PRINCIPAL VARIABLES

by

George P. McCabe*

Technical Report #82-3

Department of Statistics

February 1982

*George P. McCabe is Associate Professor of Statistics and Head of Statistical Consulting, Statistics Department, Purdue University, West Lafayette, IN 47907. Part of this work was completed while the author was visiting the University of Western Australia, Perth, Western Australia.

Abstract

One of the often stated goals of principal component analysis is to reduce into a low dimensional space, most of the essential information contained in a high dimensional space. According to several reasonable criteria, principal components do this optimally. From a practical point of view, principal components suffer from the disadvantage that each component is a linear combination of all of the original variables. An alternative approach is to select a subset of variables which contain, in some sense, as much information as possible. Methods for performing such an analysis are presented and illustrated with examples.

KEY WORDS: Principal components, Variable selection.

1. INTRODUCTION

Principal components analysis is a mathematically appealing statistical tool for examining variables in high dimensional spaces. Unlike its poor relative, factor analysis, it enjoys a solid theoretical foundation and possesses many optimal properties.

Unfortunately, however, principal components are often difficult to interpret in practical situations. Although they solve a well-defined mathematical problem, they frequently fail to provide the statistical consumer with concrete useful results. This difficulty explains the popularity of various procedures which are used in an attempt to span the selected subspace with a meaningful coordinate system.

In addition, principal components suffer from another major deficiency. In general, each component is a linear combination of all of the original variables. Thus, although the dimensionality of the space may be reduced by selecting components, one must still interpret results about a large number of variables. In many applications, it is desirable not only to reduce the dimension of the space, but also to reduce the number of variables which are to be considered or measured in the future.

This deficiency has not gone unnoticed by those who have studied principal components. Srivastava and Khatri (1979) state: "The selection of principal components with the largest variances, however, may have the disadvantage that all of the original variables (or almost all of them) may enter into some of the selected principal components with nonzero weights." Some attempts to remedy this deficiency are given by Beale, Kendall and Mann (1967) and Jolliffee (1972, 1973).

In view of these difficulties, it is natural to ask whether or not the desirable characteristics of principal components can be obtained while simultaneously reducing the number of variables to be considered. This paper is an attempt to address this question.

I have been motivated to consider this problem by numerous consulting encounters in which I have attempted to convince users of statistics to try principal components on their data. After an explanation of the procedure and its properties, many have been sufficiently unimpressed to politely reply "no thank you". Other more adventurous types have run the appropriate computer programs and returned with the output seeking further guidance. I have explained that these linear combinations are the solution. To their obvious next question, "What do they mean?" I reply that such questions are not statistical and that they must interpret the results. At this point, the discussion of principal components usually terminates with a feeling of time wasted by both consultant and client.

An encounter with a forester helped me to see clearly the inadequacy of principal components with his data. There are many measured variables which can be used to characterize the size of a tree. Among these are total shaded area, length of leaves, number of leaves, height, girth at various heights, volume of roots, etc. Some of these variables are easy to measure, for example, height at three feet; some are more difficult, perhaps requiring a team of graduate students, such as height at 20 feet; some are destructive to the tree and extremely difficult to obtain, such as total root area. Experienced foresters generally measure a half-dozen or less variables which, in their judgement, contain the basic information about tree sizes. Of course, with any such selection, essential

information can be lost. However, such is also the case with the discarded principal components.

2. PERSPECTIVES ON THE PROBLEM

There are several different views of principal components which lead to different types of optimality properties. These views are basically intuitive ideas which have been translated into precise mathematical forms. To fully appreciate principal components and the alternatives presented in this paper, it is important to keep in mind the distinction between the intuitive idea and its mathematical translation, which although precise internally, may be only an approximation to the more fundamental intuitive idea.

Karl Pearson (1901) gives us one idea: "...it is desirable to represent a system of points in plane, three, or higher dimensional space by the 'best-fitting' straight line or plane." He addresses the dilemma which arises upon consideration of the fact that the regressions of Y on X and X on Y give different lines. Principal components give the perpendicular distance solution. Thus, Pearson's view is that of fitting points in a space.

Harold Hotelling (1933) offers a different view. He writes of "a fundamental set of independent variables...which determine which values the x's will take." Thus, for Hotelling, prediction of the original set of variables is a fundamental concept.

C. R. Rao (1964) presents a third view: "When a large number of measurements are available, it is natural to inquire whether they could be replaced by a fewer number of the measurements or of their functions, without loss of much information..." Note that the possibility of variable selection is mentioned, in addition to the linear combination idea.

The variable selection theme, however, is not further pursued. For Rao, the idea of minimizing information loss is fundamental.

In summary, the literature cited above gives three views: (1) fitting points in space; (2) predicting the original variables; and (3) minimizing information loss. These ideas have various mathematical translations which lead to a principal components analysis as the optimal solution.

3. NOTATION AND ASSUMPTIONS

Let X be a p-dimensional normal random vector with mean zero and positive definite covariance matrix Σ. In most applications, Σ will be in correlation form although this assumption is not necessary in what follows.

We wish to consider dimension-reducing linear transformations of X to a random variable Y. Therefore, we let

$$Y = A_k X, \tag{3.1}$$

where A_k is a $k \times p$ matrix with $k < p$. Thus, Y is a k-dimensional random variable. We can, without loss of generality, neglect a translation term in the transformation and furthermore, assume that

$$A_k' A_k = I_k, \tag{3.2}$$

where I_k is the $k \times k$ identity matrix. Note that the random variable Y is normal with mean zero and covariance matrix

$$\Sigma_Y = A_k' \Sigma A_k. \tag{3.3}$$

Suppose that we observe the random variable Y and want to predict the random variable X. We denote the predicted random variable by Z. Since we have assumed a multivariate normal distribution for X, the best predictor is simply the conditional expectation of X given Y:

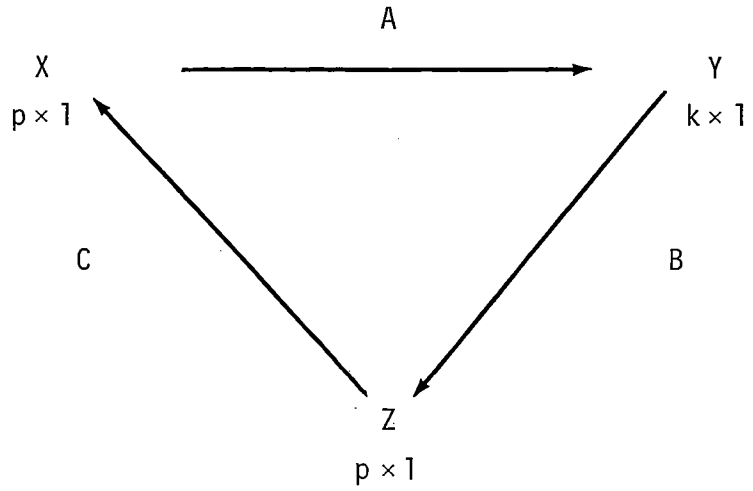
$$Z = (\Sigma A_k)(A_k'\Sigma A_k)^{-1}Y. \tag{3.4}$$

Clearly Z is a normal random variable with mean zero and covariance matrix,

$$\Sigma_Z = \Sigma A_k(A_k'\Sigma A_k)^{-1}A_k'\Sigma \ . \tag{3.5}$$

Note that Σ_Z is a $p \times p$ singular matrix, since $k < p$.

Some insight can be obtained by considering the following schematic:



The path A is determined by our choice of the matrix A_k. The path B is routine, following from the standard multivariate normal theory. Path C is one key to evaluating how well we travelled path A. The extent to which Z approximates X is a measure which can be used to select among the various choices for path A.

A basic tool for studying principal component properties is the spectral decomposition of a positive definite matrix. Thus, the matrix Σ can be decomposed as follows:

$$\Sigma = \lambda_1 g_1 g_1' + \lambda_2 g_2 g_2' + \ldots + \lambda_p g_p g_p' \ , \tag{3.6}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ are the ordered eigenvalues of Σ and $g_1, \ldots, g_p$ are the associated eigenvectors.

Let $\Lambda = \text{diag}(\lambda_1,\ldots,\lambda_p)$ and $G = (g_1,\ldots,g_p)$. Then, $G'G = I$, $G\Lambda G' = \Sigma$, $G'\Sigma G = \Lambda$, $|\Sigma| = \Pi_{i=1}^{p} \lambda_i$ and $\text{tr}(\Sigma) = \Sigma_{i=1}^{p} \lambda_i$. For convenience in what follows, let

$$G_k = (g_1,\ldots,g_k). \tag{3.7}$$

4. OPTIMAL PROPERTIES OF PRINCIPAL COMPONENTS

In most multivariate texts, principal components are introduced as orthogonal linear combinations having maximum variance subject to constraints. The derivation follows from a routine application of Lagrange multipliers. See, for example, Anderson (1958), Harris (1975), Morrison (1976) and Timm (1975).

There are many other optimal properties of principal components. A collection of these is discussed below. Proofs of most of these properties can be found in Kshirsager (1972) or Okamoto (1969). The rest follow directly from their results.

The ten properties below each involve a maximization or minimization. In all cases, this max or min is taken over matrices A_k to be used as in (3.1) and subject to the constraint (3.2) for fixed values of k. A solution to each of these problems is $A_k = G_k$, i.e. principal components. Other solutions may be obtained by premultiplying G_k by a $k \times k$ orthogonal matrix. To better understand the relationships among the criteria, the maximizing or minimizing values are also given. Also, the properties are grouped into broad intuitive categories.

4.1 Retention of Variation

$$\max \text{tr}(\Sigma_Y) = \Sigma_{i=1}^{k} \lambda_i. \tag{4.1}$$

This first criterion is the traditional view of principal components. It expresses the concept that the sum of the variances of the Y's should be large.

$$\max|\Sigma_Y| = \Pi_{i=1}^{k} \lambda_i. \tag{4.2}$$

Since the determinant of a covariance matrix is the generalized variance, this criterion is similar to the first but here, variance is expressed in a multivariate sense.

For the following, we assume that X_1 and X_2 are iid with the same distribution as X and that $Y_j = A_k X_j$ for j=1,2.

$$\max E(Y_1-Y_2)'(Y_1-Y_2) = 2\Sigma_{i=1}^{k} \lambda_i. \tag{4.3}$$

$$\max|E(Y_1-Y_2)(Y_1-Y_2)'| = 2\Pi_{i=1}^{k} \lambda_i. \tag{4.4}$$

These two criteria are related to the idea that points in the transformed space should be kept as far apart as possible, thereby retaining the variation in the original space. This idea can be formalized as separation for each component as in (4.3), or in the more general multivariate sense as in (4.4).

The four criteria above are all concerned with path A in the schematic of Section 3.

4.2 Retention of Correlational Structure

Let the norm of a matrix be defined as the sum of squares of its elements. Thus, if $M = (m_{ij})$ then $||M||^2 = \Sigma\Sigma m_{ij}^2$.

$$\min||\Sigma_X-\Sigma_Z||^2 = \Sigma_{i=k+1}^{p} \lambda_i^2. \tag{4.5}$$

This criterion is focused on path C in the schematic of Section 3. When we transform to Y and then attempt to return through Z, we would

like to reproduce as much of the original correlational structure as possible. This criterion suggests minimizing the loss as defined by the above norm.

4.3 Loss of Variation

$$\min \operatorname{tr}(\Sigma_{X|Y}) = \Sigma_{i=k+1}^{p} \lambda_i. \tag{4.6}$$

$$\min ||\Sigma_{X|Y}||^2 = \Sigma_{i=k+1}^{p} \lambda_i^2 . \tag{4.7}$$

By considering the conditional covariance matrix of X given Y, these two criteria address the variation not extracted by the transformation. This can be done by summing over the original variables as in (4.6) or in the norm sense as in (4.7). Note that the matrix $\Sigma_{X|Y}$ is singular.

4.4 Predictive Capacity

$$\max \Sigma_{j=1}^{p} \sigma_{jj} R^2(\chi_j, Y) = \Sigma_{i=1}^{k} \lambda_i. \tag{4.8}$$

In this criterion, which is due to Okamoto(1969), σ_{jj} is the (j,j)-th element of Σ and $R^2(\chi_j, Y)$ is the squared multiple correlation of χ_j with the k-dimensional vector Y. The idea here is that it is desirable to be able to predict the original variables from the retained components.

$$\min E(Z-X)'(Z-X) = \Sigma_{i=k+1}^{p} \lambda_i. \tag{4.9}$$

$$\min ||E(Z-X)(Z-X)'||^2 = \Sigma_{i=k+1}^{p} \lambda_i^2. \tag{4.10}$$

These criteria focus on path C of the schematic. A good transformation should be able to reproduce the original random variable well. The extent to which this is not done is expressed by Z - X. Thus, these criteria attempt to make this quantity small in a component by component sense (4.9) and in the norm sense (4.10).

4.4 Eigenvalues and Eigenvectors

All of the above criteria give extrema which are simple functions of the eigenvalues of Σ. Since these can be ordered, it is quite natural that principal components, with the corresponding eigenvectors as coefficient vectors, are optimal. The results are mathematically attractive, intuitively reasonable, and relatively easy to compute.

In the next section, an attempt is made to apply these criteria to the variable selection problem. Unfortunately, the results are neither mathematically attractive nor easy to compute. However, they are, hopefully, more applicable.

5. PRINCIPAL VARIABLES

Principal variables are defined as a subset of variables which possess some optimality property. To select variables, we start with the framework of (3.1) and (3.2) and add the condition that A_k is of the form,

$$A_k = (I_k, O),$$

or a matrix obtained by permuting the columns of this matrix. In the above, O is a $k \times (p-k)$ matrix of zeros.

Instead of permuting the columns of A_k, however, it is more convenient to consider permuting the elements of X. Therefore, we consider all possible partitions of X into $(X_1', X_2')'$ where X_1 is a k-vector of variables retained and X_2 is a (p-k)-vector of variables discarded. Let Σ be partitioned in the obvious way,

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where Σ_{11} is the $k \times k$ covariance matrix of X_1. Selection of a set of k

variables is thus equivalent to selection of a $k \times k$ matrix Σ_{11} from the $\binom{p}{k}$ possible choices. Note that there are $2^p - 1$ choices for all $k=1,\ldots,p$.

We now consider the optimality criteria of the previous section. Each of these will lead to an optimal choice of Σ_{11}. In addition, two other criteria will be considered.

5.1 Retention of Variation

$$\max \operatorname{tr}(\Sigma_y) = \max \operatorname{tr} \Sigma_{11}. \tag{5.1}$$

Since $\Sigma_y = \Sigma_{11}$, it follows that

$$\operatorname{tr} \Sigma_y = \Sigma_{i=1}^{k} \sigma_{ii}$$

where $\sigma_{11} \geq \sigma_{22} \geq \ldots \geq \sigma_{pp}$ be the ordered variances. Thus, X_1 consists of the variables with the k largest variances. Although this first criteria is the usual starting point for principal components, its application in the present context yields a rather uninteresting result. Moreover, if Σ is in correlation form then $\sigma_{ii} = 1$ for all i and all sets of k variables (fixed k) are equivalent.

$$\max|\Sigma_Y| = \max|\Sigma_{11}|. \tag{5.2}$$

Here, the idea is to select variables which maximize the generalized variance. Let

$$\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12},$$

be the conditional covariance of X_2 given X_1. Then,

$$|\Sigma| = |\Sigma_{11}| \cdot |\Sigma_{22.1}| .$$

Hence maximizing $|\Sigma_{11}|$ is equivalent to minimizing $|\Sigma_{22.1}|$.

$$\max E(Y_1-Y_2)'(Y_1-Y_2) = 2 \max \operatorname{tr} \Sigma_{11}. \tag{5.3}$$

$$\max|E(Y_1-Y_2)'(Y_2-Y_1)| = 2\max|\Sigma_{11}|. \tag{5.4}$$

Since the covariance matrix of $X_1 - X_2$ is simply $2\Sigma_{11}$, these results are equivalent to (5.1) and (5.2), respectively.

5.2 Retention of Correlational Structure

$$\min||\Sigma_X-\Sigma_Z||^2 = \min \Sigma_{i=1}^{p-k} \theta_i^2 \tag{5.5}$$

where $\theta_1,\ldots,\theta_{p-k}$ are the eigenvalues of $\Sigma_{22\cdot1}$. To see (5.5) we first observe that

$$Z = \begin{pmatrix} X_1 \\ EX_2|X_1 \end{pmatrix}$$

$$= \begin{pmatrix} I \\ \Sigma_{21}\Sigma_{11}^{-1} \end{pmatrix} X_1 .$$

Therefore,

$$\Sigma_Z = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix},$$

and

$$\Sigma_X - \Sigma_Z = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{22\cdot1} \end{pmatrix}.$$

Obviously,

$$||\Sigma_X-\Sigma_Z||^2 = ||\Sigma_{22\cdot1}||^2 .$$

But,

$$||\Sigma_{22\cdot1}||^2 = \mathrm{tr}(\Sigma_{22\cdot1}\Sigma_{22\cdot1}).$$

Now, let

$$\Sigma_{22\cdot1} = Q \Theta Q',$$

where the columns of Q contain the eigenvectors of $\Sigma_{22\cdot 1}$ and $\Theta = \mathrm{diag}(\theta_1,\ldots,\theta_{p-k})$. Thus,

$$\begin{aligned} \mathrm{tr}(\Sigma_{22\cdot 1}\Sigma_{22\cdot 1}) &= \mathrm{tr}(Q\Theta Q'Q\Theta Q') \\ &= \mathrm{tr}(Q\Theta^2 Q') \\ &= \mathrm{tr}\,\Theta^2 \\ &= \Sigma_{i=1}^{p-k}\theta_i^2 . \end{aligned}$$

Note that with the above notation, criteria (5.2) and (5.4) are equivalent to

$$\min \Pi_{i=1}^{p-k}\theta_i .$$

5.3 Loss of Variation

$$\min \mathrm{tr}\,\Sigma_{X|Y} = \min \Sigma_{i=1}^{p-k}\theta_i . \tag{5.6}$$

$$\min ||\Sigma_{X|Y}||^2 = \min \Sigma_{i=1}^{p-k}\theta_i^2 . \tag{5.7}$$

Since

$$\Sigma_{X_1,X} = \begin{pmatrix} \Sigma_{11} & \Sigma_{11} & \Sigma_{12} \\ \Sigma_{11} & \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{21} & \Sigma_{22} \end{pmatrix} ,$$

it immediately follows that

$$\Sigma_{X|X_1} = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{22\cdot 1} \end{pmatrix} .$$

Results (5.6) and (5.7) are consequences of this fact.

5.4 Predictive Capacity

$$\max \Sigma_{j=1}^{p} \sigma_{jj} R^2(\chi_j, Y) = \Sigma_{j=1}^{p} \lambda_j - \min \Sigma_{i=1}^{k-p} \theta_i . \tag{5.8}$$

Since the λ_j are fixed, this criteria is equivalent to (5.6). Since $Y = X_1$,

$$R^2(\chi_i, Y) = \begin{cases} 1 & \text{for } i=1,\ldots,k \\ 1 - \dfrac{\sigma_{ii \cdot Y}}{\sigma_{ii}} & \text{for } i=k+1,\ldots,p \end{cases},$$

where $\sigma_{ii \cdot Y}$ is the conditional variance of χ_i given Y. Thus,

$$\begin{aligned} \Sigma_{j=1}^{p} \sigma_{jj} R^2(\chi_j, Y) &= \Sigma_{j=1}^{p} \sigma_{jj} - \Sigma_{i=1}^{k-p} \sigma_{ii \cdot Y} \\ &= \operatorname{tr} \Sigma - \operatorname{tr} \Sigma_{22 \cdot 1} . \end{aligned}$$

The result immediately follows.

$$\min E(Z-X)'(Z-X) = \min \Sigma_{i=1}^{p-k} \theta_i , \tag{5.9}$$

$$\min ||E(Z-X)(Z-X)'||^2 = \min \Sigma_{i=1}^{p-k} \theta_i^2 . \tag{5.10}$$

Since

$$Z - X = \begin{pmatrix} 0 & 0 \\ \Sigma_{21}\Sigma_{11}^{-1} & -I \end{pmatrix} X ,$$

it follows that

$$\Sigma_{Z-X} = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{22 \cdot 1} \end{pmatrix},$$

and the results (5.9) and (5.10) are immediate.

5.5 Other Criteria

Let Z be partitioned as X: $Z = (Z_1', Z_2')'$. By definition, $Z_1 = X_1$ and $Z_2 = EX_2 | X_1$. Also let $m = \min(k, p-k)$ and $\rho_1^2, \ldots, \rho_m^2$ be the canonical

correlations between X_1 and X_2. Then,

$$\min E(Z_2-X_2)'\Sigma_{22}^{-1}(Z_2-X_2) = (p-k) - \max \Sigma_{i=1}^{m} \rho_i^2. \tag{5.11}$$

Roughly, this criteria suggests that Z_2 and X_2 should be close in the natural metric for these variables.

To prove (5.11) we first note that

$$\begin{aligned} E(Z_2-X_2)'\Sigma_{22}^{-1}(Z_2-X_2) &= \text{tr } E\Sigma_{22}^{-1}(Z_2-X_2)(Z_2-X_2)' \\ &= \text{tr } \Sigma_{22}^{-1}\Sigma_{22\cdot 1} \\ &= \text{tr } I_{p-k} - \text{tr}(\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}). \end{aligned}$$

The first term is simply (p-k). The second term can be examined by considering the eigenvalues for the matrix given. Let ρ^2 denote such an eigenvalue. The defining determinantal equation is

$$|\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}-\rho^2 I| = 0.$$

However, this equation is equivalent to

$$|\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}-\rho^2\Sigma_{22}| = 0,$$

which is the determinantal equation for the canonical correlations. Equation (5.11) follows immediately.

Criterion (5.11) therefore suggests that variables should be selected so as to maximize the squared canonical correlations between the retained and discarded variables.

A final criterion is suggested by the above. If we focus on Z, we can ask that these projections be kept apart for pairs of observations in an expected value sense. Therefore, let Z_1 and Z_2 be the p-dimensional random variables obtained from a pair of iid X variables by transformations (3.1) and (3.4).

$$\max E(Z_1-Z_2)'(Z_1-Z_2) = 2[\Sigma_{j=1}^{p} \lambda_i - \min \Sigma_{i=1}^{p-k} \theta_i]. \tag{5.12}$$

Since Z is obtained from X by the transformation

$$Z = \begin{pmatrix} I & 0 \\ \Sigma_{21}\Sigma_{11}^{-1} & 0 \end{pmatrix} X ,$$

it follows that

$$E(Z_1-Z_2)'(Z_1-Z_2) = 2 \operatorname{tr}\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix}$$

$$= 2[\operatorname{tr} \Sigma_{11} + \operatorname{tr}(\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})]$$

$$= 2[\operatorname{tr} \Sigma - \operatorname{tr} \Sigma_{22.1}].$$

Equation (5.12) follows immediately from the above equality.

Note that this criteria is equivalent to minimizing the sum of the eigenvalues of $\Sigma_{22.1}$.

5.6 Summary of Criteria

Principal components give the optimal solution for the ten criteria given in section 4. When these criteria and the two additional ones of the preceeding subsection are applied in the variable subset context, differing sets of variables can arise as optimal solutions. Each of the criteria is equivalent to one of the following:

$$\min|\Sigma_{22.1}| = \min \Pi_{i=1}^{p-k} \theta_i , \tag{5.13}$$

$$\min \operatorname{tr}(\Sigma_{22.1}) = \min \Sigma_{i=1}^{p-k} \theta_i , \tag{5.14}$$

$$\min||\Sigma_{22.1}||^2 = \min \Sigma_{i=1}^{p-k} \theta_i^2 , \tag{5.15}$$

or

$$\max \Sigma_{i=1}^{m} \rho_i^2 . \tag{5.16}$$

Note that

$$|\Phi_{22\cdot 1}| = |\Phi_{22}| \Pi_{i=1}^{m} (1-\rho_i^2) .$$

Choices among the four criteria may depend upon the particular application. Computational difficulty is another aspect which is relevant.

In principal components, results are often described by a percentage of variation explained. In this context, the trace criterion is natural. The form given by (5.8) leads to

$$P = \left(1 - \frac{\Sigma_{i=1}^{k-p} \theta_i}{\Sigma_{i=1}^{p} \lambda_i} \right) 100\% , \tag{5.17}$$

where P is the percentage of variation explained. Equivalently, (5.17) can be rewritten as

$$P = \frac{\Sigma_{j=1}^{k} \sigma_{jj} + \Sigma_{j=k+1}^{p} \sigma_{jj} R^2(\chi_j, Y)}{\Sigma_{j=1}^{k} \sigma_{jj}} \ 100\% . \tag{5.18}$$

If the correlation form of the matrix is used, this reduces to

$$P = p^{-1}(k + \Sigma_{j=k+1}^{p} R^2(\chi_j, Y)) 100\% . \tag{5.19}$$

6. COMPUTATIONS

The determinential criteria (5.13) can be evaluated easily for all possible subsets when p is not greater than about 20. The computation time grows as 2^p. The algorithm given in McCabe (1975) can be used for this purpose. For 19 variables, computations took about 110 seconds on a CDC6500 computer. The cost, at internal rates, was $2.50. The idea of giving, for example, the top 10 subsets of each size is very appealing from a practical point of view. In this way, a user can observe patterns in which a particular variable shows up consistently in the subsets or

where variables or collections of variables are interchangeable.

For the other three criteria, the computations are substantially more complicated and hence, examination of all subsets appears to be impractical. The trace criterion, however, suggests a natural step-type algorithm:

Let the selected variables be labelled $x_{(1)},x_{(2)},\ldots,x_{(p)}$. Then let $x_{(1)}$ be such that

$$\Sigma_{i=1}^{p} \sigma_{ii}R^2(x_i;x_{(1)})$$

is maximized. Let $x_{(2)}$ be such that

$$\Sigma_{i=1}^{p} \sigma_{ii}R^2(x_i;x_{(1)},x_{(2)})$$

is maximized. At step j, let $x_{(j)}$ be such that

$$\Sigma_{i=1}^{p} \sigma_{ii}R^2(x_i,x_{(1)},\ldots,x_{(j)})$$

is maximized.

Of course, if a correlation matrix is used then the σ_{ii} drop out of the above expressions. This step algorithm will not assure that the trace is minimized for any given step k. However, it can be used to select variables which explain a large proportion of the variation.

Modifications to the above algorithm could be made to increase the chances of finding the optimal subsets. Variables could be deleted or variables could be added several at a time.

7. EXAMPLES

To evaluate the procedures described in the previous sections, seven sets of real data are analyzed. First, a computer program which computes and ranks determinants for all possible subsets was used. Second,

regressions were run to determine the percentage of variation explained by the various subsets selected in the first step. These percentages can be compared to the variation explained by the first few eigenvectors.

7.1 Data Set FIS

The first data set is due to Fisher (1936) and is analyzed in Anderson (1958). There are four size measurements on Iris versicolor.

Results are presented in Table 1 for both the correlation and covariance matrices. As expected, these two matrices give different results in the rank order of the subsets. Also, the determinant and percentage of variation give different orders. The choice of matrix form and criterion for selection of subset should be made by the experimenter. Overall, however, the variable selection procedure compares favorably to the principal component analysis. For example, with the covariance matrix, the first two principal components explain 89.7% while variables 1 and 3 explain 87.3%.

7.2 Data Set J1.

Jeffers (1967) analyzed two sets of data to illustrate the use of principal components. The first of these is J1 and consists of 13 measurements on pitprops.

Table 2 gives the results. Jeffers suggests using the first four components (73.7%) or the first six (87.0%). If variables rather than components are used, then the six variables 2,3,5,11,12,13, which account for 79.2%, give slightly better performance than the first four components. The price paid for the increased simplicity of working with variables rather than components is that two extra dimensions are needed.

7.3 Data Set J2

Jeffers' (1967) example concerned a study of characteristics of winged aphids. Nineteen different measurements were analyzed, including a binary variable (number 18).

The results are presented in Table 3. Jeffers suggests that only the first two components, which explain 85.4% of the variation, have practical significance. There are clearly several sets of five variables which explain as much variation. Furthermore, the four variables 6,11,17, 19 explain 84.7%.

7.2 Data Set AH

Ahamad (1967) analyzed crime statistics using principal components. There are 18 variables measuring different types of crimes. Since there are only 14 observations in this data set, the full matrix is singular and some caution is necessary to interpret the results properly.

In Table 4, the results are given. Clearly, three or four variables contain most of the information.

7.5 Data Set CHE

Cheetham (1973) used principal components to study polymorphism in cheilostomes. The 11 variables used and details of the analysis are described in his paper. The correlation matrix is given in Table 8.

Cheetham works with the first four components which account for 77.6% of the variation. Once can get reasonably close to this figure with five variables and exceed it with a variety of sets of six variables. (Table 5)

7.6 Data Set ORH

Orheim (1981) studies the constituent elements in coal samples. The correlation matrix for the amounts of nine elements is given in Table 9.

The results are presented in Table 6. The principal components and variable selection methods compare favorably. The number of variables needed to match the component variation explained is generally one more than the number of components.

7.7 Data Set CS

Campbell and McCabe (1982) have studied various characteristics of freshman computer science majors in an attempt to discriminate between those who persist as majors and those who drop out. The pooled correlation matrix for this study is given in Table 10.

Table 7 gives results for this data set. As with the previous example, one or two extra variables give similar percentages of variation explained relative to the principal components.

7.8 Summary of Results

Tables 1 to 7 generally indicate that the principal variable method gives results comparable to principal components analysis. Since, for any given number of dimensions, principal components explain the maximum variation percentage, an equal number of variables cannot explain more. However, by considering a small number of additional variables, comparable results are obtained. The simplicity of dealing with variables rather than linear combinations would seem to justify this increase in many applications.

For interpretability, the percentage of variation explained appears to be most suitable. However, the determinantial criterion is easily computed. A reasonable compromise is to use the determinant to screen for good subsets and then evaluate them in terms of variation explained.

The final choice of variables should be left to the researcher who knows and understands the variables. Tables, such as those given herein,

facilitate this choice and allow the incorporation of other information, such as cost of measurement, into the selection process.

REFERENCES

Ahamad, B. (1967). "An Analysis of Crimes by the Method of Principal Components," Applied Statistics, Journal of the Royal Statistical Society, Ser. C., 16, 17-35.

Anderson, T. W. (1958). Introduction to Multivariate Statistical Analysis, New York: Wiley.

Beale, E. M. L., Kendall, M. G. and Mann, D. W. (1967). "The Discarding of Variables in Multivariate Analysis," Biometrika, 54, 357-366.

Campbell, Patricia F. and McCabe, George P. (1982). Personal communication.

Cheetham, Alan H. (1973). "Study of Cheilostome Polymorphism Using Principal Components Analysis," Living and Fossil Bryozoa, (Lakewood, G. P., ed.), London: Academic Press, 385-409.

Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, 7, 179-188.

Harris, R. J. (1975). A Primer of Multivariate Statistics, New York: Academic Press.

Hotelling, H. (1933). "Analysis of a Complex of Statistical Variables into Principal Components," Journal of Educational Psychology, 24, 417-441.

Jeffers, J. N. R. (1967). "Two Case Studies in the Application of Principal Component Analysis," Applied Statistics, Journal of the Royal Statistical Society, Ser. C., 16, 225-36.

Jolliffe, I. T. (1972). "Discarding Variables in a Principal Component Analysis I: Artificial Data," Applied Statistics, Journal of the Royal Statistical Society, Ser. C., 21, 160-163.

Jolliffe, I. T. (1973). "Discarding Variables in a Principal Component Analysis II: Real Data," Applied Statistics, Journal of the Royal Statistical Society, Ser. C, 22, 21-31.

Kshirsagar, A. M. (1972). Multivariate Analysis, New York: Dekker.

McCabe, George P., Jr. (1975). "Computations for Variable Selection in Discriminant Analysis," Technometrics, 17, 103-109.

Morrison, D. F. (1976). Multivariate Statistical Methods, New York: McGraw-Hill.

Okamoto, M. (1969). "Optimality of Principal Components," in Multivariate Analysis - II, (P.R. Krishnaiah ed.), New York: Academic Press, 673-685.

Orheim, Alv (1981). Personal communication.

Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space," Philosophical Magazine, Ser. 6, 2, 559-572.

Rao, C. R. (1964). "The Use and Interpretation of Principal Component Analysis in Applied Research," Sankhya, Ser. A, 26, 329-358.

Srivastava, M. S. and Khatri, C. G. (1979). An Introduction to Multivariate Statistics, New York: North Holland.

Timm, N. H. (1975). Multivariate Analysis with Applications in Education and Psychology, Monterey, California: Brooks/Cole.

Table 1.

Selected Variables for FIS Data (4 Vars)

Variables Selected	Determinant ($X10^3$)		Percentage of Variation Explained	
	Correlation	Covariance	Correlation	Covariance
1	1000.	266.	53.6	69.0
2	1000.	98.	50.8	41.4
3	1000.	221.	62.5	68.5
4	1000.	39.	59.0	47.8
1 2	723.	19.0	77.5	82.9
1 3	431.	25.4	74.1	87.3
1 4	701.	7.3	81.2	83.6
2 3	686.	14.9	81.9	80.3
2 4	559.	2.2	74.2	58.7
3 4	381.	3.3	75.5	73.1
1 2 3	285.	1.65	92.7	98.2
1 2 4	366.	.38	94.3	91.9
1 3 4	162.	.37	87.1	91.9
2 3 4	212.	.18	90.1	83.2
1 2 3 4	84.0	.019	100.	100.

Principal Components Cumulative Percentage of Variation Explained: Correlation 1 - 73.2, 2 - 86.8, 3 - 96.7; Covariance 1 - 78.1, 2 - 89.7, 3 - 98.4.

Table 2.

Selected Variables for J1 Data (13 Vars)

Subset Size	Variables Selected	Determinant($X10^3$)	Percentage of Variation Explained
1	1	1000	26.0
	2	1000	25.9
	7	1000	25.7
	10	1000	23.6
2	5 8	1000	27.8
	11 13	1000	21.0
	4 10	1000	41.1
	4 13	1000	29.3
3	6 11 12	997	38.7
	2 11 12	992	46.1
	1 11 12	992	46.2
	4 11 13	990	38.1
4	4 9 11 13	961	56.7
	2 5 11 12	950	56.4
	1 5 11 12	948	56.6
	1 9 11 12	945	52.8
5	4 9 11 12 13	883	65.1
	5 9 11 12 13	877	59.0
	1 5 11 12 13	871	66.9
	3 5 9 11 13	871	64.9
6	3 5 9 11 12 13	783	73.0
	4 5 9 11 12 13	762	73.5
	2 3 5 11 12 13	726	79.2
	3 5 8 11 12 13	696	68.8

Principal Components Cumulative Percentage of Variation Explained:

1 - 32.5, 2 - 50.7, 3 - 65.2, 4 - 73.7, 5 - 80.7, 6 - 87.0.

Table 3.

Selected Variables for J2 Data (19 Vars)

Subset Size																			Determinant ($X10^3$)	Percentage of Variation Explained
1												13							1000	70.3
													14						1000	69.2
											12								1000	69.1
			3																1000	68.3
2																17		19	999	65.5
										11						17			999	42.6
							8									17			996	74.4
																17	18		971	67.8
3										11						17		19	817	71.4
								9		11						17			789	76.6
					5					11								19	750	69.6
							8			11						17			750	78.7
4								9		11						17		19	363	82.3
					5			9		11								19	350	81.8
						6				11						17		19	345	84.7
					5			9		11							18		326	81.4
5					5			9		11						17		19	127	85.1
					5			9		11							18	19	115	84.5
						6		9		11						17		19	110	88.2
					5		8	9		11						17			108	88.4
6					5			9	10	11						17		19	33	89.9
					5	6		9		11						17		19	33	90.5
					5		8	9		11						17		19	33	90.3
					5	6		9		11							18	19	31	90.4

Principal Components Cumulative Percentage of Variation Explained:

1 - 73.0, 2 - 85.4, 3 - 89.4, 4 - 92.0.

Table 4.

Selected Variables for AH Data (18 Vars)

Subset Size													Determinant ($X10^3$)	Percentage of Variation Explained
1						7							1000	70.5
				5									1000	70.5
							8						1000	69.4
					6								1000	68.4
2	1				6								999	76.7
	1											18	999	74.9
	1									16			999	73.8
							10		13				999	56.2
3	1									16	17		964	90.6
	1								14		17		939	88.6
	1	2									17		925	91.1
	1		4								17		921	88.6
4	1								13	14	17		562	95.3
	1						10		13	14			552	92.9
	1				6				13		17		492	94.6
	1		4				10		13				489	91.4

Principal Components Cumulative Percentage of Variation Explained:

1 - 71.7, 2 - 87.8, 3 - 93.3, 4 - 96.7.

Table 5.

Selected Variables for CHE Data (11 Vars)

Subset Size												Determinant ($X10^3$)	Percentage of Variation Explained
1		2										1000	27.9
	1											1000	26.3
					6							1000	24.5
				5								1000	23.6
2		3					9					999	37.8
		2				7						999	49.3
	1									11		999	38.7
		3						10				999	39.0
3		3					9			11		962	48.1
			4			7	8					958	51.0
						7	8			11		957	44.0
				5			9			11		953	53.0
4		3					8	9		11		895	57.8
		3				7	8			11		890	59.9
			4			7	8			11		886	60.3
	1						8	9		11		843	64.8
5		3				7	8	9		11		760	73.8
			4			7	8	9		11		715	72.1
	1	3					8	9		11		688	76.2
	1		4				8	9		11		676	75.4
6		3	4			7	8	9		11		483	80.2
			4		6	7	8	9		11		463	81.8
		3			6	7	8	9		11		458	82.0
	1	3	4				8	9		11		430	82.3

Principal Components Cumulative Percentage of Variation Explained:

1 - 33.3, 2 - 59.1, 3 - 68.7, 4 - 77.6, 5 - 84.5, 6 - 89.0.

Table 6.

Selected Variables for ORH Data (9 Vars)

Subset Size										Determinant ($X10^3$)	Percentage of Variation Explained
1	1									1000	36.6
		2								1000	35.5
					5					1000	35.3
			3							1000	25.2
2				4			7			999	36.6
	1			4						999	52.8
					5			8		999	53.6
	1							8		999	55.0
3			3	4					9	933	55.4
				4		6			9	924	54.2
				4			7		9	921	50.3
		2		4			7			916	65.5
4		2		4			7		9	803	77.3
	1			4			7		9	774	77.5
				4	5	6			9	762	78.1
		2		4		6			9	735	78.2
5		2		4		6	7		9	567	87.5
	1			4		6	7		9	557	88.1
				4	5	6	7		9	546	86.6
	1		3	4			7		9	541	87.9
6	1		3	4		6	7		9	265	93.5
		2	3	4		6	7		9	263	92.8
			3	4	5	6	7		9	259	92.0
		2		4		6	7	8	9	228	92.1

Principal Components Cumulative Percentage of Variation Explained:

1 - 42.3, 2 - 62.8, 3 - 76.1, 4 - 85.8, 5 - 92.5, 6 - 96.4.

Table 7.

Selected Variables for CS Data (11 Vars)

Subset Size												Determinant ($X10^3$)	Percentage of Variation Explained
1			3									1000	21.2
									10			1000	20.2
							8					1000	19.2
					6							1000	18.7
2						7			10			999	30.9
				5				9				999	19.9
		2		4								999	21.9
			3	4								999	30.7
3						7		9	10			998	40.2
			3			7		9				996	41.2
				4	5			9				996	29.3
			3	4				9				994	40.0
4			3	4	5			9				988	50.1
			3	4		7		9				987	50.5
			3		5	7		9				981	51.0
				4	5			9		11		981	41.7
5			3	4	5	7		9				972	60.3
		2		4	5			9		11		962	53.4
		2		4	5	7		9				955	51.0
				4	5	7		9	10			949	58.9
6		2	3	4	5	7		9				904	69.9
		2		4	5	6	7	9				902	67.2
		2		4	5		8	9		11		878	68.5
		2		4	5	6		9		11		876	68.2

Principal Components Cumulative Percentage of Variation Explained:

1 - 28.0, 2 - 43.0, 3 - 53.2, 4 - 62.5, 5 - 70.9, 6 - 79.0.

Table 8.

Correlation Matrix for CHE Data

	2	3	4	5	6	7	8	9	10	11
1.	.221	.368	.268	.701	.289	.891	.256	-.238	-.378	-.035
2.		.490	.484	.537	.825	.020	-.230	.510	.439	.270
3.			.550	.430	.484	.164	.068	.010	.036	.188
4.				.423	.465	.082	.087	.184	.165	.236
5.					.444	.405	.211	.205	.039	.054
6.						.131	-.135	.292	.268	.301
7.							.174	-.328	-.456	-.077
8.								-.235	-.248	-.095
9.									.594	.053
10.										.292

Table 9.

Correlation Matrix for ORH Data

		SI	S	CA	TI	FE	SE	SR	BA
1.	AL	.961	.419	-.010	.926	.373	.328	.030	.304
2.	SI		.454	-.071	.879	.370	.280	-.032	.269
3.	S			-.058	.425	.657	.465	.061	.225
4.	CA				-.050	.195	.005	.629	.103
5.	TI					.336	.416	.024	.272
6.	FE						.424	.093	.185
7.	SE							.113	.261
8.	SR								.489
9.	BA								

Table 10.

Correlation Matrix for CS Data

	SATV	HSR	HSS	HSMAS	HSMAG	HSSCS	HSSCG	HSENS	HSENG	SEX
1. SATM	.376	.174	.048	.268	.352	.216	.179	-.036	.073	-.183
2. SATV		.217	-.010	.119	.138	.097	.243	.070	.279	.026
3. HSR			.015	.078	.582	.032	.599	-.039	.682	.287
4. HSS				-.018	-.033	-.069	-.066	.063	-.114	.089
5. HSMAS					.187	.095	.110	-.009	.132	-.023
6. HSMAG						.094	.531	.047	.452	.194
7. HSSCS							.150	.035	-.006	-.266
8. HSSCG								.063	.564	.122
9. HSENS									.028	.084
10. HSENG										.354
11. SEX										

*Variable names: SATM-SAT Math, SATV-SAT Verbal, HSR-high school rank, HSS-high school size, HSMAS-high school math semesters, HSMAG-high school math grades, HSSCS-high school science semesters, HSSCG-high school science grades, HSENS-high school English semesters, HSENG-high school English grades, SEX-sex (0=male, 1=female)