepsilon = -1/(p-1)$ and this $C^{(1)}$ are substituted in (3.6) when discarding one variable completely among p variables.

3.1.3 Variable selection with RV-coefficient

Here let us consider the RV-coefficient between unperturbed and perturbed PC score matrices to find a variable which have the smallest effect on the relative positions of PC scores in the configuration.

Denote the unperturbed and perturbed PC score matrices by A and $\widetilde{A}$, respectively. Substituting A and $\widetilde{A}$ in (2.4), we can obtain the RV-coefficient between them as

$$RV(A, \widetilde{A}) = \frac{tr(AA'\widetilde{A}\widetilde{A}')}{\left\{tr(AA')^2 \cdot tr(\widetilde{A}\widetilde{A}')^2\right\}^{1/2}} = \frac{tr(T\widetilde{T})}{\left\{tr(T^2)\cdot tr(\widetilde{T}^2)\right\}^{1/2}}. \tag{3.9}$$

Then, if $\widetilde{T}$ is expanded as $\widetilde{T} = T + \varepsilon T^{(1)} + (\varepsilon^2/2)T^{(2)} + O(\varepsilon^3)$, we obtain

$$RV(A, \widetilde{A}) = 1 - \frac{\varepsilon^2}{2}\left[\frac{tr(T^{(1)2})}{tr(T^2)} - \frac{tr(TT^{(1)})}{tr(T^2)}\right] + O(\varepsilon^3) \tag{3.10}$$

(see, Appendix A.1; Castaño-Tostado and Tanaka, 1991), where

$$T^{(1)} = \sum_{j=1}^{r}\sum_{k=1}^{r}(\boldsymbol{u}_j'C^{(1)}\boldsymbol{u}_k)\boldsymbol{u}_j\boldsymbol{u}_k' + \sum_{j=1}^{r}\sum_{k=r+1}^{p}\lambda_j(\lambda_j-\lambda_k)^{-1}(\boldsymbol{u}_j'C^{(1)}\boldsymbol{u}_k)(\boldsymbol{u}_j\boldsymbol{u}_k' + \boldsymbol{u}_k\boldsymbol{u}_j') \tag{3.11}$$

(Tanaka, 1988).

Our variable selection procedure is to discard a variable which has the largest RV-coefficient computed by (3.10) successively.

3.2 Variable selection procedure

Our proposed procedure is a backward elimination. In each step, we compute the RV-coefficient by (3.10) for each one among existing variables in turn, and discard a variable which has the largest RV-coefficient. In the next step, we renew T and repeat the same actions. When the number of remaining variables is equal to the preassigned dimensionality r, we stop the procedure. The process of the procedure is summarized as follows:

1) Apply PCA to the original data and put $q := p$;

2) Specify $r (r < p)$;

3) Compute RV-coefficient between the unperturbed and perturbed PC score matrices, where the perturbed matrix is based on the data matrix with $q-1$ variables obtained by omitted each one among q variables in turn;

4) Find a variable which has the largest RV-coefficient in 3);

5) Apply PCA to the matrix without the variable found in 4);

6) Let $q := q - 1$, and return to 3) unless $q = r$.

3.3 Numerical examples

3.3.1 Plan of evaluation

To evaluate the proposed method, first we compare our results with Jolliffe's and McCabe's ones. In practice we apply "Crime rates data (Ahamad, 1697)" which was analyzed by both Jolliffe and McCabe (their results and discussions were summarized by Jolliffe (1986)). Next, we apply our method to the artificial data sets generated by Jolliffe (1972) and then evaluate our method by following Jolliffe's aspect. Finally, we show a result of analyzing "Automobile data (Becker, et al., 1988)".

In these evaluations we apply some additional patterns of procedure. When we compute $\tilde{T}$, while our procedure proposed in section 3.2 uses an approximation with the perturbation theory, we can obtain $\tilde{T}$ exactly by omitting a variable in practice and recomputing the eigenvalue problem. In this case we can get the RV-coefficient by (3.9). This makes it possible that we compare our proposed method with the exact method, although we lose the advantages of using perturbation which are mentioned above. On the other hand, while we renew T which consists of selected q variables successively in

Table 3.1: Four patterns based on how to obtain T and $\tilde{T}$

T	$\tilde{T}$	
	Perturbed	Exact
Successive	SP (proposed)	SE
Original	OP	OE

each step, we can compute the RV-coefficient with fixed T which consists of all the original p variables. This makes it possible that we evaluate the goodness of the successive procedure. Then four patterns can be considered as shown in Table 3.1. We also apply these four patterns at the same time in the following examples.

3.3.2 Crime rates data

This data set given by Ahamad (1967) consists of measurements of the crime rates in England and Wales for 18 different categories of crime (the variables) for 14 years, 1950–63 (Appendix B.1). Jolliffe (1986) commented about the data set as follows: the sample size $n = 14$ is very small and smaller than the number of variables; furthermore the data are time series, and the 14 observations are not independent, so that the effective sample size is even smaller than 14. This potential problem and other criticisms of Ahamad's analysis caused his and also our motivation to select a subset of variables.

The data seems to have the following 4 clusters of variables, {V3}, {V1, V13}, {V10, V17} and {V2, V4, V5, V6, V7, V8, V9, V11, V12, V14, V15, V16, V18}, by biplot of variables (Figure 3.1) and cluster analysis of variables (Figure 3.2).

Now, we show the result of applying our method to this data set. At the first step in our procedure, we applied PCA to the standardized data set and obtained the eigenvalues and cumulative proportions, $\lambda_1 = 11.937(71.42\%) > \lambda_2 = 2.531(86.56\%) > \lambda_3 = 0.885(91.86\%) > \lambda_4 = 0.638(95.67\%) > \lambda_5 = 0.298(97.45\%) > \cdots$ in order of magnitude. Then we specified $r = 2$ and the result of our proposed method SP is shown in Table 3.2 which indicates the process of discarding and the RV-coefficients in each step. Table 3.3 shows the order of variables rejected by four proposed methods. You can select as many variables as you want from right to left in Table 3.3, starting the last variable. To compare our results with Jolliffe's and McCabe's ones, all the results are summarized in Table 3.4 which is created by modifying Jolliffe(1986)'s Table. In Table 3.4, "3 variables" and "4 variables" in our methods are the last 3 avariables, respectively, in Table 3.3.

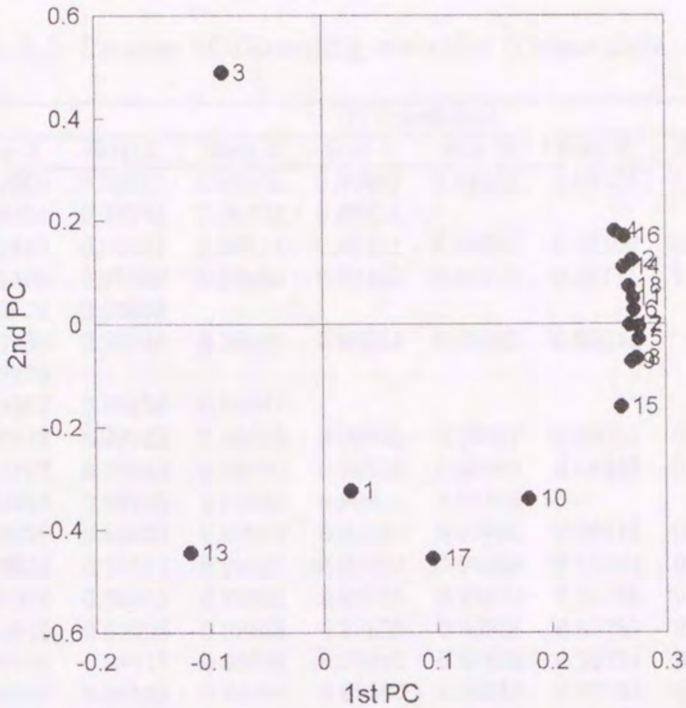


Figure 3.1: Profile plot of variables (Crime data)

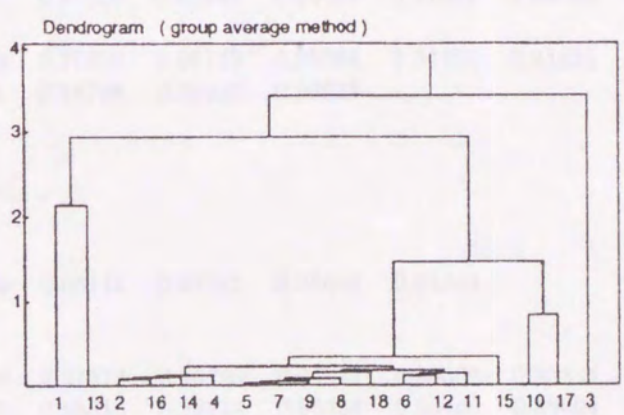


Figure 3.2: Dendrogram obtained by cluster analysis of variables (Crime data)

Table 3.2: Process of discarding variables (Crime data, $r = 2$)

Variable	*RV*-coefficient							
	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8
V1	0.99669	0.99612	0.99540	0.99441	0.99317	0.99152	0.98937	0.98615
V2	0.99954	0.99944	0.99933	0.99921				
V3	0.99649	0.99591	0.99516	0.99421	0.99303	0.99160	0.98953	0.98655
V4	0.99899	0.99880	0.99859	0.99833	0.99773	0.99711	0.99592	0.99474
V5	0.99976	0.99968						
V6	0.99956	0.99946	0.99932	0.99914	0.99882	0.99844		
V7	0.99976							
V8	0.99967	0.99958	0.99947					
V9	0.99948	0.99935	0.99916	0.99890	0.99867	0.99831	0.99777	
V10	0.99733	0.99688	0.99630	0.99550	0.99468	0.99346	0.99195	0.98932
V11	0.99953	0.99945	0.99934	0.99915	0.99888			
V12	0.99939	0.99928	0.99912	0.99894	0.99862	0.99812	0.99730	0.99636
V13	0.99634	0.99574	0.99498	0.99402	0.99269	0.99092	0.98878	0.98571
V14	0.99919	0.99901	0.99880	0.99858	0.99806	0.99756	0.99644	0.99518
V15	0.99919	0.99905	0.99883	0.99852	0.99828	0.99781	0.99718	0.99594
V16	0.99929	0.99915	0.99898	0.99880	0.99836	0.99784	0.99701	0.99592
V17	0.99628	0.99565	0.99486	0.99380	0.99251	0.99081	0.98850	0.98509
V18	0.99941	0.99928	0.99910	0.99890	0.99855	0.99806	0.99743	0.99628
Rejected variable	V7	V5	V8	V2	V11	V6	V9	V12

Variable	*RV*-coefficient							
	Step 1	Step 10	Step 11	Step 12	Step 13	Step 14	Step 15	Step 16
V1	0.98191	0.97514	0.96445	0.94404	0.91692	0.85894	0.42896	0.77072
V2								
V3	0.98244	0.97695	0.96745	0.95584	0.94433	0.91836		
V4	0.99231	0.98798	0.98293	0.96623				
V5								
V6								
V7								
V8								
V9								
V10	0.98603	0.98117	0.97015	0.95682	0.94454			
V11								
V12								
V13	0.98119	0.97374	0.96480	0.94017	0.90302	0.85834	0.73554	0.49142
V14	0.99263	0.98816	0.98214	0.96334	0.90445	0.89694	0.81906	
V15	0.99388	0.99046						
V16	0.99390	0.99016	0.98562					
V17	0.98076	0.97435	0.96271	0.94864	0.93170	0.86114	0.50699	0.48870
V18	0.99459							
Rejected variable	V18	V15	V16	V4	V10	V3	V14	V1

Table 3.3: Result of discarding variables (Crime data, $r = 2$)
(The number in the table indicates the variable's number)

Method	Step 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
SP	7	5	8	2	11	6	9	12	18	15	16	4	10	3	14	1	13 17
OP	7	5	8	2	11	6	9	12	18	16	15	4	3	10	14	1	13 17
SE	7	5	8	2	11	6	9	12	18	15	1	16	10	4	17	13	3 14
OE	7	11	5	1	17	2	6	3	16	18	8	12	9	13	10	14	4 15

Table 3.4: Subsets of selected variables (Crime data)
(Each row corresponds to a selected subset with × denoting a selected variable.)

Method		Variables 1	3	4	5	7	10	13	14	15	16	17
McCabe using criterion												
Three variables	best	×									×	×
	second best	×							×			×
Four variables	best	×						×	×			×
	second best	×					×	×	×			
Jolliffe using criteria B2 B4												
Three variables	B2	×				×		×				
	B4	×	×		×							
Four variables	B2	×				×	×	×				
	B4	×	×		×							×
Using *RV* coefficient												
Three variables	SP	×						×				×
	OP	×	×									×
	SE		×					×	×			
	OE			×					×	×		
Four variables	SP	×						×	×			×
	OP	×	×				×					×
	SE		×					×	×			×
	OE			×			×		×	×		

Note. This table is created by modifying Jolliffe(1986)'s table.

Jolliffe (1986) stated that while the results of Jolliffe's and McCabe's methods were a little different from each other, variable V1 was a member of all the selected variables and variables V10, V13 and V17 were selected by both types of method. From this point of view, our method SP and OP selected variable V1, V13 and V17. Moreover SP and OP selected the same subset of variables as McCabe's best subset when the size of subset is 4.

In comparison with the clusters observed in the profile plot of variables, McCabe's best and second subset with size 3, Jolliffe's B4 with size 3 and 4, our SE with size 3 and 4 selected one variable from each cluster, while our SP selected variables V1 and V13 which are close to each other in the profile plot of variables.

Comparing our 4 methods with each other, similar results were obtained in SP and OP, and SP had the same variables as ones by SE in the first half steps. However OE selected variable V15 which was selected by neither Jolliffe nor McCabe and selected variables from the same cluster.

3.3.3 Jolliffe's artificial data

Jolliffe (1972) generated a large number of artificial data sets, conforming to one of five predetermined models. Each model was constructed in such a way that certain variables were linear combinations of other variables, except for a random disturbance, and hence were redundant (see the definition in Table 3.5). He then tested his various rejection methods on the data sets to see whether the variables they rejected were redundant ones. In all his models, there were some categories of choice regarding how well the retained variables are, which were labeled as "best", "good", "moderate" or "bad". Table 3.5 indicates the definition of the constructed variables for each of models I–IV, and "best" and "good" subsets for them. Model V was more complicated, and we omitted it.

According to his models I–IV, we generated 100 samples of size $n = 100$ for each of these models. The table 3.6 shows the results of applying our methods to these data sets as a monte carlo simulation. The dimensionality r is 3 for model I–III and 4 for model IV according to the number of variables should be retained (i.e., m in Table 3.5).

As results, SE selected "best" and "good" subsets at the highest rate (75%) totally, and in order of magnitude, SP, OP and OE had 65.5%, 64.25% and 51.0%, respectively. SE also selected 100% of "best" and "good" subsets in model I–III, while all the methods selected 100% of those subsets in model III. On the other hand, SP had the highest rate (42.75%) by observing the rates in "best" subset, and only SP and OP selected "best" or "good" subsets from every models. From this simulation, it can be stated that SP

Table 3.5: Definition of constructed artificial variables and subsets of variables should be retained (Jolliffe, 1972)

Variable		Model I	Model II	Model III	Model IV
1	v_1	z_1	z_1	z_1	z_1
2	v_2	z_2	z_2	z_2	z_2
3	v_3	z_3	z_3	z_3	$z_2 + z_3$
4	v_4	$z_1 + 0.5z_4$	$z_1 + 0.5z_4$	$z_1 + 0.8z_2 + 0.6z_4$	z_4
5	v_5	$z_2 + 0.7z_5$	$z_2 + 0.7z_5$	$z_2 + 0.7z_5$	$z_4 + 0.75z_5$
6	v_6	$z_3 + z_6$	$z_2 + z_6$	$z_3 + 0.5z_6$	$2z_4 + 0.75z_5 + 1.5z_6$
7	v_7				z_7
8	v_8				$z_7 + 0.5z_8$
9	v_9				$2z_7 + 0.5z_8 + z_9$
10	v_{10}				$3z_7 + z_8 + z_9 + z_{10}$
n, p		100, 6	100, 6	100, 6	100, 10
m		3	3	3	4
best		(1, 4), (2, 5)	{1, 2, 3}	{1, 2, 3}	(1), (2, 3), (4, 5, 6)
		(3, 6)	{2, 3, 4}	{1, 2, 6}	(7, 8, 9, 10)
good		—	{1, 3, 5}, {1, 3, 6}	{1, 5, 6}, {1, 3, 5}	—
			{3, 4, 5}, {3, 4, 6}	{2, 4, 6}, {2, 3, 4}	
				{3, 4, 5}, {4, 5, 6}	

v_i : name of variable
z_i : standardized normal variates
n, p : number of observations, number of variables
m : number of variables should be retained
{ } : subset of variable should be retained
( ) : any subset containing one variable from each ( ) should be retained

has selection power on the average. There is a room for improvement, however, since the rates were not stable between the models.

3.3.4 Automobile data

As the third example we applied to "Automobile data (Becker et al., 1988)" which has 74 observation on 10 variables checking automobile's capacities (Appendix B.2). This data has almost 4 clusters, namely, the variable "price" {V1}, the variables related to "performance" {V2, V10}, those related to "size" {V6, V7, V8, V9} and those related to "width" {V3, V4, V5} by observing the profile plot of variables (Figure 3.3).

The result of applying our methods to the standardized data set is shown in Table 3.7 and Table 3.8. The dimensionality $r = 2$ because $\lambda_1 = 6.526(66.15\%) > \lambda_2 = 1.012(76.41\%) > \lambda_3 = 0.825(84.78\%) > \lambda_4 = 0.417(89.00\%) > \cdots$.

While the all variables related to "size" were discarded in the beginning steps, Price

Table 3.6: Results of Monte Carlo variable selection with RV-coefficient (Jolliffe's artificial data, 100 samples with $n = 100$ in each model)

Method		Model I	II	III	IV	sum	%	best & good sum(%)
SP (proposed)	best	96	0	37	38	171	42.75	65.5
	good	—	28	63	—	91	22.75	
OP	best	96	0	25	40	161	40.25	64.25
	good	—	21	75	—	96	24.0	
SE	best	100	3	35	0	138	34.5	75.0
	good	—	97	65	—	162	40.5	
OE	best	100	4	64	0	168	42.0	51.0
	good	—	0	36	—	36	9.0	

Figure 3.3: Profile plot of variables (Automobile data)

Table 3.7: Process of discarding variables (Automobile data, $r = 2$)

Variable	RV-coefficient							
	Step1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8
V1	0.91949	0.85203	0.81040	0.71565	0.85942	0.76091	0.78826	0.49736
V2	0.99311	0.98992	0.98481	0.97142	0.95473	0.93208		
V3	0.98514	0.97811	0.97102	0.93805	0.94507	0.87005	0.80463	0.42412
V4	0.97737	0.97220	0.96554	0.96780	0.95719	0.90940	0.86213	
V5	0.98996	0.98627	0.98155	0.97406	0.96756			
V6	0.99797							
V7	0.99707	0.99532						
V8	0.99450	0.99162	0.98667	0.97730				
V9	0.99656	0.99424	0.99074					
V10	0.98847	0.98085	0.97085	0.95599	0.94236	0.91627	0.81608	0.58268
Rejected variable	V6	V7	V9	V8	V5	V2	V4	V10

Table 3.8: Results of discarding variables (Automobile data, $r = 2$)

Method	Step								
	1	2	3	4	5	6	7	8	
SP	V6	V7	V9	V8	V5	V2	V4	V10	V1 V3
OP	V6	V7	V9	V8	V2	V5	V10	V4	V1 V3
SE	V6	V6	V9	V8	V5	V10	V2	V4	V1 V3
OE	V6	V4	V10	V1	V3	V8	V2	V5	V7 V9

Variables: V1 Price, V2 Miles/g, V3 Headroom, V4 Rear Seat, V5 Trunk, V6 Weight, V7 Length, V8 Turning, V9 Displacement, V10 Gear Ratio

(V1), Headroom (V3) and Gear ratio (V10) were selected from each of the other three clusters, which are reasonable variables to buy or evaluate automobiles.

3.4 Discussion

In these three numerical examples the proposed method gave reasonable results of variable selection in PCA.

In the comparison of our 4 methods with each other, since SP (successive T and perturbed $\tilde{T}$) and OP (original T and perturbed $\tilde{T}$) gave similar results, it is good enough to use the successive method to select q variables among the existing p variables. Comparing SP with SE (successive T and exact $\tilde{T}$), while in the first half steps SP could select the same variables as those selected by SE, SP selected variables in different order from SE in the last half steps. This means that some errors by perturbation exists.

On the other hand, it cannot be stated that OE (original T and exact $\widetilde{T}$) could select a reasonable subset. That is because the method does not proceed in such a way that the remaining variables have the weight representing the rejected variables in a backward procedure. This is also observed from the fact that SE did not select reasonable variables well in model IV which has a lot of redundant variables in the artificial data example.

Thus, the proposed SP is enough method to select variables.

In addition, Krzanowski (1987a, b) has studied variable selection in PCA (see, section 2.1). His criterion is to compare the configurations of PC score based on the original data matrix with that based on rejected matrix. Two differences exist mainly between his method and ours: his method is not successive, which means that it always uses a PC score matrix based on an original data as a comparative basis; and his criterion is not a comparison of relative positions but one of just configurations of PC scores.

4 Variable Selection with RV-coefficient in Hayashi's Third Method of Quantification

Hayashi's third method of quantification, whose algorithm is the same as that of correspondence analysis, is useful in multivariate data analysis. Actually, categorical answers are occasionally used in surveys and examinations conducted in various areas such as psychology, medical science, social science, and so on. In these surveys we often meet the problem such that there are too many variables or items for the participants. Then we consider again how to reduce as many variables as possible without loss of original information. It is desirable to propose an appropriate variable selection method in Hayashi's third method of quantification in the meaning of keeping the internal structure or information of the sample. However there are not so many variable selection methods in this analysis. For example in the small number of studies, Xia and Yang (1988) proposed three criteria for variable selection in Hayashi's third method of quantification and gave two practical procedures, referring the variable selection in factor analysis studied by Tanaka and Kodake (1981) and Tanaka (1983). But they did not discuss the fundamental problems in Hayashi's third method of quantification and its variable selection, which will be mentioned in the next section. Since Hayashi's third method of quantification can be thought the categorical version of PCA, there exists a possibility to propose a similar procedure to the variable selection in PCA proposed in chapter 3, if we can clear the existing problems.

Then, taking a similar way in the previous chapter, we propose backward procedures of variable selection in which we discard a variable which has the smallest effect on the sample score matrix among the existing variables successively (Mori and Tarumi, 1994). The procedures have solutions for the typical problems in Hayashi's third method of quantification itself and also in the selection process in this analysis. In the procedures we use the RV-coefficient (Robert and Escoufier, 1976) and the perturbation theory of eigenvalue problems as well as the exact method in computation.

Though we deal with only binary type (0 or 1) data in this study, the principles and the procedures can be applied to multiple type data.

4.1 Typical problems in treating categorical data sets and in its variable selection

There are typical problems in dealing with categorical data, especially in Hayashi's third method of quantification. The problems exist both in the analysis itself and in the process of selecting variables.

First problem which exists in Hayashi's third method of quantification itself is as follows: it is well known that categorical data has two different data forms, free-choice (FC) form and item-category (IC) form (Figure 4.1). They are equivalent with each other with respect to information contained, but typically lead to different results in Hayashi's third method of quantification (see, e.g., Yamada and Nishisato, 1993). Some authors indicated that a data form should be chosen based on the purpose of analysis or the properties of the data (see, e.g., Iwasaki, 1989; Okamoto, 1992).

To deal with this problem we propose two types of procedure for the two data forms, respectively, using the same principle in computation.

Next problem sometimes occurs in the process of selecting variables in Hayashi's third method of quantification. Hayashi's third method of quantification has the operation such as dividing by row sum and column sum of the data matrix in computation. This means that we cannot compute if a row sum becomes 0 in the selection process. As shown in Figure 4.2, such a case can arise easily where the data form is FC type. In Figure 4.2, if the third variable (column) is removed among 4 variables, then the third row sum becomes equal to 0. This problem is thought rather serious in selecting a set of variables. For this problem we can take the following actions: to continue the selection by omitting every individual (row) whose sum equals zero in each selection step; before starting the selection, to change the data form from FC form to IC form. In IC form every row sum is equal to the number of columns in any selection step (In this action we have to notice that FC and IC form give different results from each other in original analysis, and it happens that the computation stop when a column sum becomes 0. Let us show an example of the latter case in Figure 4.1: if all elements of the third variable in X_{FC} are 1, all elements of the left column of the third variable in X_{IC} are 0); and to introduce a perturbation as discarding a variable to avoid the case where the row sum equals 0, that is, 0 is weighted on a variable of interest instead of discarding exactly .

To deal with this problem we adopt the last two actions. The second one is included in the action for the first problem as mentioned above. As for the third action, in a backward procedure the perturbation theory will be used not only to weight 0 on one variable of interest in each step but also to weight 0 on all the variables found in the former steps at

Free Choice form   Item Category form

$$X_{FC} = \begin{pmatrix} 1 & 0 & 1 & \dots & 0 \\ 1 & 1 & 1 & \dots & 0 \\ 0 & 0 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 1 & \dots & 0 \end{pmatrix} \quad \Longleftrightarrow \quad X_{IC} = \left(\begin{array}{cc|cc|cc|c|cc} 0 & 1 & 1 & 0 & 0 & 1 & \dots & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & \dots & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & 1 & 0 & 1 & \dots & 1 & 0 \end{array}\right)$$

Figure 4.1: Different data form

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \\ n \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 1 \end{pmatrix} \end{array} \quad \Longrightarrow \quad \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \\ n \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 4 \end{array} \\ \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 1 \end{pmatrix} \end{array}$$

Figure 4.2: A Case where row sum = 0 (When the third columns is discarded, the third row sum becomes 0.)

the same time. This is the third purpose to use the perturbation theory in addition to the two purposes described in chapter 3, but it seems to be very fundamental in Hayashi's third method of quantification.

4.2 Formulation

4.2.1 Formulation of Hayashi's third method of quantification

Suppose we have a set of n samples (individuals) on p categories (variables). This is expressed as an n rows $\times$ p columns matrix X_{FC} in FC form and an $n \times 2p$ matrix X_{IC} in IC form, which have only binary data, 0 or 1. For the sake of convenience, let us denote the data matrix by $n \times m$ matrix X. They have the same number of variables, m, but $m = p$ in X_{FC} and $m = 2p$ in X_{IC}.

In Hayashi's third method of quantification, consider an eigenvalue problem of

$$C \equiv D_r^{-1/2} X D_c^{-1} X' D_r^{-1/2}, \tag{4.1}$$

where

$$D_r = diag(f_1, \ldots, f_n) \quad (f_i \text{ is the } i\text{-th row sum}),$$
$$D_c = diag(g_1, \ldots, g_m) \quad (g_l \text{ is the } l\text{-th column sum}),$$

that is,

$$(C - \lambda_j I)\boldsymbol{u}_j = 0 \quad (j = 1, \ldots, m), \tag{4.2}$$

where λ_j are the eigenvalues ordered from the largest to the smallest as $\lambda_1, \lambda_2, \ldots, \lambda_m$ and $\boldsymbol{u}_j$ are their associated eigenvectors normalized as $\boldsymbol{u}_j'\boldsymbol{u}_j = 1$.

The sample score matrix of the r largest eigenvalues is given by U_1 ($U_1 = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r)$, $r \leq m$), then we denote

$$A = U_1, \tag{4.3}$$

$$P = U_1 U_1' = AA'. \tag{4.4}$$

The aim of this study is to observe the behavior of this P when a variable is discarded.

4.2.2 Introduction of perturbation

For the sake of convenience for the formulation below, to generalize the eigenvalue problem (4.2), we change the matrix C to

$$C = D_{r(w)}^{-1/2} X W D_c^{-1} X' D_{r(w)}^{-1/2}, \tag{4.1$'$}$$

where $W = diag(w_1, \ldots, w_m)$ is a diagonal matrix which has weights on each column, $w_\alpha (\alpha = 1, \ldots, m)$, as diagonal elements and $D_{r(w)}$ is $diag(X'W\mathbf{1})$.

Now let the weights w_α be changed from 1 to as follows by introducing the perturbation:

$$w_\alpha = 1 \longrightarrow w_\alpha = \begin{cases} 1 - \varepsilon & \alpha \neq l \\ 1 + (m-1)\varepsilon & \alpha = l \end{cases} \quad (1 \leq l \leq m). \tag{4.5}$$

According to the perturbation in sown (4.5), the matrix C is changed to

$$C \longrightarrow \tilde{C} = C + \varepsilon C^{(1)}. \tag{4.6}$$

Here let us denote the elements of C and X by $c_{ii'}$ ($i, i' = 1, \ldots, n$) and x_{ik} ($i = 1, \ldots, n$; $k = 1, \ldots, m$), respectively. When the l-th column is discarded, elements of $C^{(1)}$ is given by

$$c_{ii'}^{(1)} = -\frac{m}{2} c_{ii'} \left(\frac{x_{il}}{f_i} + \frac{x_{i'l}}{f_{i'}} \right) + m \frac{x_{il} x_{i'l}}{g_l \sqrt{f_i f_{i'}}}, \tag{4.7}$$

where

$$c_{ii'} = \sum_{k=1}^{m} \frac{x_{ik} x_{i'k}}{g_k \sqrt{f_i f_{i'}}} \tag{4.8}$$

(see, e.g., Mori and Tarumi, 1993).

As the case of discarding m_j $(1 < m_j < m)$ columns, the l_1-th, $\ldots$, and the l_{m_j}-th columns, are discarded at the same time, the elements of $C^{(1)}$ are changed from (4.7) to

$$c_{ii'}^{(1)} = -\frac{m}{2}c_{ii'}\left(\frac{\sum_{k=l_1}^{l_{m_j}} x_{ik}}{f_i} + \frac{\sum_{k=l_1}^{l_{m_j}} x_{i'k}}{f_{i'}}\right) + m\sum_{k=l_1}^{l_{m_j}} \frac{x_{ik}x_{i'k}}{g_k\sqrt{f_i f_{i'}}}, \tag{4.9}$$

which is the simple sum of the $C^{(1)}$s expressed as (4.7), i.e., $\sum_{k=l_1}^{l_{m_j}} C_k^{(1)}$ where $C_k^{(1)}$ is $C^{(1)}$ of k-th variable.

In particular, $\varepsilon = -1/(m-1)$ and $C^{(1)}$ in (4.7) or (4.9) are substituted when discarding one column or m_j columns completely among m columns.

4.2.3 Variable selection with RV-coefficient

Here let us use the RV-coefficient to find a variable which has the smallest effect on the configuration of the sample score matrix when it is discarded. Now the unperturbed and perturbed sample score are denoted by A and $\widetilde{A}$, then the RV-coefficient between A and $\widetilde{A}$ is given by

$$RV(A,\widetilde{A}) = \frac{tr(AA'\widetilde{A}\widetilde{A}')}{\left\{tr(AA')^2 \cdot tr(\widetilde{A}\widetilde{A}')^2\right\}^{1/2}} = \frac{tr(P\widetilde{P})}{\left\{tr(P^2)\cdot tr(\widetilde{P}^2)\right\}^{1/2}}. \tag{4.10}$$

Since U is orthogonal,

$$RV(A,\widetilde{A}) = \frac{tr(P\widetilde{P})}{q} \tag{4.10$'$}$$

(Castaño-Tostado and Tanaka, 1990).

If $\widetilde{P}$ is expanded as $\widetilde{P} = P + \varepsilon P^{(1)} + (\varepsilon^2/2)P^{(2)} + O(\varepsilon^3)$, we obtain

$$RV(A,\widetilde{A}) = 1 - \frac{\varepsilon^2}{2}\cdot\frac{tr(P^{(1)2})}{q} + O(\varepsilon^3), \tag{4.11}$$

where

$$P^{(1)} = \sum_{j=1}^{r}\sum_{k=r+1}^{m}(\lambda_j - \lambda_k)^{-1}(\boldsymbol{u}_j' C^{(1)}\boldsymbol{u}_k)(\boldsymbol{u}_j\boldsymbol{u}_k' + \boldsymbol{u}_k\boldsymbol{u}_j'), \tag{4.12}$$

(Castaño-Tostado and Tanaka, 1990, 1991; Tanaka, 1988) using $C^{(1)}$ in (4.7) or (4.9).

Our variable selection procedure is to discard a variable which has the largest RV-coefficient (4.11) successively.

4.3 Variable selection procedures

Now we show our variable selection procedures. As stated in section 4.1 we proposed two types of selection procedure according to the given data form. They are backward eliminations. In these two types of procedure, furthermore, we can consider some more patterns of procedure depending on the following aspects:

(a) Whether the variables found in the former steps remain or omit in the next step;

(b) How to discard a variable, that is, whether $\tilde{A}$ is obtained by the perturbation theory or exact method;

(c) Which matrix is used as A, the original data matrix, which means that A is fixed in any step, or the discarded matrix found in the former step, which means A is obtained successively in every step.

These possible patterns are summarized in Table 4.1. "Remain" in aspect (a) is considered as a means to avoid the first problem. Aspects (b) and (c) are considered as to evaluate each other.

From the property of aspect (a), it is nonsense that we obtain $\tilde{A}$ exactly because the aim of aspect (a) is to discard variables approximately by introducing perturbation. Moreover, we always use the original data matrix as A in every step since all the variables found in the former steps are remained in the next step. Then we consider only the pattern "perturbation"–"original" in "remain" category. The q variables selected in the q-th step by this strategy, additional speaking, are the same as those obtained by checking all the combinations of q variables. This set of variables has the largest sum of q RV-coefficients among others in the first step.

In the pattern "omit"–"perturbation" there is only one strategy in spite of the way to obtain A. That is because the term $P\ (= AA')$ is not contained in eq.(4.11) to compute $RV(A, \tilde{A})$.

Now we show the details of the typical two procedures, FC-R2 and IC-O1.

(FC-R2) For a free-choice form:

1) Apply Hayashi's third method of quantification to the original data and put $q := m(= p)$;

2) Specify $r (r < m)$;

Table 4.1: Considerable patterns of procedure

Aspect			Data form	
(a)	(b)	(c)	Free-Choice	Item-Category
remain (and weighting 0)	perturbation	successive	—	—
		original	FC-R2	IC-R2
	exact	successive	—	—
		original	—	—
omit	perturbation	successive	FC-O1*	IC-O1**
		original		
	exact	successive	FC-O3*	IC-O3**
		original	FC-O4*	IC-O4**

Note. * these methods have the risk of the second problem.
**in these methods there are some cases where certain column sum(s) = 0 in the first step.

3) Compute RV-coefficient between the unperturbed and perturbed sample score matrices, where the perturbed matrix is based on the data matrix without each one among q variables and $m-q$ variables found in the former steps (i.e., $m_j(=m-q+1)$ variables are discarded at the same time by using (4.9));

4) Find a variable which has the largest RV-coefficient in 3);

5) Let $q := q - 1$, and return to 3) unless $q = r$.

While the above procedure is described exactly as a backward elimination, note that it is enough to compute RV-coefficients just once in the first step as stated in the previous paragraph.

(IC-O1) For an item-category form:

1) Apply third method of quantification to the original data and put $q_1 := p, q_2 = m(= 2p)$;

2) Specify $r(r < m)$;

3) Compute RV-coefficient between the unperturbed and perturbed sample score matrices, where the perturbed matrix is based on the data matrix without each one among q_1 variables in turn (i.e., $m_j(=2)$ columns contained in each one among q_1 variables are discarded at the same time by using (4.9));

4) Find a variable which has the largest RV-coefficient in 3). Suppose it is the j-th variable;

5) Apply the third method of quantification to the matrix without m_j columns in the j-th variable found in 4);

6) Put $q_1 := q_1 - 1$ and $q_2 := q_2 - m_j$. If $q_2 - m_{j'} > r$ in regard to any $j'(j' = 1, \ldots, q_1)$ then return to 3).

As mentioned in section 4.1, pay attention to that X_{IC} has a variable whose elements are all 0 or all 1, when all participants have the same response. Unfortunately we cannot apply Hayashi's third method of quantification in this case because the column sum = 0. For such a case we may adopt the same strategy as 3) in FC-R2, i.e., IC-R2, or start the procedure after omitting such a variable.

4.4 Numerical Examples

As an illustration of our procedures we applied our methods to two data sets. One is a set of "Spirits data (Arima and Ishimura, 1987)" and the other is "Fatigue data (Maehashi et al, 1992)".

4.4.1 Spirits data

The data consists of 20 samples on 7 categories, that is the response that 20 college female students were asked whether or not they like each of 7 kinds of alcoholic drinks (Appendix B.3). Arima and Ishimura showed that Whisky (V1), Wine (V3), Japanese Sake (V4) and Cocktail (V7) are close to each other in the profile plot of variables, but the others are separated (Figure 4.3).

The eigenvalues and cumulative proportions given by Hayashi's third method of quantification are shown in Table 4.2, and we applied our method with $r = 3$. As subsets of variables with size four, {V1, V2, V4, V6} were selected by our method FC-R2 as shown in Table 4.3, and {V2, V3, V4, V5} by IC-O1 in Table 4.4. It seems to be shown that the proposed methods give reasonable results of variable selection in Hayashi's third method of quantification.

4.4.2 Fatigue data

Maehashi et al.(1993) tried to make a questionnaire of subjective symptoms of fatigue for school-children based on one for adults which has been already developed. The questionnaire for adults consists of 30 variables (questions) about subjective symptom of fatigue

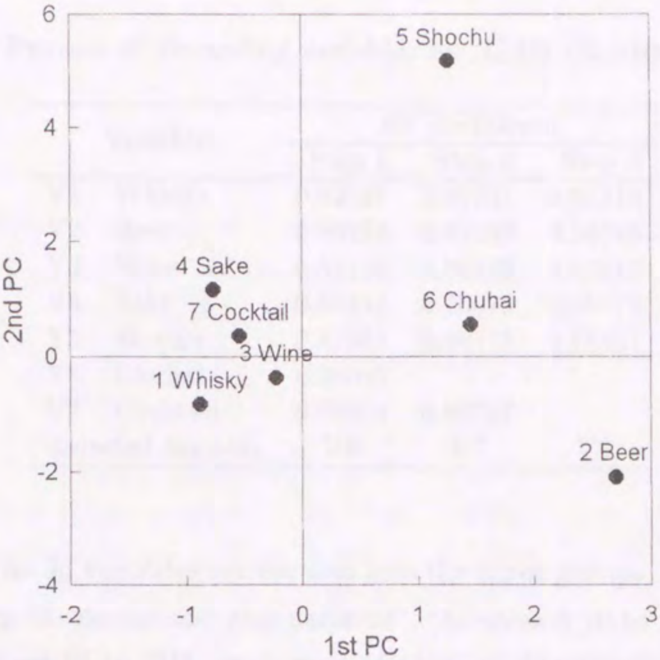


Figure 4.3: Profile plot of variables (Spirits data)

Table 4.2: Eigenvalues and their cumulative proportions (Spirits data)

	1	2	3	4	5	6
Eigenvalue	0.38537	0.27705	0.24709	0.12984	0.10886	0.03281
Cumulative proportion (%)	32.63	56.09	77.01	88.00	97.22	100.00

Table 4.3: Process of discarding variables by FC-R2 (Spirits data, $r = 3$)

Variable		RV-coefficient		
		Step 1	Step 2	Step 3
V1	Whisky	0.99909	0.99828	0.99919
V2	Beer	0.99919	0.99879	0.99881
V3	Wine	0.99947	0.99971	
V4	Sake	0.99848	0.99788	0.99897
V5	Shochu	0.99945	0.99923	0.99919
V6	Chuhai	0.99881	0.99900	0.99800
V7	Cocktail	0.99960		
Rejected variable		V7	V3	V5

Table 4.4: Process of discarding variables by IC-O1 (Spirits data, $r = 3$)

Variable		*RV*-coefficient		
		Step 1	Step 2	Step 3
V1	Whisky	0.92527	0.89731	0.91219
V2	Beer	0.89350	0.86198	0.54040
V3	Wine	0.85116	0.86330	0.83011
V4	Sake	0.63447	0.70076	0.56472
V5	Shochu	0.87965	0.86175	0.75451
V6	Chuhai	0.96045		
V7	Cocktail	0.90864	0.89767	
Rejected variable		V6	V7	V1

(Appendix B.4). The 30 variables are divided into the three groups. The first 10 variables belong to the group "I. *drowsiness and dullness*", the second 10 to "II. *difficulty concentration*" and the third 10 to "III. *projection of physical disintegration*". The conductors have the participants answer "yes" or "no" for each question, and analyze their fatigue condition or their change between the condition before physical movements (PM) and that after PM. Maehashi et al.(1993) conducted this questionnaire to school-children and gathered answers from more than 1500 children. Since it became clear in the survey that the number of variables was too large for children, they decided to reduce the number of variables. Then they selected the most effective 15 variables subjectively by examining the gathered answers, hearing from teachers, applying cluster analysis of variables and so on. The selected variables were {V2, V4, V5, V6, V7, V13, V14, V15, V18, V19, V21, V25, V27, V30} under the condition such that 5 variables were chosen certainly from each group.

We applied our method to this data as a simulation in spite that the confidence of this data was not so high because participants were all young children. The target number of variables selected was the same as the previous study, 15.

It was not so easy to decide the dimensionality r because the eigenvalues were changed very slightly (Table 4.5). Then we applied FC-R2 with $r = 15$, that is the maximum dimensionality to select 15 variables, to the 100 samples extracted from the 6th grade students' data. The data had no row whose sum equals zero.

First we selected 15 among 30 variables directly under no condition. Table 4.6 shows the results with not RV-coefficients but coefficients of ε^2 in eq.(4.11). It is often more convenient to observe the coefficients of ε^2 than to check the small changes of RV-coefficients directly. We discarded variables in order indicated in the "Order" columns of Table

Table 4.5: Eigenvalues and their proportions (Fatigue data)

	Before PM			After PM		
	Eigenvalue	Prop.*	Cum.**	Eigenvalue	Prop.*	Cum.**
1	0.39011	10.06	10.06	0.35353	9.14	9.14
2	0.30816	7.94	18.00	0.30853	7.97	17.11
3	0.26527	6.84	24.84	0.26965	6.97	24.08
4	0.25579	6.59	31.43	0.25474	6.58	30.66
5	0.20737	5.34	36.77	0.23242	6.01	36.67
6	0.20559	5.30	42.07	0.21756	5.62	42.29
7	0.18608	4.80	46.87	0.20301	5.25	47.54
8	0.18237	4.70	51.57	0.20210	5.22	52.76
9	0.16968	4.37	55.94	0.18664	4.82	57.59
10	0.15621	4.03	59.97	0.16663	4.31	61.89
11	0.14636	3.77	63.74	0.16104	4.16	66.05
12	0.14203	3.66	67.40	0.14164	3.66	69.71
13	0.13964	3.60	71.00	0.13523	3.49	73.21
14	0.12399	3.20	74.20	0.11699	3.02	76.23
15	0.11055	2.85	77.05	0.11137	2.88	79.11
16	0.10411	2.68	79.73	0.10527	2.72	81.83
17	0.09403	2.42	82.15	0.09266	2.39	84.23
18	0.09205	2.37	84.53	0.08678	2.24	86.47
19	0.08562	2.21	86.73	0.08121	2.10	88.57
20	0.07278	1.88	88.61	0.07445	1.92	90.49
21	0.06924	1.78	90.39	0.07083	1.83	92.32
22	0.06586	1.70	92.09	0.05953	1.54	93.86
23	0.06180	1.59	93.68	0.05139	1.33	95.19
24	0.05787	1.49	95.18	0.04422	1.14	96.33
25	0.04728	1.22	96.39	0.03730	0.96	97.30
26	0.04398	1.13	97.53	0.03486	0.90	98.20
27	0.03933	1.01	98.54	0.03060	0.79	98.99
28	0.03253	0.84	99.38	0.02550	0.66	99.65
29	0.02403	0.62	100.00	0.01365	0.35	100.00

Note. * Prop.: Proportion
**Cum. : Cumulative proportion

Table 4.6: Result of variable selection (all the 30 variables, Fatigue data, $r = 15$)

Variable (Symptom of fatigue)		Before PM		After PM	
		Coef.* of ε^2	Order	Coef.* of ε^2	Order
V1	your head feeling weary	0.15790	13	0.22226	
V2	feeling exhausted	0.11277	11	0.24175	
V3	feeling your legs tired	0.18613	15	0.91161	
V4	feeling like yawning	1.51598		0.10591	14
V5	feeling mentally sluggish	0.07180	9	0.54385	
V6	feeling sleepy	1.86834		4.59097	
V7	feeling your eyes tired	0.02232	5	0.20025	
V8	feeling unable to coordinate	0.37860		0.12580	15
V9	feeling unsteady on your feet	0.02132	4	0.01686	2
V10	feeling to lie down	0.61316		0.79442	
V11	feeling distracted	1.30453		0.06374	10
V12	feeling uncommunicative	0.00148	1	0.66224	
V13	feeling irritated	0.22233		0.00750	1
V14	feeling restless	0.34641		0.01983	4
V15	feeling to lose interest	0.43269		0.17449	
V16	feeling of forgetfulness	1.00999		0.05242	9
V17	making many mistakes	0.25589		0.03723	5
V18	feeling worried	0.19169		0.01744	3
V19	feeling unable to be still	0.14962	12	0.07798	11
V20	feeling to lose your temper	0.01303	3	0.04732	7
V21	headaches	0.46788		2.42720	
V22	stiff neck	0.63853		2.13637	
V23	backaches	0.03070	8	0.75962	
V24	difficult to breathe	0.18543	14	0.04304	6
V25	thirsty	0.43226		0.09117	13
V26	hoarse voice	0.02906	7	0.16128	
V27	feeling dizzy	0.24581		0.08112	12
V28	eyes twitching	0.09458	10	0.23025	
V29	hands and legs trembling	0.00649	2	0.05090	8
V30	feeling sick	0.02438	6	0.44342	

Note. ** Coef.: Coefficient

4.6. Since using FC-R2, the selected 15 variables are the best subset whose sum of RV-coefficients is the largest among others with size of 15. The selected variables are {V4, V6, V8, V10, V11, V13, V14, V15, V16, V17, V18, V21, V22, V25, V27} before PM and {V1, V2, V3, V5, V6, V7, V10, V12, V15, V21, V22, V23, V26, V28, V30} after PM. Comparing two results, they are almost reversible. It seems that variables in group I and III, which are related to the physical fatigue, play important roles before PM and variables in group II does after PM when all the variables are included in the analysis. Then we separated the data in the three variable groups and analyzed each data separately under the constraint to choose 5 variables in each group.

The results are indicated in Table 4.7.a to Table 4.7.c. Each table indicates the variables in discarding order for both pre-PM and post-PM with their coefficients of ε^2 and RV-coefficients. The number of individuals in each data set was decreased from 100 by omitting every individual whose row sum equals to zero when 30 variables were divided into the three groups. Observing Table 4.7.a trough Table 4.7.c, it can be stated that {V2, V3, V8, V10} in the group I, {V16} in II and {V21, V22, V30} in III can be reasonable candidates. But we have to notice that most selected variables were different from each other between pre-PM and post-PM in the group II. This suggested that variables in group II plays a particular role to describe one's fatigue conditions, then more consideration is necessary.

4.5 Discussion

In this chapter we studied variable selection methods in Hayashi's third method of quantification in which we can select variables which have small effect on the configuration of the sample score matrix. Our methods were proposed so as to analyze both free-choice and item-category data form, and also to avoid the computational disadvantage in selection process.

We applied our methods to two data sets as numerical examples. In the first example our methods could select reasonable variables from variable clusters observed the profile plot of variables. In second example they could select interpretable variables among all the variables. Unfortunately selected variables depend upon which method is applied, FC type or IC type. One of this reason is that Hayashi's third method of quantification gives different results for the different data form. Other one is that the perturbation theory is utilized in our procedures. It has the risk of errors yielded by approximation, while it is very useful when the computation cannot be done exactly.

Table 4.7.a: Result of variable selection (Group I, Fatigue data)

	Before PM ($n = 87$)			After PM ($n = 96$)		
Order of discarding	Coefficient of ε^2	RV-coefficient	Discarded variable	Coefficient of ε^2	RV-coefficient	Discarded variable
1	0.05385	0.99967	V9	0.07209	0.99955	V7
2	0.06717	0.99959	V5	0.09434	0.99942	V5
3	0.09163	0.99943	V7	0.18132	0.99888	V1
4	0.17667	0.99891	V4	0.33149	0.99795	V4
5	0.26565	0.99836	V1	0.35446	0.99781	V6
6	0.29103	0.99820	V3	0.43149	0.99734	V2
7	0.38260	0.99764	V8	0.45542	0.99719	V10
8	0.41799	0.99742	V2	0.52339	0.99677	V3
9	0.68865	0.99575	V6	0.59378	0.99633	V8
10	1.32330	0.99183	V10	0.96370	0.99405	V9

Table 4.7.b: Result of variable selection (Group II, Fatigue data)

	Before PM ($n = 75$)			After PM ($n = 38$)		
Order of discarding	Coefficient of ε^2	RV-coefficient	Discarded variable	Coefficient of ε^2	RV-coefficient	Discarded variable
1	0.00205	0.99999	V12	0.26691	0.99835	V20
2	0.02347	0.99986	V20	0.36013	0.99778	V17
3	0.08083	0.99950	V14	0.45030	0.99722	V15
4	0.15892	0.99902	V18	0.48693	0.99699	V11
5	0.36785	0.99773	V13	0.48801	0.99699	V19
6	1.17554	0.99274	V17	0.49767	0.99693	V13
7	1.47930	0.99087	V19	1.29570	0.99200	V18
8	1.62155	0.98999	V11	1.41009	0.99130	V16
9	2.14230	0.98678	V15	3.20509	0.98022	V14
10	2.52255	0.98443	V16	5.56904	0.96562	V12

Table 4.7.c: Result of variable selection (Group III, Fatigue data)

	Before PM ($n = 46$)			After PM ($n = 54$)		
Order of discarding	Coefficient of ε^2	RV-coefficient	Discarded variable	Coefficient of ε^2	RV-coefficient	Discarded variable
1	0.01947	0.99988	V29	0.04039	0.99975	V26
2	0.02776	0.99983	V28	0.05222	0.99968	V29
3	0.17559	0.99892	V26	0.13858	0.99914	V25
4	0.21118	0.99870	V23	0.15151	0.99906	V24
5	0.22051	0.99864	V25	0.39582	0.99756	V27
6	0.36280	0.99776	V21	0.50072	0.99691	V22
7	0.45058	0.99722	V24	0.52278	0.99677	V28
8	0.76452	0.99528	V27	0.66085	0.99592	V23
9	0.77627	0.99521	V22	1.20090	0.99259	V21
10	0.96761	0.99403	V30	1.29214	0.99202	V30

There exist some other problems in dealing with categorical data. It may be possible to apply variable selection method in PCA to categorical data sets regarding them as a continuous data sets. Furthermore another criteria will be considered to select variables.

5 Principal Component Analysis Based on a Subset of Variables: Variable Selection and Sensitivity Analysis

In this section we discusses principal components (PCs) which are computed using only a selected subset of variables but represent all the variables including those not selected. If we can find such PCs which represent all the variables very well, we may say that those PCs provide a multidimensional rating scale which has high validity and is easy to apply practically. To find such PCs we borrows the ideas of Rao(1964)'s PCA of instrumental variables and Robert and Escoufier(1976)'s approach based on RV-coefficient. We shall call this type of PCA as the generalized PCA, when we need to discriminate it from the ordinary PCA.

Suppose that we have found such PCs. But there is a possibility that those PCs were obtained by chance depending heavily upon a few "influential" individuals. To provide a solution to this question we propose a method of sensitivity analysis by deriving influence functions related with the generalized PCA. We also discuss the influence of variables to the results of analysis.

5.1 Formulation

5.1.1 Formulation based on Rao(1964)'s principal component analysis of instrumental variables

To derive PCs which are obtained as linear combinations of a part of variables but represent the whole variables well we can use PCA of instrumental variables proposed by Rao (1964) by assigning the part of variables as instrumental variables. Let X be an $n \times p$ observation matrix with n individuals and p variables, where X is decomposed into an $n \times q$ submatrix X_1 and an $n \times (p-q)$ submatrix X_2, i.e., $X = (X_1, X_2)$. Denote the population and sample covariance matrices of $X = (X_1, X_2)$ by

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}. \tag{5.1}$$

Suppose we wish to make r linear combinations $Y = X_1 A$ which jointly reproduce the original p variables as well as possible in the following sense, where A is a $q \times r$ matrix.

Criterion 1. The predictive efficiency for X is maximized by using a linear predictor in terms of Y.

The formulation can be described both in the population and in the sample. Here we shall formulate in the population. It is known that the residual covariance matrix of X after subtracting the best linear predictor is expressed as

$$\Sigma_{res} = \Sigma - \Sigma_1' A (A'\Sigma_{11} A)^{-1} A'\Sigma_1, \tag{5.2}$$

where $\Sigma_1 = (\Sigma_{11}, \Sigma_{12})$. Thus, the problem becomes to minimize the residual matrix Σ_{res} or to maximize Σ_{Reg}, the covariance matrix due to regression, which is given by the second term of the right side of eq.(5.2). Note that the diagonal elements of Σ_{Reg} correspond to the so-called "communalities" in factor analysis. If it is formulated as the maximization problem of $tr(\Sigma_{Reg})$ among other possibilities, the solution is obtained as a matrix A whose columns consist of the eigenvectors associated with the largest r eigenvalues of the following eigenvalue problem:

$$[(\Sigma_{11}^2 + \Sigma_{12}\Sigma_{21}) - \lambda\Sigma_{11}]\boldsymbol{a} = 0. \tag{5.3}$$

Assume that the q eigenvalues are ordered from the largest to the smallest as $\lambda_1, \lambda_2, \ldots, \lambda_q$ and the associated eigenvectors are denoted by $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q$. Then, the solution A is expressed as

$$A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r),$$

and the maximized value of the criterion $tr(\Sigma_{Reg})$ is given by

$$\max \quad tr(\Sigma_{Reg}) = \sum_{i=1}^{r} \lambda_i. \tag{5.4}$$

This means that the proportion

$$P = \sum_{i=1}^{r} \lambda_i / tr(\Sigma) \tag{5.4$'$}$$

of the original variations is explained by the r PCs.

Just like ordinary PCA the solution of the eigenvalue problem (5.3) is not scale invariant, and therefore sometimes it is better to apply the above method to standardized data rather than raw data. In such cases the covariance matrices in the above formulation are replaced by the corresponding correlation matrices, and the proportion P indicates the average squared multiple correlation between each of the original variables and r PCs.

The above is the formulation based on the population. The sample version is obtained by replacing the population covariance matrices (Σ) by the corresponding sample covariance matrices (S) and by attaching hats ($\widehat{\ }$) to the derived quantities, i.e., $\widehat{\lambda}$, $\widehat{\boldsymbol{a}}$ and $\widehat{P}$.

5.1.2 Formulation based on Robert and Escoufier(1976)'s approach

Let $\widetilde{X}$ and $\widetilde{Y}$ be the centered matrices corresponding to X and Y, respectively. Robert and Escoufier (1976) wish to make r linear combinations $Y = X_1 A$ which approximate the original p variables as well as possible in the following sense:

Criterion 2. The configurations of X and Y are made as close as possible in the sense that

$$\left\| \frac{\widetilde{X}\widetilde{X}'}{[tr(\widetilde{X}\widetilde{X}')^2]^{1/2}} - \frac{\widetilde{Y}\widetilde{Y}'}{[tr(\widetilde{Y}\widetilde{Y}')^2]^{1/2}} \right\| \tag{5.5}$$

is minimized, where $\|\cdot\|$ indicates L_2 or Euclidean norm.

This criterion is equivalent to the following.

Criterion 2′. The RV-coefficient between X and Y, which is defined as

$$RV(X, Y) = \frac{tr(\widetilde{X}\widetilde{X}'\widetilde{Y}\widetilde{Y}')}{\{tr(\widetilde{X}\widetilde{X}')^2 \cdot tr(\widetilde{Y}\widetilde{Y}')^2\}^{1/2}} \tag{5.5′}$$

is maximized.

The solution of this formulation is again obtained by solving the sample version of the eigenvalue problem (5.3) (see, Robert and Escoufier, 1976). Precisely speaking, the coefficient matrix A in this case is given by $\widehat{A} = (\widehat{\boldsymbol{a}}_1, \ldots, \widehat{\boldsymbol{a}}_r)$, where $\widehat{\boldsymbol{a}}_i$ is the eigenvector associated with the i-th largest eigenvalue $\widehat{\lambda}_i$, normalized so that $\widehat{\boldsymbol{a}}_i' S_{11} \widehat{\boldsymbol{a}}_j = \delta_{ij}\widehat{\lambda}_i$, δ_{ij} being Kronecker's δ. The maximized $RV(X, Y)$ in (5.5′) is given by

$$RV = \left\{ \sum_{i=1}^{r} \widehat{\lambda}_i^2 / tr(S^2) \right\}^{1/2}. \tag{5.6}$$

5.2 Rotation of axes

To consider the meaning of each PC the notation of loadings, or more precisely, correlation loadings plays an important role. The correlation loadings in the present case are defined as the correlations between the original variables and derived PCs, i.e.,

$$L = corr(X, Y) = (\Sigma_D)^{-1/2} \Sigma_1' A \{(A' \Sigma_{11} A)_D\}^{-1/2}, \tag{5.7}$$

where subscript D indicates "diagonal", namely, a matrix with subscript D is a diagonal matrix having the same diagonal elements as the corresponding matrix without subscript D.

If the loading matrix L can be interpreted properly, we may apply an appropriate "rotation of axes" as in factor analysis. Here suppose that $Y = X_1A$ is rotated to $Y^* = YT = X_1AT$, where T is an $r \times r$ orthonormal matrix. Then, the loading matrix L is transformed to

$$L^* = corr(X, Y^*) = (\Sigma_D)^{-1/2}\Sigma_1' AT\{(T'A'\Sigma_{11}AT)_D\}^{-1/2}. \tag{5.8}$$

When A consists of the eigenvectors of the eigenvalue problem (5.3), it satisfies the condition that $A'\Sigma_{11}A$ is diagonal. Moreover, if they are normalized as $\boldsymbol{a}_i'\Sigma_{11}\boldsymbol{a}_j = \delta_{ij}$, $A'\Sigma_{11}A$ becomes an identity matrix. In this case the untransformed loadings L and the transformed loadings L^* are simply expressed as

$$L = (\Sigma_D)^{-1/2}\Sigma_1' A, \tag{5.9}$$

$$L^* = (\Sigma_D)^{-1/2}\Sigma_1' AT = LT, \tag{5.9$'$}$$

respectively. Thus, for letting a rotation of the loading matrix L correspond to the same rotation of the coefficient matrix A, we have to define in such a way that the length of each eigenvector is equal to unity. In the formulation based on Rao's instrumental variables, the lengths of the eigenvectors are not specified. The above property suggests that we should define

$$\boldsymbol{a}_i'\Sigma_{11}\boldsymbol{a}_i = 1, \quad i = 1, \ldots, r. \tag{5.10}$$

We can apply various analytical rotation techniques which have been developed for factor analysis to the loading matrix in order to obtain easy-to-interpret components.

5.3 Some properties

Let $\mathcal{A}$ and $\bar{\mathcal{A}}$ be a subset of the variables used for composing PCs and its complement in the set Ω of the original p variables. Denote the values of the criteria P and RV based on a subset of $\mathcal{A}$ by $P(\mathcal{A})$ and $RV(\mathcal{A})$, respectively. Then, the following properties hold.

P1°. $0 \leq P(\mathcal{A}) \leq 1$ for any $\mathcal{A}$.

P2°. $P(\mathcal{A}) \geq P(\mathcal{A}')$ for any $\mathcal{A} \supset \mathcal{A}'$.

P3°. Suppose that $\mathcal{A}'$ is made from $\mathcal{A}$ by removing completely redundant variables in the sense that the removed variables can be expressed as linear combinations of the remaining variables. Then

$$P(\mathcal{A}') = P(\mathcal{A}).$$

R1°. $0 \leq RV(\mathcal{A}) \leq 1$ for any $\mathcal{A}$.

R2°. $RV(\mathcal{A}) \geq RV(\mathcal{A}')$ for any $\mathcal{A} \supset \mathcal{A}'$.

Properties **P1°** $\sim$ **P3°** can be proved using the theory of linear models (see, Appendix A.2). Property **R1°** is obvious from the definition of RV-coefficient, and property **R2°** can be shown based on the fact that any A for variables in subset $\mathcal{A}'$ is a member of the set of all possible A for variables in subset $\mathcal{A}$ with zero elements for the variables in subset $\mathcal{A} - \mathcal{A}'$.

5.4 Variable selection procedure

It is desirable that we can find PCs which are based on a small number of variables but represent all the variables very well. Obviously we can find the best subset for such PCs, if we try all possible subsets. But it is usually impractical to do so, because it requires very high computing cost. Therefore, as a practical strategy we propose the following two-stage procedure. This procedure is described on the basis of Criterion 1, but it can be easily modified to the procedure based on Criterion 2 by replacing the proportion P by RV.

A. Initial fixed-variable stage

Step A-1 Compute the covariance matrix of the whole variables X and assign q variables to subset $\mathcal{A}$, which consists of the variables X_1 to be used for composing PCs, and the remaining $p - q$ variables to subset $\bar{\mathcal{A}}$. Usually assign all variables to subset $\mathcal{A}$, i.e., $q = p$.

Step A-2 Solving the eigenvalue problem (5.3), obtain the eigenvalues $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_q$ ($\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_q$) and the associated eigenvectors $\widehat{\boldsymbol{a}}_1, \ldots, \widehat{\boldsymbol{a}}_q$.

Step A-3 Looking carefully at the eigenvalues and the cumulative proportions, determine the number r of PCs to be used. An appropriate rotation technique may be applied to study whether meaningful factors are obtained.

B. Variable selection stage (backward method)

Step B-1 Based on the results of Stage A, start with a preassigned subset $\mathcal{A}$ of q variables and the fixed number of PCs r.

Stage B-2 Remove each one among the q variables in $\mathcal{A}$ in turn, and solve q eigenvalue problems of $q-1$ variables. Find the best subset of size $q-1$ in which the proportion P is the largest, and actually remove the corresponding variable. Put $q := q-1$.

Step B-3 If both the proportion P and the number of variables in $\mathcal{A}$ are larger than preassigned values, go back to Step B-2. Otherwise stop the procedure.

5.5 Sensitivity analysis

As shown in section 5.1, PCs are obtained as linear combinations whose coefficients are given by the eigenvectors associated with the largest r eigenvalues of the generalized eigenvalue problem (5.3). For the purpose of sensitivity analysis we need to evaluate the change of the solution of this eigenvalue problem corresponding to a small perturbation introduced to the data or the model. The former treats the influence of individuals and the latter treats the influence of variables on the results of analysis.

5.5.1 Influence of individuals

For the sake of simplicity we shall denote the influence function by attaching superscript (1) and discriminate the EIF and SIF by attaching $\hat{}$(hat) and $\tilde{}$(tilde) to the influence function. For example, $\theta^{(1)}$, $\widehat{\theta}^{(1)}$ and $\widetilde{\theta}^{(1)}$ indicate the theoretical, empirical and sample influence functions for θ.

Using the lemma in section 2.3 influence functions are obtained for quantities characterizing the results of the PCA as functions of the influence function for the covariance matrix. It is well known (see, e.g., Critchley, 1985) that the TIF for the covariance matrix is given by

$$\Sigma^{(1)} = (\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)' - \Sigma \tag{5.11}$$

and the corresponding EIF for sample point $\boldsymbol{x}_i$ is expressed as

$$S^{(1)} = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' - S \tag{5.11$'$}$$

where the sample covariance matrix S is defined as the sum of squares and products matrix divided by n.

(a) Influence functions for eigenvalues, proportions and RV-coefficients

$$\widehat{\lambda}_j^{(1)} = \widehat{\boldsymbol{a}}_j'(C^{(1)} - \widehat{\lambda}_j D^{(1)})\widehat{\boldsymbol{a}}_j, \quad j = 1, \ldots, q \tag{5.12}$$

where

$$C^{(1)} = S_{11}^{(1)} S_{11} + S_{11} S_{11}^{(1)} + S_{12}^{(1)} S_{21} + S_{12} S_{21}^{(1)}, \tag{5.13}$$

$$D^{(1)} = S_{11}^{(1)}. \tag{5.14}$$

$$\begin{aligned} P^{(1)} &= \left[\sum_{j=1}^{r} \widehat{\lambda}_j / tr(S)\right]^{(1)} \\ &= \sum_{j=1}^{r} \widehat{\lambda}_j^{(1)} / tr(S) - \sum_{j=1}^{r} \widehat{\lambda}_j tr(S^{(1)})/(tr(S))^2, \end{aligned} \tag{5.15}$$

$$\begin{aligned} RV^{(1)} &= \left[\left\{\sum_{j=1}^{r} \widehat{\lambda}_j^2 / tr(S^2)\right\}^{1/2}\right]^{(1)} \\ &= \left\{\sum_{j=1}^{r} \widehat{\lambda}_j^2 / tr(S^2)\right\}^{-1/2} \left\{\sum_{j=1}^{r} \widehat{\lambda}_j \widehat{\lambda}_j^{(1)} / tr(S^2) - \sum_{j=1}^{r} \widehat{\lambda}_j^2 tr(SS^{(1)})/(tr(S^2))^2\right\}. \end{aligned} \tag{5.16}$$

(b) Influence functions for coefficient vectors

$$\widehat{\boldsymbol{a}}_j^{(1)} = \sum_{k \neq j} (\widehat{\lambda}_j - \widehat{\lambda}_k)^{-1} \{\widehat{\boldsymbol{a}}_j' (C^{(1)} - \widehat{\lambda}_j D^{(1)}) \widehat{\boldsymbol{a}}_k\} \widehat{\boldsymbol{a}}_k - (1/2)(\widehat{\boldsymbol{a}}_j' D^{(1)} \widehat{\boldsymbol{a}}_j) \widehat{\boldsymbol{a}}_j, \quad j = 1, \ldots, q. \tag{5.17}$$

(c) Influence function for the configuration of loadings

$$\begin{aligned} (\widehat{L}\widehat{L}')^{(1)} &= \{S_D^{-1/2} S_1' \widehat{A}\widehat{A}' S_1 S_D^{-1/2}\}^{(1)} \\ &= E^{(1)} \widehat{A}\widehat{A}' E' + E(\widehat{A}\widehat{A}')^{(1)} E' + E\widehat{A}\widehat{A}' E^{(1)'} \end{aligned} \tag{5.18}$$

where

$$E = S_D^{-1/2} S_1', \tag{5.19}$$

$$E^{(1)} = -(1/2) S_D^{-3/2} S_D^{(1)} S_1' + S_D^{-1/2} S_1^{(1)'}, \tag{5.20}$$

$$\begin{aligned} (\widehat{A}\widehat{A}')^{(1)} = &- \sum_{j=1}^{r} \sum_{k=1}^{r} (\widehat{\boldsymbol{a}}_j' D^{(1)} \widehat{\boldsymbol{a}}_k) \widehat{\boldsymbol{a}}_j \widehat{\boldsymbol{a}}_k' \\ &+ \sum_{j=1}^{r} \sum_{k=r+1}^{q} (\widehat{\lambda}_j - \widehat{\lambda}_k)^{-1} \{\widehat{\boldsymbol{a}}_j' (C^{(1)} - \widehat{\lambda}_j D^{(1)}) \widehat{\boldsymbol{a}}_k\} (\widehat{\boldsymbol{a}}_j \widehat{\boldsymbol{a}}_k' + \widehat{\boldsymbol{a}}_k \widehat{\boldsymbol{a}}_j'). \end{aligned} \tag{5.21}$$

Those influence functions indicate the measures of influence on (a) the amount of variation explained the j-th PC, (b) the coefficient vector for the j-th PC, and (c) the configuration of loadings which plays an important role for the interpretation of the obtained PCs, respectively.

The above are formulas for our generalized PCA based on covariance matrix. But as mentioned in section 5.1.1 our procedure is sometimes applied to the correlation matrix R instead of the covariance matrix S. In such cases S and $S^{(1)}$ in (a) through (c) should be replaced by R and $R^{(1)}$, respectively, where $R^{(1)}$ is obtained as

$$R^{(1)} = S_D^{-1/2} S^{(1)} S_D^{-1/2} - (1/2) S_D^{-1} S_D^{(1)} R - (1/2) R S_D^{(1)} S_D^{-1}. \tag{5.22}$$

5.5.2 Influence of variables

To evaluate the influence of variables we shall perturb slightly the weight of a specified variable from 1 to $1-\varepsilon$ without changing the other weights and evaluate the effect on the result of analysis.

Suppose we wish to evaluate the influence of the j-th variable in subset $\mathcal{A}$ of q variables and perturb the weight of the variable as stated above. Then, the covariance matrices change as follows:

$$\Sigma_{11} \longrightarrow \Sigma_{11} - \varepsilon(J_j \Sigma_{11} + \Sigma_{11} J_j) + O(\varepsilon^2), \tag{5.23}$$

$$\Sigma_{12} \longrightarrow \Sigma_{12} - \varepsilon J_j \Sigma_{12}, \tag{5.24}$$

$$\Sigma_{21} \longrightarrow \Sigma_{21} - \varepsilon \Sigma_{21} J_j, \tag{5.25}$$

where J_j indicates a $q \times q$ diagonal matrix with unity in the j-th element and zeros in the other elements. Hence, if we express the generalized eigenvalue problem (5.3) as $(C - \lambda D)\boldsymbol{a} = 0$, C and D change to $C + \varepsilon C^{(1)} + O(\varepsilon^2)$ and $D + \varepsilon D^{(1)} + O(\varepsilon^2)$, respectively, where

$$C^{(1)} = -J_j C - C J_j - 2\Sigma_{11} J_j \Sigma_{11}, \tag{5.26}$$

$$D^{(1)} = -J_j \Sigma_{11} - \Sigma_{11} J_j. \tag{5.27}$$

Next, if we wish to evaluate the influence of the j-th variable in subset $\bar{\mathcal{A}}$ of $p-q$ variables and perturb the weight of this variable. Then, the covariance matrices change as $\Sigma_{11} \to \Sigma_{11}$, $\Sigma_{12} \to \Sigma_{12} - \varepsilon \Sigma_{12} K_j$ and $\Sigma_{21} \to \Sigma_{21} - \varepsilon K_j \Sigma_{21}$, where K_j indicates a $(p-q) \times (p-q)$ diagonal matrix with unity in the j-th element and zeros in the other elements. In this case, C and D change to $C + \varepsilon C^{(1)}$ and $D + \varepsilon D^{(1)}$, respectively, where

$$C^{(1)} = -2\Sigma_{12} K_j \Sigma_{21}, \tag{5.28}$$

$$D^{(1)} = 0. \tag{5.29}$$

Table 5.1: Process of removing variables based on P (Alate data)

Step	q	Removed variable	P	P_q
0	19	—	0.85270	1.00000
1	18	V13	0.85268	0.99970
2	17	V12	0.85254	0.99818
3	16	V7	0.85242	0.99678
4	15	V3	0.85225	0.99457
5	14	V15	0.85197	0.98834
6	13	V1	0.85154	0.98302
7	12	V9	0.85107	0.97263
8	11	V8	0.85057	0.96609
9	10	V2	0.85022	0.96154
10	9	V10	0.84931	0.95232
11	8	V4	0.84800	0.94794
12	7	V16	0.84655	0.94153
13	6	V11	0.84287	0.90106
14	5	V6	0.83899	0.88817
15	4	V19	0.83459	0.86881
16	3	V17	0.82743	0.85316
17	2	V18	0.79525	0.79525

Based on the lemma in the section 2.3 we can easily compute the differential coefficients of the eigenvalues $\lambda_1, \ldots, \lambda_r$ and of the related quantities P and RV, and use these differential coefficients for the evaluation of the influence of variables. For simplicity we denote these differential coefficients by attaching superscript (1) as in the case of influence functions.

5.6 Numerical examples

5.6.1 Alate adelges data

As the first numerical example we analyzed a data set of alate adelges (winged aphids), which was analyzed originally by Jeffers (1967) using ordinary PCA and later by some authors including Jolliffe (1986) and Krzanowski (1987a, b) using PCA with variable selection functions. We applied our generalized PCA based on correlation matrix to the data given in Krzanowski (1987a). The data set consists of 40 individuals and 19 variables (Appendix B.5).

At the first stage ordinary PCA was applied to the standardized data set and the

Table 5.2: Coefficients for PCs and correlation loadings (Alate data)

Variable	Coefficients		Loadings		R^2		
	I	II	I	II	G.PCA* (9 var.)	O.PCA** (19 var.)	O.PCA** (9 var.)
V1	—	—	0.93096	-0.02305	0.867214	0.872455	0.824270
V2	—	—	0.95652	-0.10425	0.925806	0.933979	0.888962
V3	—	—	0.96424	-0.04842	0.932109	0.939231	0.909786
V4	0.33089	-0.08076	0.96752	-0.13799	0.955135	0.950924	0.931712
V5	0.08338	0.44547	0.60449	0.62352	0.754182	0.752580	0.795746
V6	0.25511	0.12090	0.89412	0.27049	0.872614	0.869308	0.868889
V7	—	—	0.93944	0.24381	0.941994	0.951654	0.926786
V8	—	—	0.85792	-0.35358	0.861052	0.876022	0.821168
V9	—	—	0.87722	-0.06696	0.774002	0.788993	0.734576
V10	—	—	0.91345	0.03403	0.835547	0.857939	0.811928
V11	-0.10380	0.21218	-0.48701	0.31711	0.337739	0.336011	0.399714
V12	—	—	0.97215	-0.01747	0.945387	0.947376	0.894147
V13	—	—	0.97998	-0.04530	0.962411	0.964052	0.898638
V14	0.84672	-0.23918	0.97363	-0.10377	0.958721	0.954642	0.884840
V15	—	—	0.93556	0.01200	0.875414	0.880678	0.843413
V16	0.16178	0.44342	0.75077	0.60520	0.929931	0.927411	0.918809
V17	0.11956	0.49225	0.40895	0.83985	0.872580	0.870261	0.883324
V18	-0.14005	0.37386	-0.69968	0.54363	0.785080	0.781275	0.843330
V19	0.17517	-0.31543	0.74711	-0.43803	0.750045	0.746567	0.786389
Average	—				0.849314	0.852703	0.835075

Note. * G.PCA: Generalized PCA
**O.PCA: Ordinary PCA

same results was obtained as in Jeffers (1967). The eigenvalues and cumulative proportions were $\widehat{\lambda}_1 = 13.838(72.83\%) > \widehat{\lambda}_2 = 2.363(85.27\%) > \widehat{\lambda}_3 = 0.748(89.21\%) > \widehat{\lambda}_4 = 0.505(91.86\%) > \cdots$ in order of magnitude, and on the basis of these values it was decided to extract two PCs.

Then the generalized PCA was applied using the backward procedure based on Criterion 1. The process of removing variables is shown in Table 5.1. In the last two columns, P indicates the proportion given by the sample version of (5.5) and P_q indicates the proportion defined by the same equation excepting r replaced by q, namely, the proportion obtained by using all PCs. This table shows that the proportion P (in this case the average squared multiple correlation) changes very slightly until step 10, in which the number of variables is 9. This means that 10 among 19 variables are almost redundant for composing PCs to be used to reproduce the original variables.

Table 5.2 shows the coefficients for PCs and the correlation loadings in Step 10. The

Table 5.3: Comparison of $\bar{R}^2$ of 4 variables selected by various methods (Alate data)

Method	Selected variables				$\bar{R}^2_g$	$\bar{R}^2_o$
Criterion 1 (P)	5	14	17	18	0.8346	0.7975
Criterion 2 (Robert & Escoufier's RV)	5	6	14	19	0.8234	0.8055
Jolliffe's **B2**	5	8	11	14	0.8160	0.7886
Jolliffe's **B4**	5	11	13	17	0.8321	0.7886
McCabe	5	9	11	18	0.7547	0.7236
Krzanowski	5	12	14	18	0.8309	0.8150
SP in chapter 3	9	11	17	19	0.7675	0.7573
SE in chapter 3	5	8	17	18	0.7893	0.7698

last three columns (R^2 part) indicate how well each variable is reproduced using the generalized PCs based on the 9 variables, the ordinary PCs of all the 19 variables and the ordinary PCs of the same 9 variables, respectively. The generalized PCs based on the 9 variables can reproduce all the 19 variables almost equally well as the ordinary PCs of the 19 variables. The ordinary PCs of the same 9 variables can also reproduce the 19 variables well, but the degrees of reproducibility are a little inferior to those of the generalized PCs. In particular it is noticed that the ordinary PCs reproduce some of the variables composing PCs better than the generalized PCs, but they do not reproduce so well the removed variables.

Moreover we shall try to compare our result with the results obtained by using sets of variables selected by other authors' methods. Table 5.3 shows the values of the average of R^2 computed for each subset of 4 variables which was obtained by a method indicated in the first column. $\bar{R}^2_g$ is computed using the generalized PCs of $X = (X_1, X_2)$ where X_1 consists of the selected 4 variables and and X_2 the other ones. $\bar{R}^2_o$ is computed the ordinary PCs of $X = X_1$ which contains selected 4 variables. Note that the values of $\bar{R}^2_g$ in the row of Criterion 2 part was recomputed based on Criterion 1 using the 4 variables obtained by their criterion (Criterion 2 in our study). It illustrates clearly that the generalized PCs obtained by Criterion 1 gives the largest value of $\bar{R}^2$ among others. Figure 5.1 is the plot of these $\bar{R}^2$ where the number of variables is changing from 19 to 4 successively. This figure shows that the generalized PCs always represents all the original variables well.

Next the generalized PCA was applied using the backward procedure based on Criterion 2. The process of removing variables is shown in Table 5.4. The last two columns of RV and RV_q have similar meanings as P and P_q in Table 5.1. The coefficient RV changes

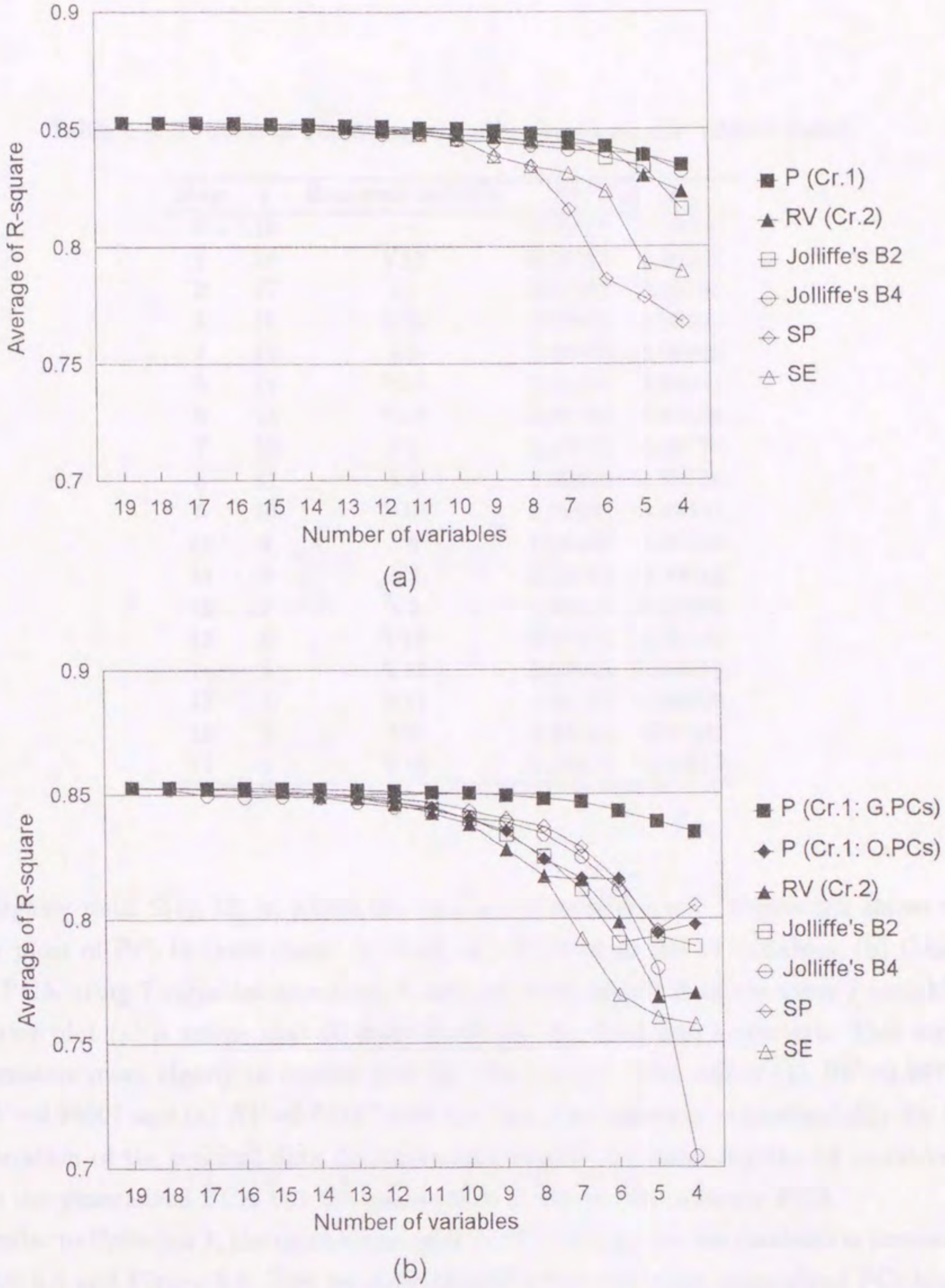


Figure 5.1: Plot of change of $\bar{R}^2$ (a) based on the generalized PCs ($\bar{R}_g^2$) and (b) based on the ordinary PCs ($\bar{R}_o^2$). Note that $\bar{R}_g^2$ of Criterion 1 is overplotted in (b). (Alate data)

Table 5.4: Process of removing variables based on RV (Alate data)

Step	q	Removed variable	RV	RV_q
0	19	—	0.99726	1.00000
1	18	V13	0.99723	0.99997
2	17	V7	0.99707	0.99981
3	16	V12	0.99692	0.99965
4	15	V3	0.99670	0.99942
5	14	V15	0.99634	0.99901
6	13	V18	0.99583	0.99836
7	12	V1	0.99521	0.99770
8	11	V4	0.99452	0.99700
9	10	V16	0.99388	0.99631
10	9	V9	0.99300	0.99530
11	8	V8	0.99219	0.99443
12	7	V2	0.99107	0.99329
13	6	V10	0.98925	0.99140
14	5	V17	0.98622	0.98818
15	4	V11	0.98163	0.98223
16	3	V6	0.97554	0.97607
17	2	V19	0.96813	0.96813

very slightly until Step 12, in which the number of variables is 7. Figure 5.2 shows the scatter plots of PCs in three cases: (a) Ordinary PCA of all the 19 variables, (b) Generalized PCA using 7 variables as subset $\mathcal{A}$, and (c) Ordinary PCA of the same 7 variables. In scatter plot (a) it seems that 40 individuals are classified into 4 clusters. This structure remains more clearly in scatter plot (b) than in (c). The values (a) RV=0.99726, (b) RV=0.99107 and (c) RV=0.94417 indicate that the degrees of reproducibility for the configuration of the original data decreases only slightly by removing the 12 variables if we use the generalized PCA but decreases much if we use the ordinary PCA.

Similar to Criterion 1, the comparison of RVs obtained by various methods is presented in Table 5.5 and Figure 5.3. Now we show the RVs only using the generalized PCs based on Criterion 2. The table and figure indicate that the set of variables based on Criterion 2 has the highest RV among others.

5.6.2 Mild disturbance of consciousness (MDOC) data

These data were originally analyzed by Sano et al (1977) using factor analysis, and later by Tanaka and Kodake (1981) and Tanaka (1983) using principal factor analysis with variable

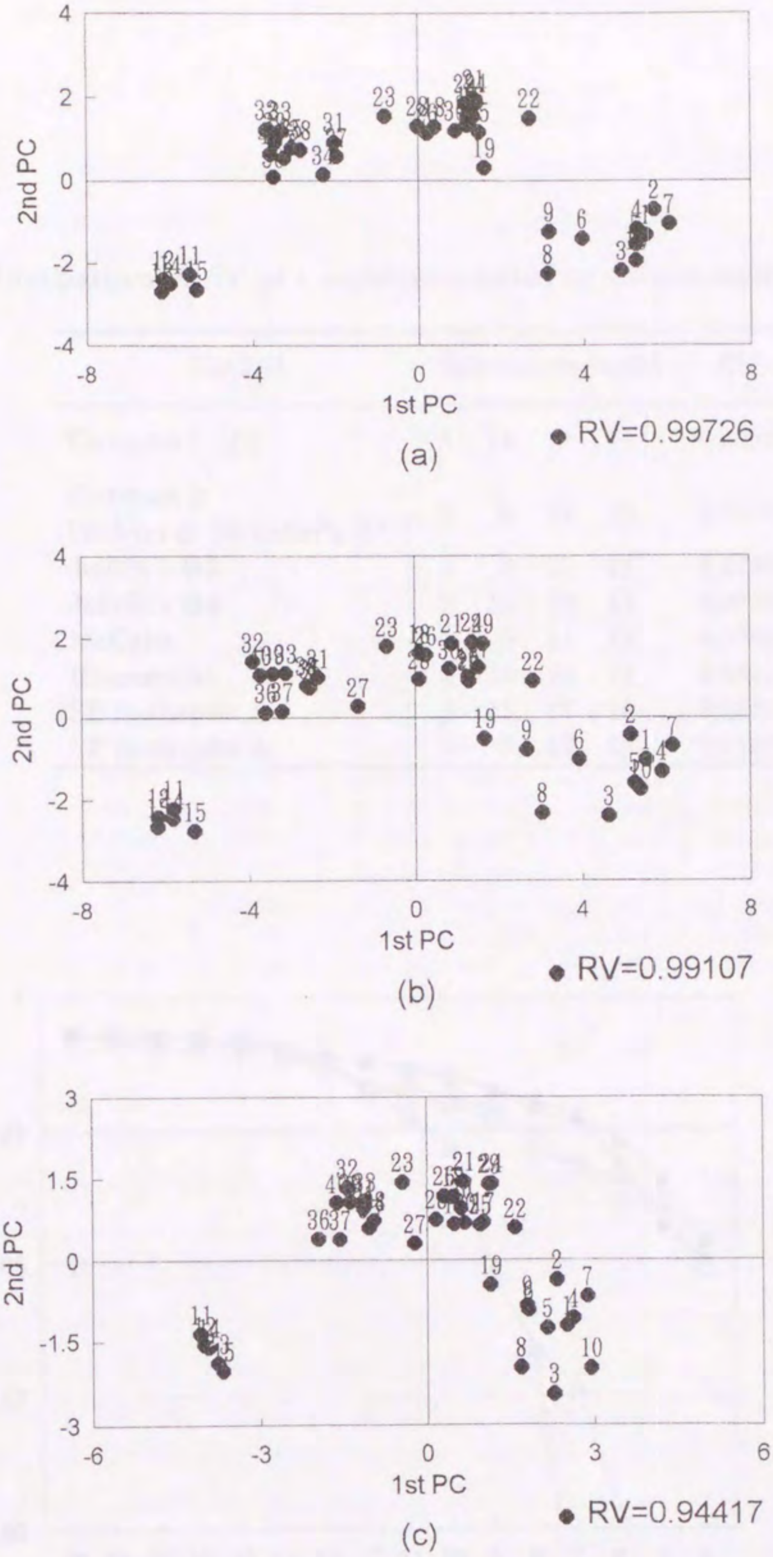


Figure 5.2: Scatter plots of PCs: (a) Ordinary PCA of all the 19 variables; (b) Generalized PCA of the selected 7 variables; (c) Ordinary PCA of the same selected 7 variables

Table 5.5: Comparison of RV of 4 variables selected by various methods (Alate data)

Method	Selected variables				RV
Criterion 1 (P)	5	14	17	18	0.9802
Criterion 2 (Robert & Escoufier's RV)	5	6	14	19	0.9816
Jolliffe's B2	5	8	11	14	0.9796
Jolliffe's B4	5	11	13	17	0.9815
McCabe	5	9	11	18	0.8788
Krzanowski	5	12	14	18	0.9812
SP in chapter 3	9	11	17	19	0.8951
SE in chapter 3	5	8	17	18	0.9187

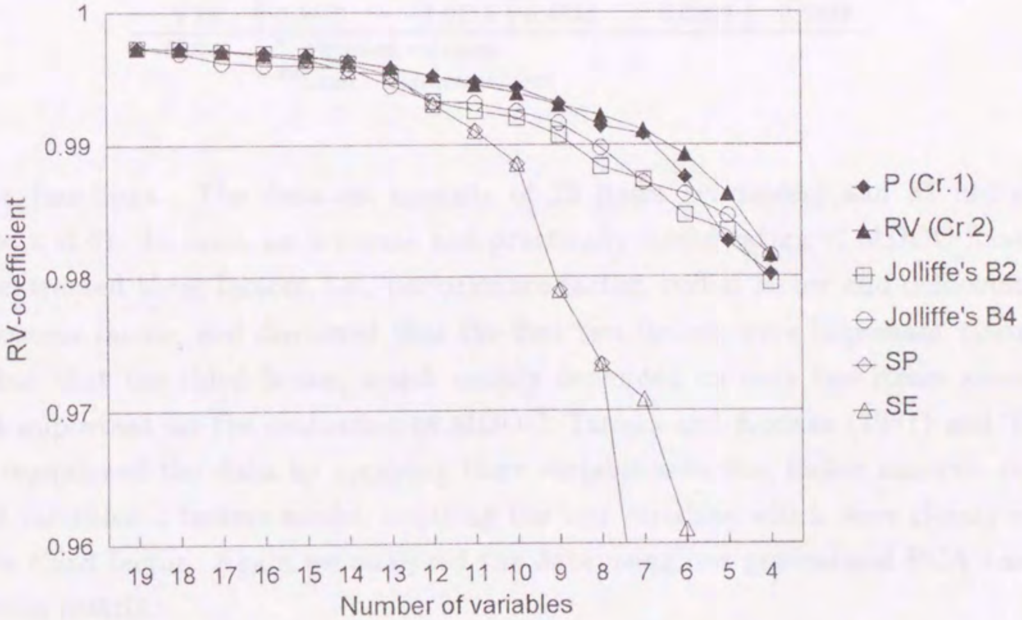


Figure 5.3: Plot of change of RV based on the generalized PCs (Alate data)

Table 5.6: Loadings obtained by ordinary PCs of 23 variables (MDOC data)

Variable	Unrotated loadings		Rotated* loadings		Com.**
	I	II	I	II	
V1	0.7199	-0.3947	0.7631	-0.3028	0.6740
V2	0.7924	-0.1204	0.8012	-0.0216	0.6423
V3	0.8453	-0.2829	0.8738	-0.1763	0.7946
V4	0.7863	0.2817	0.7455	0.3766	0.6977
V5	0.7419	0.3968	0.6872	0.4854	0.7079
V6	0.5959	0.4945	0.5303	0.5644	0.5997
V7	0.6951	0.1446	0.6719	0.2293	0.5041
V8	0.8379	0.1809	0.8091	0.2830	0.7347
V9	0.8060	0.1094	0.7863	0.2081	0.6615
V10	0.8579	-0.0037	0.8518	0.1023	0.7359
V11	0.7730	0.3188	0.7277	0.4118	0.6992
V12	0.8099	-0.2221	0.8311	-0.1204	0.7053
V13	0.8298	0.0493	0.8173	0.1514	0.6909
V14	0.7652	0.3453	0.7167	0.4371	0.7047
V15	0.8787	-0.1192	0.8867	-0.0098	0.7864
V16	0.7896	-0.2443	0.8137	-0.1449	0.6831
V17	0.8969	-0.2050	0.9153	-0.0927	0.8464
V18	0.8633	-0.2100	0.8826	-0.1017	0.7894
V19	0.8770	-0.2108	0.8964	-0.1009	0.8136
V20	0.8586	-0.2709	0.8855	-0.1627	0.8105
V21	0.4586	-0.2244	0.4828	-0.1660	0.2607
V22	0.7026	0.4239	0.6449	0.5075	0.6734
V23	0.4897	-0.0310	0.4898	0.0297	0.2407

Note. * Varimax rotation
**Com.: Communalities

selection functions. The data set consists of 25 items (variables) and 87 individuals (Appendix B.6). To make an accurate and practically useful rating of MDOC Sano et al (1977) extracted three factors, i.e., performance factor, verbal factor and deformation of consciousness factor, and discussed that the first two factors were important among the three, but that the third factor, which mainly depended on only two items among 25, was not important for the evaluation of MDOC. Tanaka and Kodake (1981) and Tanaka (1983) reanalyzed the data by applying their variable selection factor analysis starting from 23 variables–2 factors model, omitting the two variables which were closely related with the third factor. Again we analyzed the data using our generalized PCA based on correlation matrix.

At first ordinary PCA was applied to the correlation matrix. The eigenvalues and the cumulative proportions were $\widehat{\lambda}_1 = 13.878$ (60.34%) $> \widehat{\lambda}_2 = 1.579$ (67.20%) $> \widehat{\lambda}_3 = 0.9580$ (71.37%) $> \widehat{\lambda}_4 = 0.8456$ (75.05%) $> \widehat{\lambda}_5 = 0.7432$ (78.28%) $> \cdots$. Looking at these values it was decided to extract two PCs. The unrotated loadings and the varimax rotated

Table 5.7: Process of removing variables (MDOC data)

Step	q	Removed variable	P	P_q
0	23	—	0.67203	1.00000
1	22	V10	0.67159	0.99160
2	21	V23	0.67118	0.96420
3	20	V18	0.67077	0.95589
4	19	V15	0.67044	0.94760
5	18	V2	0.67000	0.93451
6	17	V8	0.66946	0.92292
7	16	V13	0.66890	0.90986
8	15	V20	0.66798	0.90074
9	14	V14	0.66687	0.88596
10	13	V21	0.66556	0.85592
11	12	V4	0.66424	0.84251
12	11	V19	0.66232	0.83146
13	10	V7	0.66036	0.80734
14	9	V5	0.65771	0.78944
15	8	V1	0.65465	0.77054
16	7	V9	0.65008	0.74329
17	6	V12	0.64438	0.72242
18	5	V11	0.63303	0.69097
19	4	V16	0.61901	0.65788
20	3	V6	0.59774	0.61215
21	2	V3	0.56865	0.56865

loading are given in Table 5.6. Note that the patterns of the varimax-rotated loadings are very similar to those obtained by the iterative principal factor analysis (see, Tanaka and Kodake, 1981, Table 4). Then the generalized PCA was applied using the backward procedure based on Criterion 1 and it was found that the loss of information was almost negligible by removing 10 variables among 23.

Table 5.8 shows the coefficients for 13 ($= 23 - 10$) variables, the loadings for all the variables and the degrees of reproductivity of variables with the generalized PCs and the ordinary PCs of all the 23 variables. This table suggests that (a) we can use the generalized PCs based on 13 variables instead of the ordinary PCs based on all the 23 variables as a two-dimensional scale, because the loadings are very similar for both sets of PCs, and (b) the loss of information is small by removing 10 variables, because this removal does not cause much decrease of R^2. In both analyses the reproducibility is very low for variables No.21 and No.23. It seems that these two variables are somewhat different from the other variables and it may be better to analyze these variables separately from the analysis of the remaining variables.

Table 5.8: Coefficients for PCs and correlation loadings (MDOC data)

Variable	Coefficients		Loadings		R^2	
	I	II	I	II	G.PCA* (13 var.)	O.PCA** (23 var.)
V1	0.08292	-0.28342	0.72210	-0.41764	0.695848	0.674034
V2	—	—	0.78937	-0.10542	0.634224	0.642309
V3	0.12143	-0.30323	0.84798	-0.29272	0.804763	0.794582
V4	0.06906	0.20113	0.78915	0.28369	0.703241	0.697668
V5	0.08825	0.21299	0.74440	0.41706	0.728074	0.707868
V6	0.05529	0.37984	0.59798	0.50911	0.616774	0.599728
V7	0.07838	0.10828	0.69755	0.14471	0.507520	0.504093
V8	—	—	0.83558	0.14926	0.720470	0.734699
V9	0.11307	0.08690	0.80869	0.10561	0.665133	0.661533
V10	—	—	0.84307	-0.00683	0.710807	0.735923
V11	0.08176	0.29802	0.77584	0.31362	0.700279	0.699147
V12	0.09501	-0.22126	0.81236	-0.22977	0.712726	0.705278
V13	—	—	0.81719	0.01588	0.668043	0.690969
V14	—	—	0.75893	0.27498	0.651582	0.704720
V15	—	—	0.87117	-0.10609	0.770188	0.786388
V16	0.09590	-0.21173	0.79220	-0.26293	0.696712	0.683112
V17	0.15343	-0.22256	0.89945	-0.20763	0.852122	0.846354
V18	—	—	0.86154	-0.17064	0.771374	0.789343
V19	0.12523	-0.17213	0.87977	-0.22544	0.824815	0.813647
V20	—	—	0.84980	-0.23080	0.775423	0.810489
V21	—	—	0.42942	-0.12516	0.200070	0.260715
V22	0.10356	0.33961	0.70508	0.42332	0.676342	0.673419
V23	—	—	0.47034	-0.01100	0.221342	0.240738
Average	—				0.665560	0.672033

Note. * G.PCA: Generalized PCA
**O.PCA: Ordinary PCA

5.6.3 Sensitivity analysis of the alate adelges data

In section 5.6.1 it was found that the generalized PCs based on the 9 variables gave almost the same information as the ordinary PCs of all the 19 variables. Then the sensitivity analysis was performed to examine whether the obtained results depended heavily upon a few individuals and/or a few variables.

Firstly the influence of individuals was studied using the influence functions derived in section 5.5.1. Figure 5.4 shows the index plots of $\widehat{P}^{(1)}$ and $||(\widehat{L}\widehat{L}')^{(1)}||$ for 40 individuals. It seems that there are no individuals which are singly influential. Then, as the next stage PCA was applied to the EIF vectors $\{(\widehat{\lambda}_1^{(1)}, \widehat{\lambda}_2^{(1)}, \widehat{\boldsymbol{a}}_{1i}^{(1)'}, \widehat{\boldsymbol{a}}_{2i}^{(1)'}), i = 1, \ldots, n\ \}$. Figure 5.5 shows the scatter plot of the first two PCs, which explain 92.17% (1st PC: 83.82%, 2nd PC: 8.35%) of all the variations of the EIF. In Figure 5.5 we can observe that individuals No.11 – 14 form a cluster of points which are located far from the origin. The

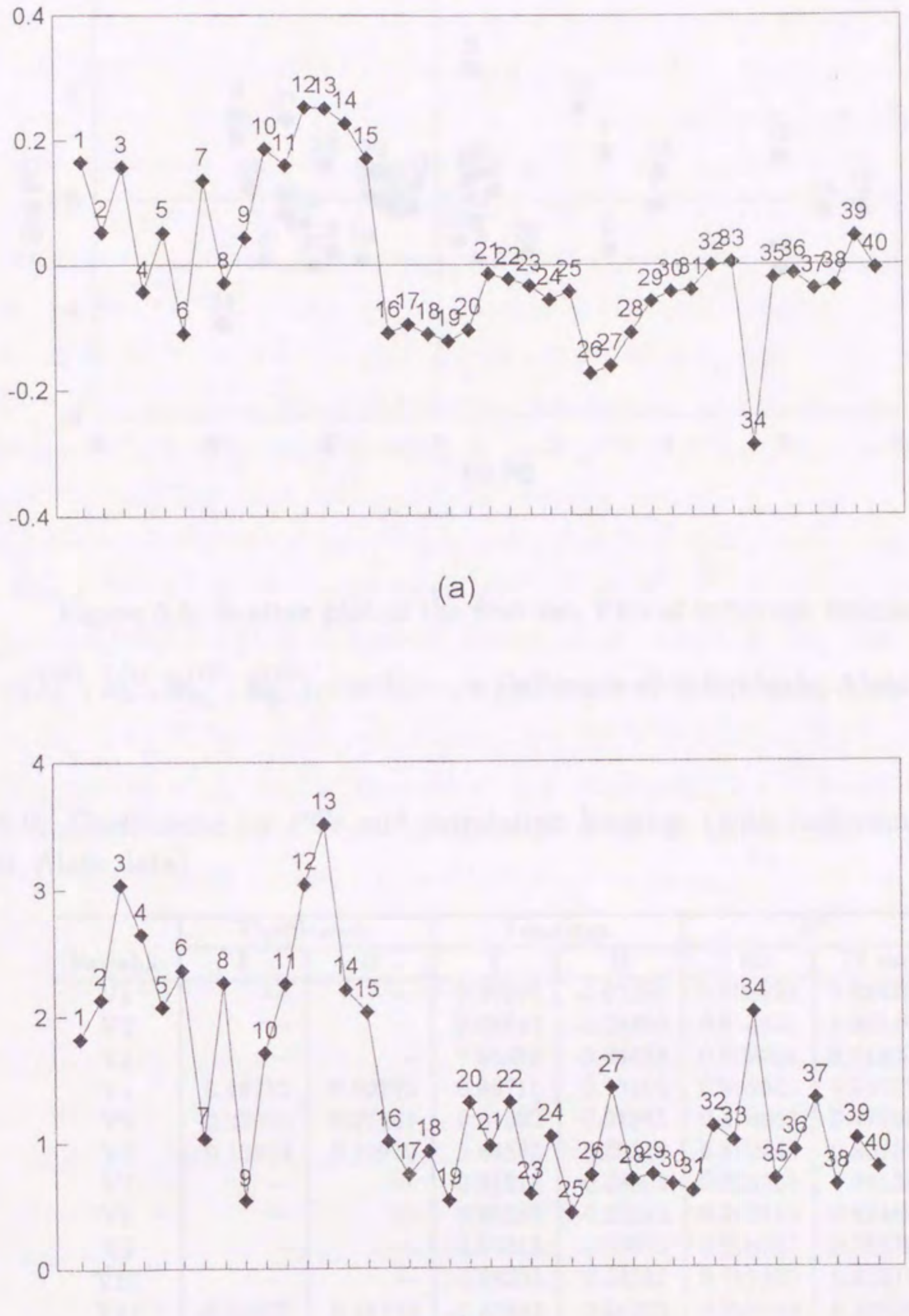


Figure 5.4: Index plots of (a) $\widehat{P}^{(1)}$ and (b) $\|(\widehat{L}\widehat{L}')^{(1)}\|$ (influence of individuals, Alate data)

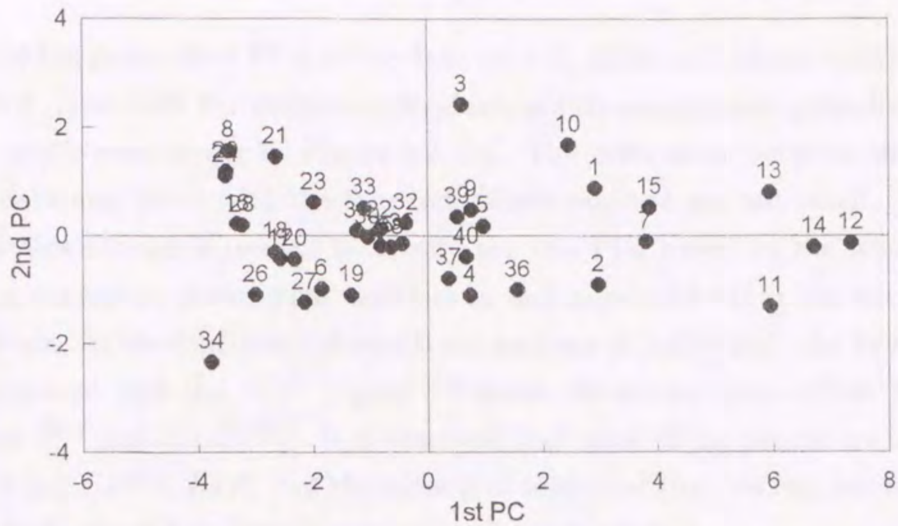


Figure 5.5: Scatter plot of the first two PCs of influence functions, $(\hat{\lambda}_1^{(1)}, \hat{\lambda}_2^{(1)}, \hat{\boldsymbol{a}}_{1i}^{(1)'}, \hat{\boldsymbol{a}}_{2i}^{(1)'})$, $i = 1, \ldots, n$ (influence of individuals, Alate data)

Table 5.9: Coefficients for PCs and correlation loadings (with individuals No.11 – 14 omitted, Alate data)

Variable	Coefficients		Loadings		R^2	
	I	II	I	II	9 var.	19 var.
V1	—	—	0.90996	-0.02208	0.828523	0.834707
V2	—	—	0.95543	-0.05456	0.915823	0.925160
V3	—	—	0.95283	-0.04478	0.909884	0.918903
V4	0.18712	0.00743	0.96715	-0.10122	0.945631	0.940029
V5	0.02694	0.27481	0.31032	0.61381	0.473063	0.471638
V6	0.13804	0.11919	0.84886	0.30312	0.812440	0.807574
V7	—	—	0.91916	0.28004	0.923268	0.941398
V8	—	—	0.90583	-0.21281	0.865810	0.873899
V9	—	—	0.85618	-0.05623	0.736207	0.755792
V10	—	—	0.88295	0.04551	0.781670	0.813118
V11	-0.06079	0.15744	-0.46553	0.36222	0.347919	0.346787
V12	—	—	0.96776	0.04898	0.938969	0.941227
V13	—	—	0.98187	0.02747	0.964811	0.966416
V14	0.49300	-0.04100	0.98126	-0.02992	0.963773	0.958061
V15	—	—	0.91446	0.02343	0.836783	0.844562
V16	0.07037	0.33077	0.61628	0.68488	0.848867	0.845429
V17	0.03619	0.35971	0.06816	0.87754	0.774726	0.772712
V18	-0.07957	0.23751	-0.71921	0.51519	0.782686	0.779455
V19	0.10413	-0.18292	0.75257	-0.39977	0.726184	0.722776
Average	—				0.809318	0.813666

results of the generalized PCA of the data set with those individuals omitted are shown in Table 5.9. Note that the omitted individuals are contained among the five points dotted at the south-west corner in Figure 5.2 (b). The differences between the results of the whole data and those with the four individuals omitted are not small. It seems that a considerable change is needed to modify the two PCs based on the whole data so that they can reproduce the original variables as well as possible using the data without those individuals. As the final stage of sensitivity analysis of individuals the SIF was computed and compared with the EIF. Figure 5.6 shows the scatter plots of the SIF against the EIF for $\widehat{P}^{(1)}$ and $||(\widehat{L}\widehat{L}')^{(1)}||$. It is observed that most of the points are located near the straight line $SIF = EIF$, and therefore it is suggested that we can use the EIF instead of the SIF, which has clear "leave-one-out" interpretation.

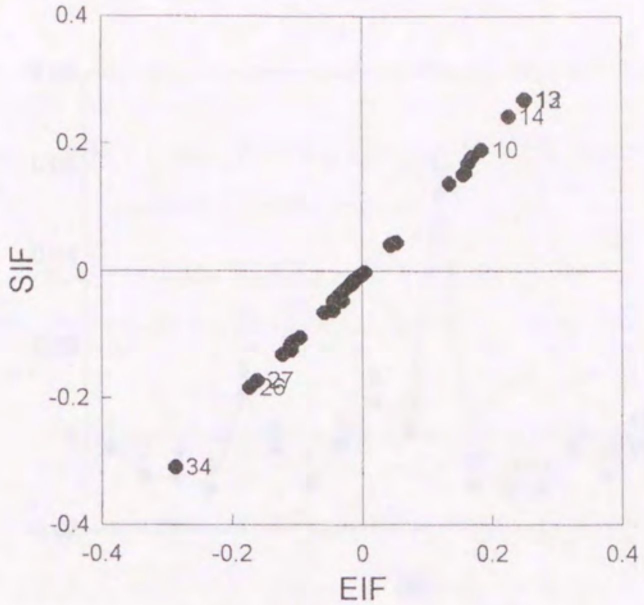
Secondly the influence of variables is evaluated with the method proposed in section 5.5.2. Figure 5.7 shows the index plots of $\widehat{P}^{(1)}$ and $\widehat{RV}^{(1)}$ for 19 variables. These plots show that variable No.11 is extremely influential compared to the other variables. The signs of $\widehat{P}^{(1)}$ and $\widehat{RV}^{(1)}$ indicate that both the proportion explained by the first two PCs and the closeness of the configurations improve much by underweighting variable No.11. It may be related with the fact that R^2 is very small (only 0.3) for this variable while it is much larger (more than 0.7) for the other variables. We should consider the possibility to analyze this variable separately from the other variables.

5.7 Concluding remarks

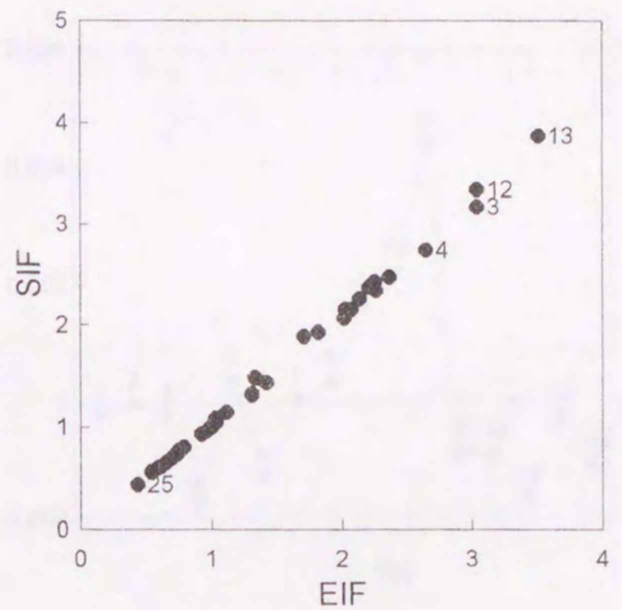
We have proposed a generalized PCA in which PCs are computed using a small number of selected variables but represent all the variables well, borrowing the ideas of Rao(1964)'s PCA of instrumental variables and Robert and Escoufier(1976)'s approach based on RV-coefficient, and also developed methods of sensitivity analysis to study the influence of individuals and variables on the results of analysis. From the numerical study in section 5.6 we can say the followings:

(1) The proposed generalized PCA is effective to obtain PCs which represent all the variables well but are computed using only a part of variables. This method will be useful specifically in the case where we wish to construct a multidimensional rating scale which has high validity and is easy to apply practically.

(2) To evaluate the influence of individuals the EIF can be used instead of the SIF, which has clear "leave-one-out" interpretation, and therefore the generalized procedure



(a)



(b)

Figure 5.6: Scatter plots of SIF against EIF of (a) $\widehat{P}^{(1)}$ and (b) $||(\widehat{L}\widehat{L}')^{(1)}||$ (influence of individuals, Alate data)

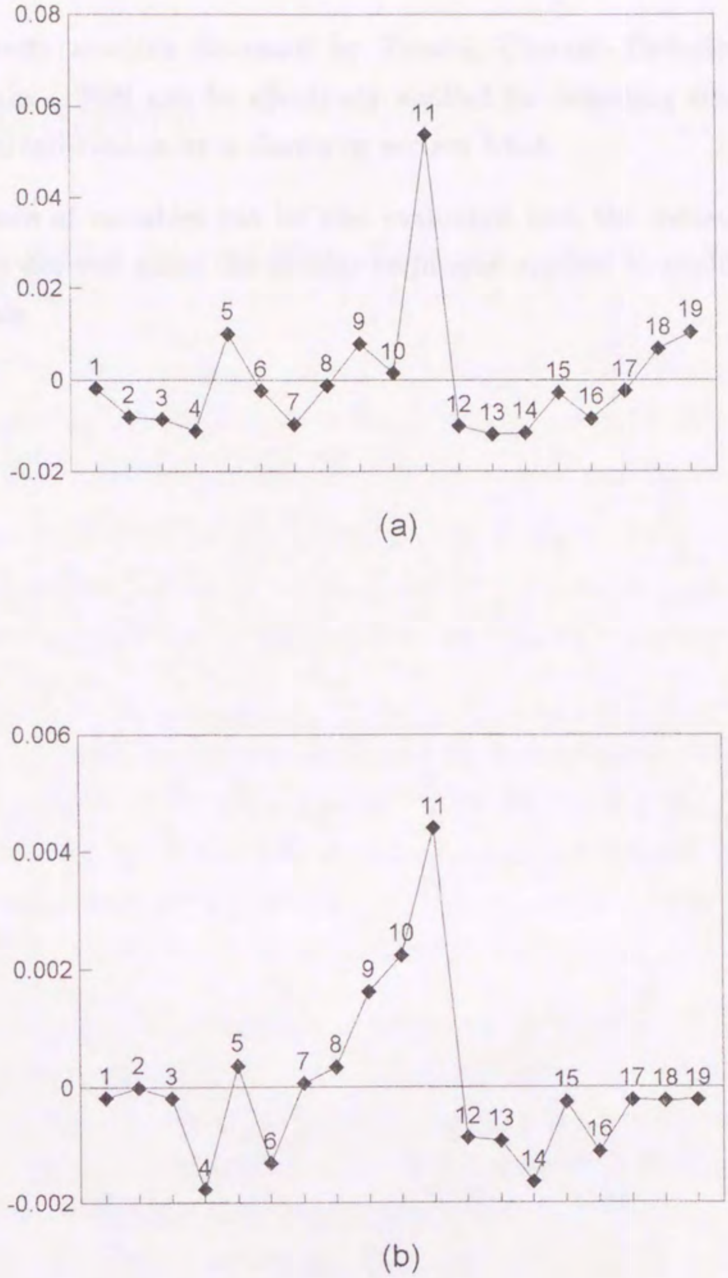


Figure 5.7: Index plots of (a) $\widehat{P}^{(1)}$ and (b) $\widehat{RV}^{(1)}$ (influence of variables, Alate data)

of sensitivity analysis discussed by Tanaka, Castaño-Tostado and Odaka (1990) and Tanaka (1992) can be effectively applied for detecting singly and/or multiply influential individuals as is shown in section 5.6.3.

(3) The influence of variables can be also evaluated with the measures in section 2.3.1, which are derived using the similar technique applied to evaluate the influence of individuals.

6 Conclusion

In this thesis we discussed the reduction of variables in the multivariate analysis without response variables. Especially we have the following two senses for variable selection:

- How to select reasonable variables which reproduce the original features as well as possible among the existing variables in principal component analysis (PCA) and Hayashi's third method of quantification;

- How to conduct PCA to extract the similar dimensions using a subset of variables to those based on the complete variables. In this case, moreover, the observation of the influence of individuals and variables are focused when such a subset of variables is found.

As mathematical tools, we used Robert and Escoufier(1976)'s RV-coefficient and the perturbation theory in the former study, and Rao(1964)'s PCA of instrumental variables, the RV-coefficient to select a subset of variables and the concept of influence functions for sensitivity analysis in the latter case.

In practice, the former study is summarized as follows:

- In principal component analysis, a backward elimination procedure has been proposed for variable selection. This procedure could discard a variable is discarded among the existing variables in each step in such a way that it causes the smallest effect on the configuration of the PC scores. The RV-coefficient was used to evaluate the difference of the configurations of the PC scores and the perturbation theory of eigenvalue problems as well as the exact method were utilized to compute the effect on the configurations. Two sets of real data and four sets of artificial data were analysed for the comparison of our method with other methods proposed so far. In these numerical examples our method made reasonable results of variable selection in PCA.

- In Hayashi's third method of quantification, which deals with categorical data sets, similar backward eliminations to that in PCA has been proposed. They were derived by modifying the selection procedure in PCA partly, and then have the same concept for selection. These procedures can treat both free-choice and item-category data forms and avoid the case where the computation cannot be done in the selection process. Evaluating these methods by analyzing two real data sets, they selected variables from each clusters observed in profile plot of variables, while there seems to exit some errors due to perturbation or data forms.

On the other hand, summary of the latter study is as follows:

- We have proposed a generalized PCA in which PCs are computed using only a selected subset of variables but represent all the variables well, using the ideas of Rao (1964) and Robert and Escoufier (1976), and also proposed methods of sensitivity analysis to evaluate the influence of individuals and variables on the results of analysis. A couple of numerical studies suggest that the proposed generalized PCA is effective from the aspect of two criteria to represent all the variables well, and that the general procedure of sensitivity analysis works well to detect influential individuals and variables in the proposed generalized PCA. These methods will be useful specifically in the case where we wish to construct a multidimensional rating scale which has high validity and is easy to apply practically.

The future considerations are follows:

- How to decide the number of variables has not been discussed in our study. It is free to retain how many variables under the dimensionality fixed at the beginning step in our procedures. For this area of study we can refer to Jolliffe (1973, 1986) or Krzanowski (1987b). It seems to be a considerable problem to decide the number of variables which should be selected.

- Only backward procedures were proposed. It is possible to make forwardstep and/or stepwise procedures to select variables in the whole studies.

- We used the criteria so as to reproduce the original features as well as possible. Further criteria can be considered. For example, especially in Hayashi's third method of quantification, a set of variables can be chosen in the sense to represent the linearity contained in the data, and to order or rank individuals as well as possible. And they should be compared with other criteria.

- Variable selection procedures can be proposed and should be proposed according to the situation and purpose of selection. It becomes necessary to summarize various procedures in multivariate analysis without response variables.

Appendix A Proofs

A.1 Proof of eq.(3.10)

Substituting $\tilde{T} = T + \varepsilon T^{(1)} + (\varepsilon^2/2)T^{(2)} + O(\varepsilon^3)$ in (3.9)

$$RV(A,\tilde{A}) = \frac{tr(AA'\tilde{A}\tilde{A}')}{\left\{tr(AA')^2 \cdot tr(\tilde{A}\tilde{A}')^2\right\}^{1/2}} = \frac{tr(T\tilde{T})}{\left\{tr(T^2)\cdot tr(\tilde{T}^2)\right\}^{1/2}},$$

the numerator is

$$\begin{aligned} NUM &= tr(T\tilde{T}) = tr(T^2) + \varepsilon \cdot tr(TT^{(1)}) + \frac{\varepsilon^2}{2} tr(TT^{(2)}) + O(\varepsilon^3) \\ &= \left\{tr(T^2)\right\}^{-1}\left\{1 + \varepsilon\frac{tr(TT^{(1)})}{tr(T^2)} + \frac{\varepsilon^2}{2}\cdot\frac{tr(TT^{(2)})}{tr(T^2)} + O(\varepsilon^3)\right\}, \end{aligned}$$

and the denominator is

$$\begin{aligned} DEN &= \left\{tr(T^2)\cdot tr(\tilde{T}^2)\right\}^{-1/2} \\ &= \left\{tr(T^2)\right\}^{-1/2}\left[tr\left\{T + \varepsilon T^{(1)} + \frac{\varepsilon^2}{2}T^{(2)} + O(\varepsilon^3)\right\}^2\right]^{-1/2} \\ &= \left\{tr(T^2)\right\}^{-1/2}\left[tr(T^2) + 2\varepsilon\cdot tr(TT^{(1)}) + \varepsilon^2\left\{tr(T^{(1)2}) + tr(TT^{(2)})\right\} + O(\varepsilon^3)\right]^{-1/2} \\ &= \left\{tr(T^2)\right\}^{-1}\left[1 + \varepsilon\frac{2tr(TT^{(1)})}{tr(T^2)} + \varepsilon^2\frac{tr(T^{(1)2}) + tr(TT^{(2)})}{tr(T^2)} + O(\varepsilon^3)\right]^{-1/2} \\ &= \left\{tr(T^2)\right\}^{-1} g(\varepsilon). \end{aligned}$$

The expansion of $g(\varepsilon)$ by the perturbation is expressed as

$$g(\varepsilon) = g(0) + \varepsilon g^{(1)}(0) + (\varepsilon^2/2)g^{(2)}(0) + O(\varepsilon^3).$$

Since the first differential coefficient is

$$g^{(1)}(\varepsilon) = -\frac{1}{2}\{g(\varepsilon)\}^{-3/2}\left[\frac{2tr(TT^{(1)})}{tr(T^2)} + \varepsilon\frac{2\left\{tr(T^{(1)2}) + tr(TT^{(2)})\right\}}{tr(T^2)} + O(\varepsilon^3)\right],$$

then we get

$$g^{(1)}(0) = -\frac{tr(TT^{(1)})}{tr(T^2)}.$$

Also, since the second differential coefficient is

$$g^{(2)}(\varepsilon) = \frac{3}{4}\{g(\varepsilon)\}^{-5/2}\{g(\varepsilon)\}^2 - \frac{1}{2}\{g(\varepsilon)\}^{-3/2}\left[\frac{2\left\{tr(T^{(1)2}) + tr(TT^{(2)})\right\}}{tr(T^2)} + O(\varepsilon^3)\right],$$

then we get

$$\begin{aligned} g^{(2)}(0) &= \frac{3}{4}\left\{\frac{2tr(TT^{(1)})}{tr(T^2)}\right\}^2 - \frac{1}{2}\cdot\frac{2\left\{tr(T^{(1)2}) + tr(TT^{(2)})\right\}}{tr(T^2)} \\ &= 3\left\{\frac{tr(TT^{(1)})}{tr(T^2)}\right\}^2 - \frac{tr(T^{(1)2}) + tr(TT^{(2)})}{tr(T^2)}. \end{aligned}$$

Hence,

$$g(\varepsilon) = 1 - \varepsilon\frac{tr(TT^{(1)})}{tr(T^2)} + \frac{\varepsilon^2}{2}\left\{3\left(\frac{tr(TT^{(1)})}{tr(T^2)}\right)^2 - \frac{tr(T^{(1)2}) + tr(TT^{(2)})}{tr(T^2)}\right\} + O(\varepsilon^3).$$

Thus, we have

$$\begin{aligned} RV(A,\tilde{A}) &= \frac{NUM}{DEN} \\ &= \left[1 + \varepsilon\frac{tr(TT^{(1)})}{tr(T^2)} + \frac{\varepsilon^2}{2}\frac{tr(TT^{(2)})}{tr(T^2)} + O(\varepsilon^3)\right] \\ &\quad \times\left[1 - \varepsilon\frac{tr(TT^{(1)})}{tr(T^2)} + \frac{\varepsilon^2}{2}\left\{3\left(\frac{tr(TT^{(1)})}{tr(T^2)}\right)^2 - \frac{tr(T^{(1)2}) + tr(TT^{(2)})}{tr(T^2)}\right\} + O(\varepsilon^3)\right] \\ &= 1 - \frac{\varepsilon^2}{2}\left[\frac{tr(T^{(1)2})}{tr(T^2)} - \left\{\frac{tr(TT^{(1)})}{tr(T^2)}\right\}^2\right] + O(\varepsilon^3). \end{aligned}$$

We can see the last equation in Castaño-Tostado and Tanaka(1991)'s paper.

A.2 Proof of properties P1° –P3°

(1) A case where $q = p$:

Since $\Sigma_{11} = \Sigma_1 = \Sigma$, the eigenvalue problem (5.3) is the same as $(\Sigma - \lambda I)A = 0$. Then

$$\sum_{i=1}^{r} \lambda_i \leq \sum_{i=1}^{q} \lambda_i = \sum_{i=1}^{p} \lambda_i = tr(\Sigma) = \sum_{i=1}^{p} \sigma_{ii},$$

where σ_{ii} is the variance of the i-th variable or the i-th diagonal element of Σ.

(2) A case where $q < p$, and $p - q$ variables are completely redundant :

In multiple regression with the $p - q$ variables as dependent variables, since the residuals

$$\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = 0,$$

then,

$$\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \Sigma_{22}.$$

Thus,

$$\begin{aligned}
\sum_{i=1}^{r} \lambda_i \leq \sum_{i=1}^{q} \lambda_i &= tr(\Sigma_1'\Sigma_{11}^{-1}\Sigma_1) \\
&= tr(\Sigma_{11}^{-1}\Sigma_1\Sigma_1') = tr[\Sigma_{11}^{-1}(\Sigma_{11}^2 + \Sigma_{12}\Sigma_{21})] = tr(\Sigma_{11}) + tr(\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}) \\
&= tr(\Sigma_{11}) + tr(\Sigma_{22}) \\
&= \sum_{i=1}^{q} \sigma_{ii} + \sum_{i=q+1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \sigma_{ii}.
\end{aligned}$$

The above shows that the sum of r eigenvalues is always equal to $\sum_{i=1}^{p} \sigma_{ii}$ whenever the completely redundant variables are removed. This is a proof of property **P3°**.

(3) The case where remove variables which are not completely redundant :

In this case, since $\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ is an non negative definite,

$$tr(\Sigma_{22}) \geq tr(\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}).$$

Then

$$\begin{aligned}
\sum_{i=1}^{r} \lambda_i \leq \sum_{i=1}^{q} \lambda_i &= tr(\Sigma_1\Sigma_{11}^{-1}\Sigma_1') = tr(\Sigma_{11}) + tr(\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}) \\
&\leq tr(\Sigma_{11}) + tr(\Sigma_{22}) = \sum_{i=1}^{p} \sigma_{ii},
\end{aligned}$$

that is

$$\sum_{i=1}^{r} \lambda_i \leq \sum_{i=1}^{p} \sigma_{ii},$$

which means that the sum of r eigenvalues is not greater than that of variances of the original matrix.

From above (1)-(3), $P \leq 1$ is obtained, that is, a proof of **P1°**.

(4) The case where $q < q* < p$ and the number of variables in X_1 is q^* : (Figure A.1)

Obviously

$$\sum_{i=1}^{q*} \lambda_i^* = tr(\Sigma_{11}^*) + tr(\Sigma_{21}^* \Sigma_{11}^{*-1} \Sigma_{12}^*).$$

The first term in the right hand side is

$$\sum_{i=1}^{q*} \sigma_{ii} = \sum_{i=1}^{q} \sigma_{ii} + \sum_{i=q+1}^{q*} \sigma_{ii},$$

and the second term is sum of variation due to regression with q^* variables as independent variables and $p - q^*$ as dependent ones.

(5) The case where $q < q* < p$ and the number of variables in X_1 is q : (Figure A.2)

Since $q^* - q$ variables are reduced,

$$\begin{aligned} \sum_{i=1}^{q} \lambda_i &= tr(\Sigma_{11}) + tr(\Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}) \\ &= tr(\Sigma_{11}) + tr[(\Sigma_{21(q*-q)} \Sigma_{11}^{-1} \Sigma_{12(q*-q)})] + tr[(\Sigma_{21(p-q*)} \Sigma_{11}^{-1} \Sigma_{12(p-q*)})]. \quad \text{(A.1)} \end{aligned}$$

The first term of the right hand side is

$$\sum_{i=1}^{q} \sigma_{ii}.$$

The value of the second term is less than or equal to $\sum_{i=q+1}^{q*} \sigma_{ii}$ because it is sum of variation due to regression with q variables as independent variables and $q^* - q$ as dependent variables. The third is sum of variation due to regression with q variables as independent variables and $p - q*$ as dependent variables, which is not greater than the second term of eq.(A.1) because the number of independent variables is reduced.

From (4) and (5),

$$\sum_{i=1}^{q} \lambda_i \leq \sum_{i=1}^{q*} \lambda_i^*,$$

then property **P2°** is always obtained.

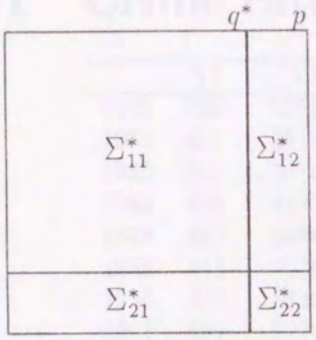


Figure A.1

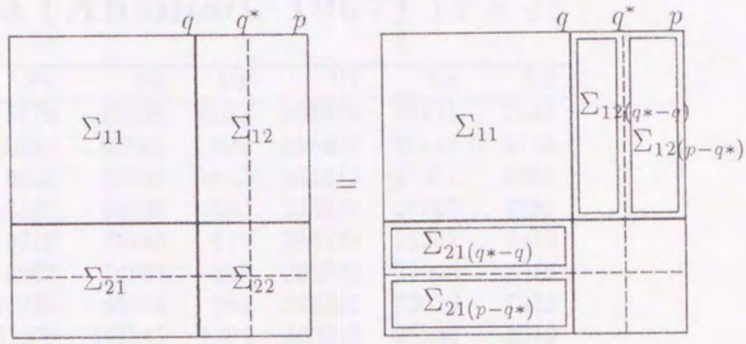


Figure A.2

Appendix B Sets of Raw Data

B.1 Crime rates data (Ahamad, 1967) 14×18

	V1	V2	V3	V4	V5	V6	V7	V8	V9
1950	529	5258	4416	8178	92839	1021	301078	25333	7586
1951	455	5619	4876	9223	95946	800	355407	27216	9716
1952	555	5980	5443	9026	97941	1002	341512	27051	9188
1953	456	6187	5680	10107	88607	980	308578	27763	7786
1954	487	6586	6357	9279	75888	812	285199	26267	6468
1955	448	7076	6644	9953	74907	823	295035	22966	7016
1956	477	8433	6196	10505	85768	965	323561	23029	7215
1957	491	9774	6327	11900	105042	1194	360985	26235	8619
1958	453	10945	5471	11823	131132	1692	409388	29415	10002
1959	434	12707	5732	13864	133962	1900	445888	34061	10254
1960	492	14391	5240	14304	151378	2014	489258	36049	11696
1961	459	16197	5605	14376	164806	2349	531430	39651	13777
1962	504	16430	4866	14788	192302	2517	588566	44138	15783
1963	510	18655	5435	14722	219138	2483	635627	45923	17777

	V10	V11	V12	V13	V14	V15	V16	V17	V18
1950	4518	3790	118	20844	9477	24616	49007	2786	3126
1951	4993	3378	74	19963	10359	21122	55229	2739	5495
1952	5003	4173	120	19056	9108	23339	55635	2598	4145
1953	5309	4649	108	17772	9278	19919	55688	2639	4551
1954	5251	4903	104	17379	9176	20585	57011	2587	4343
1955	2184	4086	92	17329	9460	19198	57118	2607	4836
1956	2559	4040	119	16677	10997	19064	63289	2311	5932
1957	2965	4689	121	17539	12817	16432	71014	2310	7148
1958	3607	5376	164	17344	14289	24543	69864	2371	9772
1959	4083	5598	160	18047	14118	26853	69751	2544	11211
1960	4802	6590	241	18801	15866	31266	74336	2719	12519
1961	5606	6924	205	18525	16399	29922	81753	2820	13050
1962	6256	7816	250	16449	16852	34915	89709	2614	14141
1963	6935	8634	257	15918	17003	40434	89149	2777	22896

V1	Homicide	V2	Woundings	V3	Homosexual offences
V4	Heterosexual offences	V5	Breaking and entering	V6	Robbery
V7	Larceny	V8	Frauds and false pretences	V9	Peceiving
V10	Malicious injuries to property	V11	Forgery	V12	Blackmail
V13	Assault	V14	Malicious damage	V15	Revenue laws
V16	Intoxication laws	V17	Indecent exposure	V18	Taking motor vehicle without consent

B.2 Automobile data (Becker et al., 1988) 74×10

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
C1	4099	22	2.5	27.5	11	2930	186	40	121	3.58
C2	4749	17	3	25.5	11	3350	173	40	258	2.53
C3	3799	22	3	18.5	12	2640	168	35	121	3.08
C4	9690	17	3	27	15	2830	189	37	131	3.2
C5	6295	23	2.5	28	11	2070	174	36	97	3.7
C6	9735	25	2.5	26	12	2650	177	34	121	3.64
C7	4816	20	4.5	29	16	3250	196	40	196	2.93
C8	7827	15	4	31.5	20	4080	222	43	350	2.41
C9	5788	18	4	30.5	21	3670	218	43	231	2.73
C10	4453	26	3	24	10	2230	170	34	111	2.87
C11	5189	20	2	28.5	16	3280	200	42	196	2.93
C12	10372	16	3.5	30	17	3880	207	43	231	2.93
C13	4082	19	3.5	27	13	3400	200	42	231	3.08
C14	11385	14	4	31.5	20	4330	221	44	425	2.28
C15	14500	14	3.5	30	16	3900	204	43	350	2.19
C16	15906	21	3	30	13	4290	204	45	350	2.24
C17	3299	29	2.5	26	9	2110	163	34	98	2.93
C18	5705	16	4	29.5	20	3690	212	43	250	2.56
C19	4504	22	3.5	28.5	17	3180	193	41	200	2.73
C20	5104	22	2	28.5	16	3220	200	41	200	2.73
C21	3667	24	2	25	7	2750	179	40	151	2.73
C22	3955	19	3.5	27	13	3430	197	43	250	2.56
C23	6229	23	1.5	21	6	2370	170	35	119	3.89
C24	4589	35	2	23.5	8	2020	165	32	85	3.7
C25	5079	24	2.5	22	8	2280	170	34	119	3.54
C26	8129	21	2.5	27	8	2750	184	38	146	3.55
C27	3984	30	2	24	8	2120	163	35	98	3.54
C28	5010	18	4	29	17	3600	206	46	318	2.47
C29	5886	16	3.5	26	16	3870	216	48	318	2.71
C30	6342	17	4.5	28	21	3740	220	46	225	2.94
C31	4296	21	2.5	26.5	16	2130	161	36	105	3.37
C32	4389	28	1.5	26	9	1800	147	33	98	3.15
C33	4187	21	2	23	10	2650	179	42	140	3.08
C34	5799	25	3	25.5	10	2240	172	36	107	3.05
C35	4499	28	2.5	23.5	5	1760	149	34	91	3.3
C36	11497	12	3.5	30.5	22	4840	233	51	400	2.47
C37	13594	12	2.5	28.5	18	4720	230	48	400	2.47
C38	13466	14	3.5	27	15	3830	201	41	302	2.47
C39	3995	30	3.5	25.5	11	1980	154	33	86	3.73
C40	3829	22	3	25.5	9	2580	169	39	140	2.73
C41	5379	14	3.5	29.5	16	4060	221	48	302	2.75
C42	6303	14	3	25	16	4130	217	45	302	2.75
C43	6165	15	3.5	30.5	23	3720	212	44	302	2.26
C44	4516	18	3	27	15	3370	198	41	250	2.43
C45	3291	20	3.5	29	17	2830	195	43	140	3.08
C46	8814	21	4	31.5	20	4060	220	43	350	2.41
C47	4733	19	4.5	28	16	3300	198	42	231	2.93
C48	5172	19	2	28	16	3310	198	42	231	2.93
C49	5890	18	4	29	20	3690	218	42	231	2.73
C50	4181	19	4.5	27	14	3370	200	43	231	3.08

C51	4195	24	2	25.5	10	2720	180	40	151	2.73
C52	10371	16	3.5	30	17	4030	206	43	350	2.41
C53	12990	14	3.5	30.5	14	3420	192	38	163	3.58
C54	4647	28	2	21.5	11	2360	170	37	156	3.05
C55	4425	34	2.5	23	11	1800	157	37	86	2.97
C56	4482	25	4	25	17	2200	165	36	105	3.37
C57	6486	26	1.5	22	8	2520	182	38	119	3.54
C58	4060	18	5	31	16	3330	201	44	225	3.23
C59	5798	18	4	29	20	3700	214	42	231	2.73
C60	4934	18	1.5	23.5	7	3470	198	42	231	3.08
C61	5222	19	2	28.5	16	3210	201	45	231	2.93
C62	4723	19	3.5	28	17	3200	199	40	231	2.93
C63	4424	19	3.5	27	13	3420	203	43	231	3.08
C64	4172	24	2	25	7	2690	179	41	151	2.73
C65	3895	26	3	23	10	1830	142	34	79	3.72
C66	3798	35	2.5	25.5	11	2050	164	36	97	3.81
C67	5899	18	2.5	22	14	2410	174	36	134	3.06
C68	3748	31	3	24.5	9	2200	165	35	97	3.21
C69	5719	18	2	23	11	2670	175	36	134	3.05
C70	4697	25	3	25.5	15	1930	155	35	89	3.78
C71	5397	41	3	25.5	15	2040	155	35	90	3.78
C72	6850	25	2	23.5	16	1990	156	36	97	3.78
C73	7140	23	2.5	37.5	12	2160	172	36	97	3.74
C74	11995	17	2.5	29.5	14	3170	193	37	163	2.98

V1 Price V2 Miles/g V3 Headroom V4 Rear Seat
V5 Trunk V6 Weight V7 Length V8 Turning
V9 Displacement V10 Gear Ratio

B.3 Spirits data (Arima and Ishimura, 1987) 20×7

	V1 Whisky	V2 Beer	V3 Wine	V4 Sake	V5 Shochu	V6 Chuhai	V7 Cocktail
C1	1	0	1	0	0	0	0
C2	0	0	1	1	0	1	1
C3	0	1	1	0	0	1	0
C4	1	0	1	1	0	0	1
C5	1	0	1	1	0	0	1
C6	1	0	1	0	0	0	1
C7	0	0	1	0	1	1	1
C8	1	0	1	0	0	0	1
C9	0	0	1	0	0	0	1
C10	1	1	1	0	0	1	1
C11	1	0	1	0	0	1	0
C12	0	0	0	1	0	1	1
C13	1	0	1	0	0	1	1
C14	0	0	0	0	0	1	1
C15	0	1	1	0	0	1	0
C16	0	0	1	0	0	0	1
C17	0	0	0	0	0	1	0
C18	1	0	1	0	0	0	1
C19	1	0	1	0	0	0	1
C20	0	0	1	0	0	0	1

B.4 Fatigue data (Maehashi et al., 1993) 100×30

(a) Before physical movements

	I	II	III
	1 10	11 20	21 30
C1	1 1 0 1 1 0 0 0 0 0	0 0 0 0 0 1 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C2	0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C3	0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 1 0 1 0 0	0 0 0 0 0 0 0 0 0 0
C4	1 0 0 0 0 1 0 0 0 0	0 0 0 0 0 0 0 1 0 0	0 1 0 0 0 0 0 0 0 0
C5	0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C6	0 1 0 1 1 1 0 0 1 0	0 0 0 0 1 1 0 0 1 1	0 0 0 0 1 0 0 0 0 0
C7	0 0 0 0 0 0 0 0 0 0	1 0 0 0 0 1 1 0 0 0	0 0 0 0 0 0 0 0 0 0
C8	0 0 0 1 0 0 0 0 0 0	0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C9	0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C10	0 1 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C11	0 0 0 1 1 1 1 1 1 1	1 1 1 1 0 1 0 1 1 1	0 1 1 0 0 0 0 1 0 0
C12	0 0 1 0 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C13	0 0 0 0 0 0 0 0 0 0	0 0 1 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C14	0 0 0 0 0 0 0 0 0 0	1 0 0 0 0 1 0 1 0 0	0 0 0 1 0 0 0 0 0 0
C15	0 0 0 1 0 0 1 0 0 0	0 0 0 1 0 0 0 1 0 0	0 0 0 0 0 0 0 0 0 0
C16	0 0 0 0 0 0 0 1 0 1	0 0 0 0 0 0 0 1 0 0	0 0 0 0 0 0 0 0 0 0
C17	0 0 0 1 0 1 0 0 0 0	0 0 0 0 0 0 0 1 0 0	0 0 0 0 0 0 0 0 0 0
C18	1 0 1 0 0 0 1 0 1 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C19	0 0 0 0 0 1 1 0 0 0	0 0 0 0 0 0 0 1 0 0	0 1 0 0 1 0 0 0 0 0
C20	0 0 1 0 0 0 1 1 1 0	1 0 0 0 1 1 1 0 1 1	0 1 0 0 0 1 0 0 0 0
C21	1 0 0 0 1 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 1 0 0 0
C22	0 0 0 0 0 1 0 0 0 0	0 0 0 0 0 1 0 1 1 0	0 0 0 1 0 0 0 0 0 0
C23	0 0 0 0 0 0 0 0 0 0	1 0 0 0 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C24	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 1 0 1 0 0	0 0 0 0 0 0 0 0 0 0
C25	0 0 1 1 0 0 0 0 1 1	1 0 1 1 0 1 1 0 0 0	1 0 0 0 1 1 0 0 1 0
C26	0 0 0 0 0 0 0 0 0 0	0 0 0 1 1 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C27	1 1 1 0 0 0 0 0 0 1	1 0 1 1 0 0 0 1 0 0	0 1 0 1 0 0 0 0 0 0
C28	1 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 1 0 0	0 0 0 0 0 0 0 0 0 0
C29	1 0 1 0 0 0 1 0 0 0	1 0 1 0 0 0 1 0 1 0	0 0 0 0 0 1 0 0 1 0
C30	0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C31	0 0 0 0 0 0 0 0 0 0	1 0 0 0 1 0 1 0 1 1	0 0 0 0 0 0 0 0 0 0
C32	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 1 0 0 1 0	0 0 0 0 1 0 0 0 0 0
C33	1 1 0 1 0 1 1 0 0 1	0 0 0 0 0 1 0 0 0 0	0 0 0 0 0 1 0 0 0 0
C34	1 1 1 0 1 0 1 0 0 0	0 0 0 0 0 0 1 1 0 0	1 0 0 0 0 0 0 0 0 0
C35	1 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 1 0 0	0 0 0 0 0 0 0 0 0 0
C36	0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C37	0 0 1 0 0 0 0 0 0 0	1 0 1 1 1 0 0 1 0 0	0 0 0 0 0 1 0 0 0 0
C38	0 0 0 1 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C39	0 0 0 1 0 0 0 0 0 0	1 0 0 0 0 1 1 0 1 0	0 1 0 0 0 0 0 0 0 0
C40	0 0 1 0 0 0 0 0 0 0	0 0 0 1 0 0 0 0 0 0	0 1 0 0 0 0 0 0 0 0
C41	0 0 0 0 0 0 0 0 0 0	1 0 1 1 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C42	1 1 0 0 0 1 1 0 0 1	1 0 1 1 1 1 0 1 0 0	0 1 0 0 1 1 0 0 0 0
C43	0 1 0 1 0 1 0 0 0 1	1 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 1
C44	1 1 0 0 0 0 1 0 0 0	0 0 1 0 1 0 0 0 0 0	1 1 0 0 1 0 0 0 0 1
C45	0 0 1 1 0 1 0 0 0 1	1 0 0 0 0 1 0 1 0 0	0 0 0 0 0 0 0 0 0 0
C46	0 0 0 1 0 1 0 0 0 1	0 0 1 0 1 1 1 0 0 1	0 0 0 0 0 0 0 1 0 0
C47	0 0 0 0 0 0 0 0 0 0	0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0

C48	1 1 0 1 0 1 1 0 0 0	1 0 1 1 1 1 0 0 1 0	0 0 0 0 0 0 0 0 0 0
C49	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C50	1 0 1 0 1 1 1 0 0 1	1 0 0 0 0 1 1 0 0 0	1 1 0 0 0 0 0 0 0 1
C51	0 0 0 1 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C52	1 1 0 1 0 1 0 1 1 1	0 0 1 0 1 0 0 0 1 1	1 1 0 1 0 1 1 0 0 0
C53	0 0 0 1 0 1 0 0 0 1	0 0 1 1 1 1 0 1 1 0	0 0 0 0 0 0 0 1 0 0
C54	0 0 0 1 0 1 0 0 0 1	1 1 1 0 1 1 0 0 1 1	0 0 0 0 1 0 0 0 0 0
C55	0 1 0 0 1 0 0 0 0 0	0 0 0 0 0 0 1 1 0 0	0 0 0 1 0 0 0 0 0 0
C56	0 0 0 1 0 1 0 0 0 1	0 0 0 0 1 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C57	0 0 0 1 0 1 0 0 0 1	0 0 0 0 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C58	0 0 1 1 0 1 0 0 0 1	0 0 0 1 0 1 0 0 1 0	0 0 1 0 0 1 0 0 0 0
C59	0 0 0 0 0 0 0 0 0 1	1 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 1 0 0
C60	0 0 0 1 0 1 0 0 0 1	0 0 0 0 1 0 0 0 1 0	0 0 0 0 0 0 0 0 0 0
C61	0 0 0 1 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C62	0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 0 0 1 0 0	0 0 0 0 0 0 0 0 0 0
C63	0 0 0 0 0 1 0 0 0 0	0 0 0 0 1 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C64	0 0 0 1 0 1 0 0 0 1	0 0 0 0 1 0 0 0 1 0	0 0 0 0 0 0 0 0 0 0
C65	0 0 0 0 0 1 0 0 0 1	0 0 0 1 0 1 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C66	0 0 0 1 0 1 0 0 0 1	0 0 0 0 0 0 0 0 1 0	0 0 0 0 0 0 0 0 0 0
C67	0 0 0 0 0 0 0 0 0 1	0 0 0 0 0 1 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C68	0 0 0 1 0 1 0 0 0 1	0 0 0 0 0 1 1 0 0 0	1 0 0 0 0 0 0 0 0 0
C69	0 0 0 1 0 1 0 0 0 1	0 0 0 0 0 0 1 0 1 0	0 0 0 0 0 0 0 0 0 0
C70	0 1 0 1 0 1 0 0 0 1	0 0 0 1 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0
C71	0 0 0 1 0 1 0 0 0 1	0 0 0 0 0 0 0 0 1 0	0 0 0 0 0 1 0 0 0 0
C72	0 0 0 1 0 1 1 0 0 1	0 0 0 0 1 1 0 0 0 0	0 1 0 0 1 0 0 0 0 0
C73	0 0 0 0 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C74	0 0 0 1 1 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C75	0 0 0 0 0 0 0 0 0 1	0 0 0 0 1 0 1 1 0 0	0 0 0 0 0 0 0 0 0 0
C76	1 1 1 1 1 1 1 0 1 1	1 1 1 0 1 1 1 1 0 1	1 1 1 1 1 0 1 1 1 1
C77	0 0 0 1 0 0 0 0 0 1	0 0 0 0 1 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C78	0 0 0 1 0 1 0 0 0 1	0 0 0 1 0 0 0 0 1 0	0 0 0 0 1 0 0 0 0 0
C79	0 0 0 1 0 1 0 0 0 0	0 0 0 0 0 0 0 0 1 0	0 0 0 0 0 0 0 0 0 0
C80	0 0 0 0 0 0 0 0 0 1	0 0 1 0 1 1 0 1 1 0	0 0 0 0 0 0 0 1 0 0
C81	0 0 0 1 0 1 0 0 0 1	0 0 0 0 0 0 0 0 1 0	0 0 0 0 0 1 0 0 0 0
C82	0 0 0 1 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C83	0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 1 0 0 0 0 0 0 0 0
C84	0 0 1 0 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 1 0 0 0
C85	0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 1 0 0 0 0 0 0 0 0
C86	0 0 0 1 0 1 0 0 0 1	0 0 0 0 0 0 0 0 1 0	0 0 1 1 0 0 0 0 0 0
C87	0 0 0 0 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C88	0 0 0 1 0 1 0 0 0 0	0 0 0 0 0 0 0 1 0 0	0 0 0 0 0 0 0 0 0 0
C89	0 0 0 1 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C90	0 1 0 0 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C91	0 0 0 1 0 0 0 0 0 0	0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C92	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C93	0 0 0 0 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C94	0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C95	0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	1 0 0 0 0 0 0 0 0 0
C96	0 0 0 0 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C97	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 1 0 0 0	0 0 0 0 0 0 0 0 0 0
C98	1 1 1 0 1 1 1 1 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 1 0 0 1
C99	0 0 0 1 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C100	0 0 0 1 0 0 0 0 0 0	1 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0

(b) After physical movements

	I	II	III
	1 10	11 20	21 30
C1	1 1 0 1 1 0 0 0 0 0	1 0 0 0 0 1 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C2	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 1 0 0 0 0 0 0 0 0
C3	0 0 0 0 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C4	0 1 1 0 0 0 0 0 0 1	0 0 0 0 0 1 0 0 1 1	0 1 0 0 0 0 0 0 0 0
C5	0 1 0 1 1 1 0 0 1 0	0 0 0 0 1 1 0 0 1 1	0 0 0 0 1 0 0 0 0 0
C6	1 1 1 0 0 0 1 0 1 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C7	0 1 1 0 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C8	0 1 1 0 0 1 1 0 0 1	0 0 0 0 0 0 0 0 0 0	0 1 1 0 0 0 0 0 0 0
C9	1 0 0 0 0 0 0 0 1 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C10	0 1 1 0 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C11	1 0 0 0 0 0 0 1 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 1 0 0
C12	1 0 0 0 1 0 1 1 1 0	1 1 1 1 0 1 0 1 0 1	0 0 0 0 0 0 0 0 0 0
C13	0 1 0 0 0 0 0 0 0 0	1 1 0 0 0 0 0 1 0 0	0 0 0 0 0 0 1 0 0 1
C14	0 0 0 0 0 0 0 0 1 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C15	0 1 1 0 0 0 1 1 0 0	0 0 0 0 0 0 0 0 0 0	0 1 0 0 1 0 0 0 1 0
C16	1 1 1 0 1 1 1 0 0 1	1 0 0 0 0 0 0 0 0 0	0 0 0 0 0 1 0 0 0 0
C17	0 1 0 0 0 0 0 0 0 0	0 0 1 1 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C18	0 0 1 0 0 1 1 0 1 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C19	0 1 0 0 1 1 0 1 0 0	0 0 0 0 0 0 0 1 0 0	0 0 0 0 0 0 1 0 0 0
C20	0 1 1 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C21	1 1 1 0 1 1 0 0 1 1	1 1 1 1 1 1 1 1 0 1	1 1 0 1 1 1 1 0 1 0
C22	0 1 1 0 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C23	1 1 1 0 1 1 1 1 0 1	1 1 1 1 0 0 0 1 1 1	0 1 0 1 0 0 1 0 0 0
C24	0 1 1 0 0 0 0 1 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C25	1 0 0 0 0 1 1 0 0 0	1 0 1 0 0 0 0 1 1 0	0 1 0 0 1 1 0 0 1 0
C26	0 0 0 0 0 0 1 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C27	0 1 1 0 0 0 0 0 0 0	1 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C28	0 0 0 0 0 1 0 0 0 0	0 0 0 0 0 0 0 0 1 0	0 0 0 0 0 0 1 0 0 0
C29	0 1 1 0 1 0 1 0 0 1	0 0 0 0 0 0 0 0 0 0	0 1 0 1 0 0 0 0 0 0
C30	1 1 1 0 1 0 0 0 0 0	0 0 0 0 0 0 1 1 0 0	0 1 0 0 0 0 0 0 0 0
C31	0 0 0 0 0 1 1 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C32	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C33	0 0 1 1 0 1 0 0 0 1	1 0 1 1 1 0 0 1 0 0	0 0 0 0 1 1 0 0 0 0
C34	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C35	1 1 0 0 1 1 1 0 0 0	1 0 0 1 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C36	0 1 1 0 0 1 0 0 0 1	0 0 0 0 0 0 0 0 1 1	0 0 0 0 0 0 0 0 0 0
C37	0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C38	0 1 1 0 1 1 0 0 0 1	0 0 0 0 0 0 0 0 1 1	0 1 0 1 1 0 0 0 0 1
C39	1 0 0 1 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 1 0 0 0 0 0 0 0 0
C40	0 1 1 0 0 0 1 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 1 0 0 0
C41	0 0 1 0 0 0 1 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C42	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 1 1 0 0 0	0 0 0 0 0 0 0 0 0 0
C43	0 0 0 1 0 1 1 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C44	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C45	0 0 1 0 0 0 1 0 0 0	0 1 0 0 1 1 1 0 0 1	0 0 0 0 0 0 0 0 0 0
C46	1 1 0 0 1 1 0 1 0 1	0 0 0 0 1 1 0 0 0 1	1 0 0 1 0 0 1 0 0 0
C47	0 0 1 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0

C48	0 1 0 0 0 0 1 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C49	0 0 0 0 1 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C50	0 0 1 0 1 0 1 0 1 0	0 0 1 0 0 0 0 0 0 0	0 1 1 1 0 0 0 1 1 0
C51	0 1 0 0 0 0 1 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C52	1 0 0 1 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C53	0 1 0 1 0 1 0 0 0 1	0 0 1 1 0 1 0 1 1 1	0 0 0 0 0 0 0 0 0 0
C54	0 0 0 1 0 1 0 0 0 1	0 1 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C55	0 0 0 0 0 0 0 0 0 1	0 0 0 0 1 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C56	1 1 1 0 1 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 1 0 1 0 0 0 0 0
C57	0 1 0 0 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C58	0 1 0 0 0 0 0 0 0 1	0 0 0 0 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C59	0 0 0 1 0 1 1 0 0 0	1 0 0 0 0 0 0 0 0 0	0 1 0 0 1 0 0 0 0 0
C60	0 0 1 0 0 0 1 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C61	0 0 0 0 1 1 0 0 0 0	0 0 0 0 1 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C62	0 1 1 0 0 0 1 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C63	0 0 0 1 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C64	0 1 0 0 0 1 1 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C65	0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C66	0 1 0 0 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C67	0 1 0 1 0 1 1 0 0 0	0 0 0 0 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0
C68	0 1 1 1 0 1 0 0 0 0	1 0 0 0 0 1 0 0 0 0	0 1 0 0 0 0 0 0 0 0
C69	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C70	0 1 0 1 1 1 1 0 0 1	0 0 0 0 1 1 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C71	0 0 0 0 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C72	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C73	0 1 1 0 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	1 0 0 1 0 0 0 0 0 0
C74	1 1 1 1 1 1 1 0 0 1	1 0 0 0 1 0 1 0 0 0	1 1 0 0 1 0 0 1 0 1
C75	0 0 0 1 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C76	0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C77	0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0
C78	0 0 0 0 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C79	0 0 0 0 1 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 1 0 0 0 0
C80	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C81	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C82	0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	0 1 0 0 0 0 0 0 0 0
C83	0 0 0 0 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C84	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C85	0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C86	0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	1 0 0 1 0 0 0 0 0 0
C87	0 0 0 0 0 1 1 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C88	1 1 1 1 0 1 1 1 1 1	1 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 1 0 0 0
C89	1 0 1 0 1 1 0 0 1 0	1 0 0 0 1 0 0 1 0 0	0 0 0 0 0 0 0 0 0 1
C90	1 0 0 0 1 0 1 0 0 1	0 0 0 0 0 0 0 0 0 0	1 0 0 0 0 0 0 0 0 0
C91	1 1 1 1 1 1 1 1 0 1	0 0 0 0 0 0 0 0 0 0	1 1 0 0 1 0 1 0 0 1
C92	1 1 1 1 1 0 1 1 0 0	0 1 0 1 0 0 0 0 0 0	1 0 0 1 1 0 0 1 0 1
C93	0 1 0 0 1 1 1 0 0 1	0 0 0 0 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 1
C94	0 1 0 0 0 0 1 0 0 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C95	0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 0 0 0 0	1 0 0 0 0 0 0 0 0 0
C96	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
C97	0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 0 1 0 0	0 0 0 0 1 0 0 0 1 0
C98	0 1 1 0 0 1 1 0 0 1	0 0 0 0 0 0 0 0 0 0	0 1 0 0 0 0 0 0 0 0
C99	0 1 1 0 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 1	0 0 0 1 0 0 0 0 0 1
C100	0 0 0 0 0 1 1 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0

I. *drowsiness and dullness*	II. *difficulty concentration*	III. *projection of physical disintegration*
1. your head feeling weary	11.feeling distracted	21.headaches
2. feeling exhausted	12.feeling uncommunicative	22.stiff neck
3. feeling your legs tired	13.feeling irritated	23.backaches
4. feeling like yawning	14.feeling restless	24.difficult to breathe
5. feeling mentally sluggish	15.feeling to lose interest	25.thirsty
6. feeling sleepy	16.feeling of forgetfulness	26.hoarse voice
7. feeling your eyes tired	17.making many mistakes	27.feeling dizzy
8. feeling unable to coordinate	18.feeling worried	28.eyes twitching
9. feeling unsteady on your feet	19.feeling unable to be still	29.hands and legs trembling
10.feeling to lie down	20.feeling to lose your temper	30.feeling sick

B.5 Alate adelges data (Jeffers, 1967) 40×19

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
C1	21.2	11	7.5	4.8	5	2	2	2.8	2.8	3.3
C2	20.2	10	7.5	5	5	2.3	2.1	3	3	3.2
C3	20.2	10	7	4.6	5	1.9	2.1	3	2.5	3.3
C4	22.5	8.8	7.4	4.7	5	2.4	2.1	3	2.7	3.5
C5	20.6	11	8	4.8	5	2.4	2	2.9	2.7	3
C6	19.1	9.2	7	4.5	5	1.8	1.9	2.8	3	3.2
C7	20.8	11.4	7.7	4.9	5	2.5	2.1	3.1	3.1	3.2
C8	15.5	8.2	6.3	4.9	5	2	2	2.9	2.4	3
C9	16.7	8.8	6.4	4.5	5	2.1	1.9	2.8	2.7	3.1
C10	19.7	9.9	8.2	4.7	5	2.2	2	3	3	3.1
C11	10.6	5.2	3.9	2.3	4	1.2	1	2	2	2.2
C12	9.2	4.5	3.7	2.2	4	1.3	1.2	2	1.6	2.1
C13	9.6	4.5	3.6	2.3	4	1.3	1	1.9	1.7	2.2
C14	8.5	4	3.8	2.2	4	1.3	1.1	1.9	2	2.1
C15	11	4.7	4.2	2.3	4	1.2	1	1.9	2	2.2
C16	18.1	8.2	5.9	3.5	5	1.9	1.9	1.9	2.7	2.8
C17	17.6	8.3	6	3.8	5	2	1.9	2	2.2	2.9
C18	19.2	6.6	6.2	3.4	5	2	1.8	2.2	2.3	2.8
C19	15.4	7.6	7.1	3.4	5	2	1.9	2.5	2.5	2.9
C20	15.1	7.3	6.2	3.8	5	2	1.8	2.1	2.4	2.5
C21	16.1	7.9	5.8	3.7	5	2.1	1.9	2.3	2.6	2.9
C22	19.1	8.8	6.4	3.9	5	2.2	2	2.3	2.4	2.9
C23	15.3	6.4	5.3	3.3	5	1.7	1.6	2	2.2	2.5
C24	14.8	8.1	6.2	3.7	5	2.2	2	2.2	2.4	3.2
C25	16.2	7.7	6.9	3.7	5	2	1.8	2.3	2.4	2.8
C26	13.4	6.9	5.7	3.4	5	2	1.8	2.8	2	2.6
C27	12.9	5.8	4.8	2.6	5	1.6	1.5	1.9	2.1	2.6
C28	12	6.5	5.3	3.2	5	1.9	1.9	2.3	2.5	3
C29	14.1	7	5.5	3.6	5	2.2	2	2.3	2.5	3.1
C30	16.7	7.2	5.7	3.5	5	1.9	1.9	2.5	2.3	2.8
C31	14.1	5.4	5	3	5	1.7	1.6	1.8	2.5	2.4
C32	10	6	4.2	2.5	5	1.6	1.4	1.4	2	2.7
C33	11.4	4.5	4.4	2.7	5	1.8	1.5	1.9	1.7	2.5
C34	12.5	5.5	4.7	2.3	5	1.8	1.4	1.8	2.2	2.4
C35	13	5.3	4.7	2.3	5	1.6	1.4	1.8	1.8	2.5
C36	12.4	5.2	4.4	2.6	5	1.6	1.4	1.8	2.2	2.2
C37	12	5.4	4.9	3	5	1.7	1.5	1.7	1.9	2.4
C38	10.7	5.6	4.5	2.8	5	1.8	1.4	1.8	2.2	2.4
C39	11.7	5.5	4.3	2.6	5	1.7	1.5	1.8	1.9	2.4
C40	12.8	5.7	4.8	2.8	5	1.6	1.4	1.7	1.9	2.3

	V11	V12	V13	V14	V15	V16	V17	V18	V19
C1	3	4.4	4.5	3.6	7	4	8	0	3
C2	5	4.2	4.5	3.5	7.6	4.2	8	0	3
C3	1	4.2	4.4	3.3	7	4	6	0	3
C4	5	4.2	4.4	3.6	6.8	4.1	6	0	3
C5	4	4.2	4.7	3.5	6.7	4	6	0	3
C6	5	4.1	4.3	3.3	5.7	3.8	8	0	3.5
C7	4	4.2	4.7	3.6	6.6	4	8	0	3
C8	3	3.7	3.8	2.9	6.7	3.5	6	0	3.5
C9	3	3.7	3.8	2.8	6.1	3.7	8	0	3
C10	0	4.1	4.3	3.3	6	3.8	8	0	3
C11	6	2.5	2.5	2	4.5	2.7	4	1	2
C12	5	2.4	2.3	1.8	4.1	2.4	4	1	2
C13	4	2.4	2.3	1.7	4	2.3	4	1	2
C14	5	2.4	2.4	1.9	4.4	2.3	4	1	2
C15	4	2.5	2.5	2	4.5	2.6	4	1	2
C16	4	3.5	3.8	2.9	6	4.5	9	1	2
C17	3	3.5	3.6	2.8	5.7	4.3	10	1	2
C18	4	3.5	3.4	2.5	5.3	3.7	10	1	2
C19	4	3.3	3.6	2.7	6	4.2	8	1	3
C20	4	3.7	3.7	2.8	6.4	4.3	10	1	2.5
C21	5	3.6	3.6	2.7	6	4.5	10	1	2
C22	4	3.8	4	3	6.5	4.5	10	1	2.5
C23	5	3.4	3.4	2.6	5.4	4	10	1	2
C24	5	3.5	3.7	2.7	6	4.1	10	1	2
C25	4	3.8	3.7	2.7	5.7	4.2	10	1	2.5
C26	4	3.6	3.6	2.6	5.5	3.9	10	1	2
C27	5	2.8	3	2.2	5.1	3.6	9	1	3
C28	5	3.3	3.5	2.6	5.4	4.3	8	1	2
C29	5	3.6	3.7	2.8	5.8	4.1	10	1	2
C30	5	3.4	3.6	2.7	6	4	10	1	2.5
C31	5	2.7	2.9	2.2	5.3	3.6	8	1	2
C32	6	2.8	2.5	1.8	4.8	3.4	8	1	2
C33	5	2.7	2.5	1.9	4.7	3.7	8	1	2
C34	4	2.8	2.6	2	5.1	3.7	8	0	2
C35	4	2.7	2.7	2.1	5	3.6	8	1	2
C36	5	2.7	2.5	2	5	3.2	6	1	2
C37	5	2.7	2.7	2	4.2	3.7	6	1	2
C38	4	2.7	2.6	2	5	3.5	8	1	2
C39	5	2.6	2.5	1.9	4.6	3.4	8	1	2
C40	5	2.3	2.5	1.9	5	3.1	8	1	2

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

B.6 MDOC data (Sano et al., 1977) 87×23

	1 10	11 20	23
C1	4 4 4 4 4 4 4 4 4 4	4 4 4 4 4 3 4 3 4 4	3 2 2
C2	2 3 2 3 4 4 4 3 3 3	2 1 2 2 2 3 3 2 2 2	2 1 1
C3	3 3 2 2 1 1 1 2 2 2	2 3 2 1 2 2 2 2 2 2	1 1 1
C4	4 4 4 4 1 1 4 4 4 4	4 4 4 4 2 4 3 4 4 3	1 1 1
C5	2 4 2 4 4 4 1 2 3 4	4 2 1 2 2 2 2 2 2 2	2 1 1
C6	4 4 4 4 4 4 4 4 4 4	4 4 4 4 4 3 4 3 4 4	4 2 2
C7	2 1 2 3 1 4 4 2 3 2	2 2 2 1 2 1 1 2 1 1	2 1 1
C8	4 4 3 4 4 4 4 4 4 4	4 3 4 4 4 4 4 3 3 3	1 2 1
C9	1 2 2 3 1 4 1 2 2 3	4 1 2 4 2 2 2 1 2 1	1 1 1
C10	2 4 3 3 4 4 4 3 3 3	4 3 2 4 3 4 3 3 3 4	1 2 1
C11	1 1 1 4 4 4 1 2 1 2	3 1 1 4 2 1 1 1 2 2	2 2 1
C12	2 3 4 3 4 4 4 2 4 3	4 4 4 4 3 3 3 4 3 2	4 1 1
C13	1 1 1 3 1 1 1 1 1 1	2 1 1 1 2 2 1 1 1 1	1 1 1
C14	2 3 3 4 4 4 4 4 4 4	3 1 1 2 3 3 2 3 3 2	1 2 1
C15	1 1 1 2 4 1 1 2 2 2	3 1 1 2 2 2 2 1 1 1	1 1 1
C16	1 1 2 3 1 1 1 1 1 1	1 1 1 2 1 1 1 1 1 1	1 1 1
C17	1 4 3 3 1 1 4 2 1 1	4 1 1 1 2 2 1 2 1 1	1 1 1
C18	1 1 1 3 1 4 4 4 3 3	4 2 2 4 2 1 2 1 1 1	1 2 1
C19	1 4 2 3 1 1 4 2 4 2	2 1 1 2 2 2 1 2 1 1	1 1 1
C20	1 1 1 1 1 1 1 2 1 2	1 1 1 1 1 1 1 1 1 1	1 1 1
C21	1 1 1 1 1 4 1 1 1 1	4 1 1 2 1 1 1 1 2 1	1 1 1
C22	2 2 1 1 1 1 1 1 1 1	1 1 1 2 1 1 2 1 1 1	2 1 2
C23	3 2 3 3 1 4 4 3 4 4	2 3 1 2 2 3 2 3 2 2	2 1 2
C24	2 2 1 2 1 4 1 1 1 2	3 2 2 2 2 3 1 1 2 2	1 1 2
C25	1 2 1 2 1 1 4 1 2 2	3 1 1 2 2 2 1 2 1 1	1 1 1
C26	2 1 1 2 1 1 1 2 2 2	1 2 2 1 1 2 1 1 1 1	1 1 1
C27	2 1 1 1 1 1 1 1 2 2	1 1 2 1 1 1 1 1 1 2	2 1 2
C28	3 4 3 4 4 4 4 3 4 4	4 3 4 4 3 3 4 3 3 3	4 2 1
C29	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1
C30	1 1 1 1 1 1 1 1 1 2	1 1 1 2 1 1 1 1 1 1	2 1 1
C31	1 1 1 4 4 4 4 4 4 2	4 1 2 3 1 1 1 2 1 1	1 2 1
C32	3 4 3 4 4 4 4 3 3 4	4 3 4 4 3 1 3 2 1 2	4 2 1
C33	4 4 4 4 4 4 4 4 4 4	4 4 4 4 4 3 4 3 3 3	3 2 2
C34	2 4 2 3 4 4 1 2 4 2	4 1 3 4 3 3 2 2 2 2	1 2 2
C35	1 1 1 3 1 1 1 3 3 3	3 2 1 1 1 2 1 1 1 1	1 1 1
C36	1 1 1 2 1 1 1 1 1 2	2 1 1 1 2 1 1 2 1 1	1 1 1
C37	3 4 4 4 4 4 4 3 4 3	4 3 4 4 3 3 3 4 3 3	4 2 1
C38	4 4 4 4 4 4 4 4 4 4	4 4 4 4 4 3 4 3 4 4	1 2 2
C39	2 2 2 3 1 1 1 2 3 3	3 3 2 2 3 3 3 3 2 2	3 1 1
C40	2 2 2 2 1 1 1 2 1 2	1 1 1 2 2 2 2 2 2 2	2 1 1
C41	2 2 2 2 1 1 1 2 2 2	2 1 2 1 2 2 1 1 1 1	2 1 1
C42	3 3 3 3 1 1 4 2 2 3	3 2 2 3 2 2 2 2 2 2	2 1 1
C43	2 2 2 2 1 4 1 2 2 3	2 2 2 2 2 2 2 2 2 2	2 1 1
C44	1 1 1 2 1 1 1 2 2 2	1 1 1 1 2 2 1 1 1 1	2 1 1
C45	3 3 3 4 4 4 4 4 4 4	4 3 4 4 4 4 3 3 3 3	3 2 2
C46	2 2 2 2 1 1 4 2 2 2	2 2 2 2 2 3 2 2 2 2	2 1 1
C47	1 2 1 2 1 4 1 1 2 2	1 1 1 3 1 2 1 1 1 1	2 1 1
C48	1 2 1 2 4 4 4 2 2 3	3 3 2 3 1 2 2 2 2 1	1 1 1
C49	2 4 3 4 4 4 4 3 4 3	4 4 4 4 4 3 4 3 3 3	2 2 1
C50	1 1 1 1 1 1 1 1 2 1	1 1 1 1 1 1 1 1 1 1	1 1 1

C51	3 2 2 3 4 4 4 3 1 2	3 1 2 4 2 3 2 2 2 2	2 2 1
C52	3 1 2 2 1 1 1 2 1 3	2 1 2 2 2 3 2 2 2 2	2 1 1
C53	1 2 1 3 4 4 4 3 2 3	4 3 4 4 2 2 1 1 2 2	2 2 1
C54	2 1 1 1 1 4 4 1 1 2	1 3 2 1 2 2 1 1 1 1	2 1 1
C55	2 4 2 4 4 4 4 2 4 3	4 4 4 4 3 3 3 3 2 3	1 2 2
C56	3 1 2 2 1 4 4 2 3 3	3 3 2 3 2 3 1 2 2 2	2 1 1
C57	1 1 1 1 1 1 1 1 1 1	2 2 1 2 1 1 1 1 1 1	1 1 1
C58	3 3 2 3 4 4 1 3 3 3	2 3 2 2 2 2 2 3 2 2	3 2 1
C59	1 1 1 1 1 1 1 1 1 1	1 1 1 2 1 1 1 1 1 1	1 1 1
C60	1 3 2 3 4 4 4 2 2 3	4 3 4 4 3 3 3 2 3 2	4 1 1
C61	2 3 2 1 1 1 1 2 3 2	4 3 2 3 2 2 2 2 2 2	2 1 1
C62	2 1 2 2 1 1 1 1 2 2	2 2 2 1 2 2 2 2 2 2	1 1 1
C63	2 3 3 3 1 1 1 2 2 2	1 2 2 2 2 2 2 2 2 2	2 1 1
C64	1 1 2 1 1 4 1 1 1 2	1 2 2 2 2 3 2 2 2 3	2 1 1
C65	1 3 3 4 4 4 4 4 4 4	4 3 2 3 3 3 3 3 3 3	4 1 2
C66	1 2 2 3 1 1 1 2 1 2	2 2 1 2 1 2 2 1 2 1	2 2 2
C67	3 4 3 3 4 4 4 3 4 3	3 4 2 4 3 3 3 2 2 2	3 1 1
C68	4 4 3 4 4 4 4 4 4 3	4 4 4 4 4 4 3 3 3 3	1 2 2
C69	1 1 2 4 4 4 4 4 4 4	4 1 4 4 2 3 1 2 2 1	1 2 2
C70	4 4 3 4 4 4 4 4 4 4	4 4 4 4 4 4 4 4 4 4	1 2 2
C71	1 1 2 3 1 1 1 2 3 2	3 1 1 3 2 2 1 1 1 1	1 1 1
C72	4 4 3 4 4 4 1 3 4 4	4 3 4 4 3 3 3 3 3 2	1 2 1
C73	4 4 3 3 1 4 4 4 4 4	4 3 4 4 4 4 3 2 4 2	1 2 2
C74	1 1 1 3 1 4 1 2 2 2	1 1 1 1 1 2 1 1 1 1	1 1 1
C75	2 1 2 1 4 4 1 1 1 1	4 1 2 4 2 2 1 2 1 1	1 1 1
C76	3 4 3 4 4 4 4 4 4 4	4 3 3 4 3 3 3 4 3 3	4 2 2
C77	2 2 1 1 1 4 1 2 2 2	1 1 1 2 1 2 1 1 1 1	2 1 1
C78	2 2 3 4 4 4 4 4 4 4	4 3 4 4 3 3 4 3 3 3	4 2 2
C79	3 1 2 3 1 1 1 2 4 2	3 1 1 3 2 3 2 2 2 2	3 1 2
C80	2 2 1 1 1 1 1 1 1 2	1 1 1 2 1 2 1 1 1 1	2 1 1
C81	3 4 2 3 1 1 1 2 4 2	2 1 1 4 1 2 1 2 1 1	2 2 1
C82	3 4 2 3 4 4 4 3 4 3	4 3 1 4 2 3 3 3 2 3	2 2 2
C83	2 1 1 2 1 1 4 1 1 1	1 1 1 3 2 1 1 1 1 1	1 1 1
C84	2 1 1 2 1 1 1 1 1 1	2 1 1 1 1 1 1 1 1 1	2 1 1
C85	2 3 1 3 4 4 1 3 4 2	3 2 2 3 1 2 2 2 1 1	2 1 1
C86	1 1 1 2 1 4 1 2 2 2	2 1 2 2 1 1 1 1 2 1	1 1 1
C87	4 3 2 2 4 1 4 4 3 4	4 3 2 2 3 4 3 3 3 3	4 1 1

References

Ahamad, B. (1967). An analysis of crimes by the method of principal component analysis. *Appl. Statist.*, **16**, 17–35.

Arima, S. and Ishimura, S. (1987). *A story of multivariate analysis*, Tokyo Tosyo. (in Japanese)

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The S Language*. California: Wadsworth & Brooks/Cole Advanced Books & Software.

Castaño-Tostado, E. and Tanaka, Y. (1990). Some comments on Escoufier's RV-coefficient as a sensitivity measure in principal component analysis. *Comm. Statist.*, **A 19**, 4619–26.

Castaño-Tostado, E. and Tanaka, Y. (1991). Sensitivity measures of influence on the loading matrix in exploratory factor analysis. *Comm. Statist.*, **A 20**, 1329–43.

Critchley, F. (1985). Influence in principal component analysis. *Biometrika*, **72**, 627–36.

Hampel. F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383–93.

Iwasaki, M. (1989). Analysis of test score by Hayashi's third method of quantification. *Japanese J. Behaviormetrics*, **16**, 2, 13–21. (in Japanese)

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.

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

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

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

Krzanowski, W. J. (1988). *Principles of multivariate analysis: a user's perspective*. Oxford University Press.

Maehashi, A. et al. (1993). A trial of making "The Questionnaire of Subjective Symptoms of Fatigue for School-children". *J. Health Education of Children*, **2**, 51–70. (in Japanese)

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

Mori, Y. and Tarumi, T. (1993). Statistical software SAM II: Sensitivity analysis in multivariate methods. *J. Japanese Soc. Comp. Statist.*, **6**(2), 21–32.

Mori, Y. and Tarumi, Y. (1994). Variable selection in Hayashi's third method of quantification. *Proceedings of the 5th Japan-China Symposium on Statistics*, 191–194.

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

Mori, Y., Tarumi, T. and Tanaka, Y. (1994b). Variable selection with RV-coefficient in principal component analysis. *Bull. Comp. Statist. of Japan*, **7**, 47–56. (in Japanese)

Okamoto, M. (1992). Some problems on Hayashi's third method of quantification. *J. Japan Statist. Soc.*, **22**, 229–239. (in Japanese)

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**, 257–265.

Sano, K. et al. (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)

Tanaka, Y. (1983). Some criteria for variable selection in factor analysis. *Behaviormetrika*, **13**, 31–45.

Tanaka, Y. (1988). Sensitivity analysis in principal component analysis: Influence on the subspace spanned by principal components. *Comm. Statist.*, **A 17**, 3157–75. (Corrections, **A 18**(1989), 4305).

Tanaka, Y. (1989). Influence functions related to eigenvalue problems which appear in multivariate analysis. *Comm. Statist.*, **A 18**, 3991–4010.

Tanaka, Y. (1992). Sensitivity analysis in multivariate methods. *Jap. J. of Behaviormetrics.*, **19**, 3–17. (in Japanese)

Tanaka, Y., Castaño-Tostado, E. and Odaka, Y. (1990). Sensitivity analysis in factor analysis: Methods and software. *COMPSTAT 1990* (Edited by Momirović, K. and Mildner, V.), Physica-Verlag, 205–10.

Tanaka, Y. and Kodake, K. (1981). A method of variable selection in factor analysis and its numerical investigation. *Behaviormetrika*, **10**, 49–61.

Xia, L. and Yang, Y. (1988). A method of variable selection in Hayashi's third method of quantification. *J. Japanese Soc. Comp. Statist.*, **1**, 27–43.

Yamada, F. and Nishisato, S. (1993). Several mathematical properties of dual scaling as applied to dichotomous item-category data. *Japanese J. Behaviormetrics*, **20**, 1, 56–63. (in Japanese)

Acknowledgements

I would like to express my deep gratitude to Professor Tomoyuki Tarumi of Department of Environmental and Mathematical Sciences, Okayama University, for his continuous teaching, guidance and encouragement on my study. I also express my sincere appreciation to Professor Yutaka Tanaka of Department of Environmental and Mathematical Sciences, Okayama University, for his valuable teaching, advice and encouragement on my study.

Furthermore, grateful appreciation and sincere thanks are extended to the late Professor Kazumasa Wakimoto, Okayama University, who gave me the opportunity to study in the doctor course, for his teaching, suggestions and encouragement.

Finally, I am very grateful and sincere thanks to all the members and staff of Okayama Statistical Association for their kindness, assistance and various advices.