TECHNOMETRICS ©, VOL. 26, NO. 2, MAY 1984

Principal Variables

George P. McCabe
Department of Statistics
Purdue University
West Lafayette, IN 47907

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, however, principal components suffer from the disadvantage that each component is a linear combination of all of the original variables. Thus interpretation of the results and possible subsequent data collection and analysis still involve all of the variables. An alternative approach is to select a subset of variables that contain, in some sense, as much information as possible. Methods for selecting such "principal variables" 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 multivariate data. Unlike its poor relative, factor analysis, it enjoys a solid theoretical foundation and possesses many optimal properties. (See Schönemann and Steiger 1978 and Schönemann 1983 for some insight into the limitations of factor analysis.) In a series of papers, Jackson (1980,1981a,b) discusses principal components and factor analysis from an applied point of view and illustrates these ideas with examples from quality control, audiometrics, and the testing of ballistic missiles.

Rotational methods that transform the linear combinations produced by principal components or factor analysis into more interpretable linear combinations are available in most statistical software packages. The popularity of these methods is a recognition of the fact that the original linear combinations are often difficult to interpret. Thus, although principal components solve a well-defined mathematical problem, they frequently fail to provide the statistical consumer with useful results.

Principal components suffer from another 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 the original 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 that are to be considered or measured in the future. For example, suppose that the information contained in 10 variables can be effectively characterized by two principal components, each of which is a linear combination of all 10 variables. An alternative might be to use three or four of the original variables for this purpose.

Procedures for discarding or selecting variables in multivariate analysis are available in a variety of settings. The regression area has been investigated extensively (see Hocking 1976 and 1983 for an overview). Beale, Kendall, and Mann (1967) propose an approach to discarding variables called interdependence analysis. They suggest maximizing the minimum multiple correlation between the selected variables and each discarded variable. Jolliffe (1972,1973) investigates several methods for discarding variables from a principal components analysis.

An example from forestry illustrates the fundamental problem. There are many measured variables that 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, and so on. Some of these variables are easy to measure—for example, girth at three feet; some are more difficult, perhaps requiring a team of graduate students, such as girth 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 fewer variables that, in their judgment, contain the basic information about tree sizes. Of course, with any such selection, essential information can be lost. However, this is also the case with discarded principal components.

In view of the above considerations it is natural to ask whether the desirable characteristics of principal components can be obtained while simultaneously reducing the number of variables to be considered. This article is an attempt to address this question.

In Section 2, assumptions and notation are given. The main results are presented in Section 3 and illustrated with two examples in Section 4. Optimality criteria for principal components and principal variables are discussed in Sections 5 and 6, respectively. Some conclusions are summarized in Section 7.

2. PRELIMINARIES

Let X be a p-dimensional normally distributed random vector with mean zero and known positive definite covariance matrix Σ. Often 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

$$Y = A_k' X, \tag{2.1}$$

where A_k is a $p \times k$ 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

$$A_k' A_k = I_k, \tag{2.2}$$

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

$$\Sigma_Y = A_k' \Sigma A_k. \tag{2.3}$$

Suppose that we observe the random variable Y and want to estimate 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:

$$Z = (\Sigma A_k)(A_k' \Sigma A_k)^{-1} Y. \tag{2.4}$$

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

$$\Sigma_Z = \Sigma A_k (A_k' \Sigma A_k)^{-1} A_k' \Sigma. \tag{2.5}$$

Note that Σ_Z is a $p \times p$ singular matrix, since $k < p$. The transformations X to Y to Z involve a loss of information. The difference between Σ_X and Σ_Z is a measure of the loss resulting from this process.

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

$$\Sigma = \lambda_1 g_1 g_1' + \lambda_2 g_2 g_2' + \cdots + \lambda_p g_p g_p', \tag{2.6}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ 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, familiar properties are $G'G = I$, $G\Lambda G' = \Sigma$, $G'\Sigma G = \Lambda$, $|\Sigma| = \prod_{i=1}^{p} \lambda_i$, and $\text{tr}(\Sigma) = \sum_{i=1}^{p} \lambda_i$.

3. MAIN RESULTS

Dimension-reducing transformations can have a variety of desirable properties. Such properties can be viewed as optimality criteria leading to choices of the matrix A_k. For each of the 10 criteria discussed in Section 5, the optimal solution is principal components, that is, $A_k = G_k$, where G_k consists of the eigenvectors corresponding to the largest k eigenvalues:

$$G_k = (g_1, g_2, \ldots, g_k).$$

The fact that many optimality criteria lead to the principal components solution is one reason for the popularity of this analysis.

Suppose, though, that we constrain the problem so that variables are selected. Then A_k must be of the form

$$A_k = (I_k, 0)'$$

or a matrix obtained by permuting the rows of this matrix. In this matrix, 0 is a $k \times (p - k)$ matrix of zeros. Applying the principal components optimality criteria to the variable selection problem does not lead to a unique solution. The 12 criteria considered in Section 6 lead to the following four solutions:

1. max $|\Sigma_{11}|$, or equivalently, min $|\Sigma_{22 \cdot 1}|$,
2. min tr $(\Sigma_{22 \cdot 1})$,
3. min $\| \Sigma_{22 \cdot 1} \|^2$,
4. max $\sum_{i=1}^{m} \rho_i^2$.

Here, Σ_{11} is the covariance matrix of the selected variables; $\Sigma_{22 \cdot 1}$ is the conditional covariance matrix of the variables not selected given those selected; $|A|$ and tr (A) denote the determinant and trace of the matrix A, respectively; $\| A \|^2$ is the squared norm $(\sum\sum a_{ij}^2)$; and the ρ_i are canonical correlations between the variables not selected and those selected. The matrix $\Sigma_{22 \cdot 1}$ represents the information left in the variables that are not selected, so it is not surprising that the various optimality criteria yield solutions that are functions of the matrix.

Choosing among the criteria depends on the desired optimality property (Sec. 6) and on computational feasibility. The determinantal criterion, (1), 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, the computations take about 110 seconds on a CDC-6500 computer. The cost, at internal rates, is \$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, efficient algorithms are not available for examining all subsets. Of course, evaluation of these criteria for a relatively small collection of subsets is easily accomplished.

A search for near-optimal subsets defined by the trace criterion, (2), can be accomplished by a step-type algorithm. Note that (2) is equivalent to

$$\max \sum_{i=1}^{p} \sigma_{ii}\eta^2(X_i;\, Y),$$

where σ_{ii} denotes the variance of X_i and $\eta^2(X_i,\, Y)$ is the squared multiple correlation between X_i and Y. Note that here Y is simply the vector of selected variables. This form of (2) is due to Okamoto (1969).

Let $X_{(i)}$ denote the variable selected at the ith step. We choose $X_{(1)}$ to be the variable for which

$$\sum_{i=1}^{p} \sigma_{ii}\eta^2(X_i;\, X_{(1)})$$

is maximized. Clearly, $X_{(1)}$ is the best subset of size one. The subsequent steps produce a sequence of subsets, each containing the variables in the preceding set plus the variable that increases the value of the optimality criterion by the largest amount. Thus $X_{(2)}$ is such that

$$\sum_{i=1}^{p} \sigma_{ii}\eta^2(X_i;\, X_{(1)},\, X_{(2)})$$

is maximized. At step j, $X_{(j)}$ is such that

$$\sum_{i=1}^{p} \sigma_{ii}\eta^2(X_i;\, X_{(1)},\, X_{(2)},\, \ldots,\, X_{(j)})$$

is maximized. If a correlation matrix is used, then the σ_{ii} drop out of the above expressions.

Modifications to the above algorithm can be made to increase the chances of finding the optimal subsets. Variables can be deleted or added, several at a time. A "step-down" approach is also feasible.

4. EXAMPLES

To evaluate the procedures described in the previous sections, two sets of real data are analyzed. These analyses are based, as usual, on estimated covariance matrices, not the true matrices. In each case, a computer program that computes and ranks determinants (criterion (1)) for all possible subsets is used first. Then, regressions are run to determine the percentage of variation explained by the various subsets selected in the first step (criterion (2) divided by the trace of Σ). These percentages can be compared to the variation explained by the first few eigenvectors.

The first data set is due to Fisher (1936). It is analyzed in Anderson (1958) and elsewhere. There are four size measurements on 50 samples of Iris versicolor.

Table 1. *Selected Variables for Fisher Data (4 Vars)*

	Determinant (× *10*³)		*Percentage of Variation Explained*	
Variables Selected	*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.	.019	100.	100.

NOTE: 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.

Results are presented in Table 1 for both the observed 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 criteria give different orders.

In correlation form, the matrix has eigenvalues 2.93, .55, .40, and .13. Thus there is one large component that explains 73.2% of the variation. In other words, the average of the squared correlations of the first component with each of the four variables is .732.

From Table 1 it can be seen that if we pick one of the variables instead of the first principal component then the average squared correlation ranges from .508 to .625. We need two variables to explain as much variation as the first principal component. The values range from .741 to .819. It appears from this narrow range of squared multiple correlations that the variables are somewhat interchangeable. However, the determinantal criterion distinguishes the subsets of size two more clearly. The pair (1, 2) appears to be much better than the pair (3, 4). It is interesting to note that the first two variables are sepal length and width while the last two are petal length and width. In a future experiment, one might wish to consider measuring sepals only. Such a decision, of course, must be made by the experimenter, not the statistician.

The second example is from a study of the constituent elements in coal samples by Orheim (1981). Nine elements were measured in 50 samples. The correlation matrix is given in Table 2.

Table 2. *Correlation Matrix for Orheim 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								

The results of the principal variables analysis are presented in Table 3. Note that for each subset size, the subsets are listed in rank order based on the determinantal criterion. The variation explained by the first three principal components (76.1%) can be achieved by any of the subsets of four variables listed. Four components account for 85.8% of the variance. A similar amount can be achieved with five variables.

The first principal component has large weights for variables 1, 2, and 5. In the next two components, the predominant variables are (4, 8) and (3, 6, 7), respectively. It is interesting to note that variables 1, 2, and 5 never appear together in the subsets given in Table 3. On the other hand, most of the subsets contain one of these variables. The same phenomenon occurs with the sets (4, 8) and (3, 6, 7), although to a lesser extent.

Table 3. *Selected Variables for Orheim Data (9 Vars)*

Subset Size	*Determinant* (× *10*^{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

NOTE: 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.

Notice that the determinantal and trace (percentage of variation explained) criteria need not lead to the same optimal subsets for each subset size. The information presented in Table 3 should serve as a guide for the researcher who can use it in combination with other considerations to select a useful subset of variables.

A large number of other data sets have been analyzed using this methodology. The results described above are typical. Generally, the number of variables needed to match or exceed the principal component variation is only one or two more than the number of components.

5. OPTIMAL PROPERTIES OF PRINCIPAL COMPONENTS

In most multivariate texts, principal components are introduced as orthogonal linear combinations having maximum variance subject to constraints. There are many other optimal properties of principal components. A collection of these is discussed below. Proofs of most of these results can be found in Kshirsagar (1972) or Okamoto (1969). The rest follow directly from their results. In Section 6, each of these properties is applied to the variable selection problem, thereby giving sets of principal variables.

Each property involves a maximization or minimization taken over matrices A_k to be used as in (2.1) and subject to the constraint (2.2) for fixed values of k. A solution to each is $A_k = G_k$, that is, 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.

$$\max \operatorname{tr}(\Sigma_Y) = \sum_{i=1}^{k} \lambda_i. \tag{5.1}$$

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

$$\max |\Sigma_Y| = \prod_{i=1}^{k} \lambda_i. \tag{5.2}$$

Since the determinant of a covariance matrix is the generalized variance, this criterion is similar to the first. Here, however, 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 $j = 1, 2$.

$$\max E(Y_1 - Y_2)'(Y_1 - Y_2) = 2 \sum_{i=1}^{k} \lambda_i. \qquad (5.3)$$

$$\max | E(Y_1 - Y_2)(Y_1 - Y_2)' | = 2 \prod_{i=1}^{k} \lambda_i. \qquad (5.4)$$

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 separately for each component as in (5.3), or in the more general multivariate sense as in (5.4).

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

$$\min \| \Sigma_X - \Sigma_Z \|^2 = \sum_{i=k+1}^{p} \lambda_i^2. \qquad (5.5)$$

Recall that Z is the estimated value of X given Y. Therefore, the difference between Σ_X and Σ_Z represents the information lost by the transformation.

$$\min \operatorname{tr} (\Sigma_{X|Y}) = \sum_{i=k+1}^{p} \lambda_i. \qquad (5.6)$$

$$\min \| \Sigma_{X|Y} \|^2 = \sum_{i=k+1}^{p} \lambda_i^2. \qquad (5.7)$$

By considering the conditional covariance matrix of X given Y, Properties 6 and 7 focus upon the variation not extracted by the transformation. This is done by summing over the original variables as in (5.6) or in the norm sense as in (5.7). Note that the matrix $\Sigma_{X|Y}$ is singular.

$$\max \sum_{j=1}^{p} \sigma_{jj} \eta^2(X_j, Y) = \sum_{i=1}^{k} \lambda_i. \qquad (5.8)$$

In this criterion, which is due to Okamoto (1969), σ_{jj} is the (j, j)th element of Σ and $\eta^2(X_j, Y)$ is the squared multiple correlation of X_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.

$$\min E(Z - X)'(Z - X) = \sum_{i=k+1}^{p} \lambda_i. \qquad (5.9)$$

$$\min \| E(Z - X)(Z - X)' \|^2 = \sum_{i=k+1}^{p} \lambda_i^2. \quad (5.10)$$

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 (5.9) and in the norm sense (5.10).

All of the above criteria give extrema that 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. But it is hoped that they are more useful in applications where the objective is variable reduction, not dimension reduction.

6. PRINCIPAL VARIABLES

Recall that to select principal variables, we start with the framework of Section 2 and add the condition that A_k is of the form,

$$A_k = (I_k, 0)',$$

or a matrix obtained by permuting the rows of this matrix. Instead of permuting the rows of A_K, however, it is more convenient to permute 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 corresponding 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. Thus, selection of a set of k variables is 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.

$$\max \operatorname{tr} (\Sigma_Y) = \max \operatorname{tr} \Sigma_{11}. \qquad (6.1)$$

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

$$\operatorname{tr} \Sigma_Y = \sum_{i=1}^{k} \sigma_{ii},$$

where $\sigma_{11} \geq \sigma_{22} \geq \cdots \geq \sigma_{pp}$ are the ordered variances. Thus X_1 consists of the variables with the k largest variances. Although this first criterion 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.

$$\max |\Sigma_Y| = \max |\Sigma_{11}|. \qquad (6.2)$$

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

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

be the conditional covariance of X_2 given X_1. Then,

$$|\Sigma| = |\Sigma_{11}| \cdot |\Sigma_{22 \cdot 1}|.$$

Hence maximizing $|\Sigma_{11}|$ is equivalent to minimizing $|\Sigma_{22 \cdot 1}|$. In other words, maximizing the retained variation, represented by $|\Sigma_{11}|$, is equivalent to minimizing the lost variation, represented by $|\Sigma_{22 \cdot 1}|$.

$$\max E(Y_1 - Y_2)'(Y_1 - Y_2) = 2 \max \operatorname{tr} \Sigma_{11}. \qquad (6.3)$$

$$\max |E(Y_1 - Y_2)(Y_2 - Y_1)'| = 2 \max |\Sigma_{11}|. \qquad (6.4)$$

Since the covariance matrix of $X_1 - X_2$ is simply $2\Sigma_{11}$, these results are equivalent to (6.1) and (6.2), respectively.

$$\min \| \Sigma_X - \Sigma_Z \|^2 = \min \sum_{i=1}^{p-k} \theta_i^2, \qquad (6.5)$$

where $\theta_1, \ldots, \theta_{p-k}$ are the eigenvalues of $\Sigma_{22 \cdot 1}$. To see (6.5) we first observe that

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

Therefore,

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

and

$$\| \Sigma_X - \Sigma_Z \|^2 = \| \Sigma_{22 \cdot 1} \|^2.$$

But,

$$\| \Sigma_{22 \cdot 1} \|^2 = \operatorname{tr}(\Sigma_{22 \cdot 1}\Sigma_{22 \cdot 1}).$$

Now let

$$\Sigma_{22 \cdot 1} = Q\Theta Q',$$

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

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

Note that with the above notation, criteria (6.2) and (6.4) are equivalent to

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

$$\min \operatorname{tr} \Sigma_{X|Y} = \min \sum_{i=1}^{p-k} \theta_i. \qquad (6.6)$$

$$\min \| \Sigma_{X|Y} \|^2 = \min \sum_{i=1}^{p-k} \theta_i^2. \qquad (6.7)$$

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 (6.6) and (6.7) are consequences of this fact.

$$\max \sum_{j=1}^{p} \sigma_{jj}\eta^2(x_j, Y) = \sum_{j=1}^{p} \lambda_j - \min \sum_{i=1}^{k-p} \theta_i. \qquad (6.8)$$

Since the λ_j are fixed, this criterion is equivalent to (6.6). Since $Y = X_1$,

$$\begin{aligned} \eta^2(x_i, Y) &= 1 && \text{for } i = 1, \ldots, k \\ &= 1 - \frac{\sigma_{ii \cdot Y}}{\sigma_{ii}} && \text{for } i = k + 1, \ldots, p, \end{aligned}$$

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

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

The result immediately follows.

$$\min E(Z - X)'(Z - X) = \min \sum_{i=1}^{p-k} \theta_i, \qquad (6.9)$$

$$\min \| E(Z - X)(Z - X)' \|^2 = \min \sum_{i=1}^{p-k} \theta_i^2. \qquad (6.10)$$

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 (6.9) and (6.10) are immediate.

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,

$$\begin{aligned} &\min E(Z_2 - X_2)'\Sigma_{22}^{-1}(Z_2 - X_2) \\ &\qquad = (p - k) - \max \sum_{i=1}^{m} \rho_i^2. \qquad (6.11) \end{aligned}$$

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

To prove (6.11) we first note that

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

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 (6.11) follows immediately.

Criterion (6.11) therefore suggests that variables should be selected so as to maximize the squared canonical correlations between the retained and discarded variables.

The final criterion is suggested by the previous one. 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 (2.1) and (2.4).

$$\max E(Z_1 - Z_2)'(Z_1 - Z_2) = 2\left[\sum_{i=1}^{p}\lambda_i - \min\sum_{i=1}^{p-k}\theta_i\right]. \tag{6.12}$$

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

$$\begin{aligned} E(Z_1 - Z_2)'(Z_1 - Z_2) &= 2 \operatorname{tr} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix} \\ &= 2[\operatorname{tr} \Sigma_{11} + \operatorname{tr}(\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})] \\ &= 2[\operatorname{tr} \Sigma - \operatorname{tr} \Sigma_{22\cdot 1}]. \end{aligned}$$

Equation (6.12) follows immediately from the above equality. Note that this criterion is equivalent to minimizing the sum of the eigenvalues of $\Sigma_{22\cdot 1}$.

Principal components give the optimal solution for the 10 criteria given in Section 5. When these criteria and the two additional ones of this section 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:

$$\min |\Sigma_{22\cdot 1}| = \min \prod_{i=1}^{p-k} \theta_i, \tag{6.13}$$

$$\min \operatorname{tr}(\Sigma_{22\cdot 1}) = \min \sum_{i=1}^{p-k} \theta_i, \tag{6.14}$$

$$\min \|\Sigma_{22\cdot 1}\|^2 = \min \sum_{i=1}^{p-k} \theta_i^2, \tag{6.15}$$

or

$$\max \sum_{i=1}^{m} \rho_i^2. \tag{6.16}$$

Note that

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

Choices among the four criteria may depend on the particular application. If retention of variation is the primary goal, then (6.13) seems to be a natural choice. On the other hand, if predicting the discarded variables is important, then (6.14) is preferable. Computational difficulty is another aspect that 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

$$P = \left(1 - \frac{\sum_{i=1}^{k-p}\theta_i}{\sum_{i=1}^{p}\lambda_i}\right)100\%, \tag{6.17}$$

where P is the percentage of variation explained. Equivalently, (6.17) can be rewritten as

$$P = \frac{\sum_{j=1}^{k}\sigma_{jj} + \sum_{j=k+1}^{p}\sigma_{jj}\eta^2(X_j, Y)}{\sum_{j=1}^{p}\sigma_{jj}}\,100\%. \tag{6.18}$$

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

$$P = p^{-1}\left(k + \sum_{j=k+1}^{p}\eta^2(X_j, Y)\right)100\%. \tag{6.19}$$

7. CONCLUSIONS

In the examples of Section 4 and others examined by the author, the principal variables method gives results comparable to principal components analysis. Since principal components explain the maximum variation percentage for any given number of dimensions, 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 with 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 determinantal 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.

All of the derivations presented in this article are based on starting with a known covariance matrix. In practice, however, only a sample estimate is usually available. The distribution theory associated with principal variables is complex, as is the case with principal components. Cross-validation or bootstrap methods can be used to investigate the stability of results. Such an approach would most likely reinforce the empirical conclusion that there are generally many near-equivalent subsets.

The final choice of variables should be left to the researcher who knows and understands the variables. Tables, such as those with this article, facilitate this choice and allow the incorporation of other information, such as cost of measurement, into the selection process.

A FORTRAN computer program that performs the calculations described in this article is available from the author.

ACKNOWLEDGMENT

Part of this work was completed while the author was visiting the University of Western Australia in Perth, Western Australia.

[*Received September 1982. Revised October 1983.*]

REFERENCES

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

BEALE, E. M. L., KENDALL, M. G., and MANN, D. W. (1967), "The Discarding of Variables in Multivariate Analysis," *Biometrika*, 54, 357–366.

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

HOCKING, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1–49.

——— (1983). "Developments in Linear Regression Methodology: 1959–1982" (with discussion), *Technometrics*, 25, 219–249.

JACKSON, J. E. (1980), "Principal Components and Factor Analysis: Part I—Principal Components," *Journal of Quality Technology*, 12, 201–213.

——— (1981a), "Principal Components and Factor Analysis: Part II—Additional Topics Related to Principal Components," *Journal of Quality Technology*, 13, 46–58.

——— (1981b), "Principal Components and Factor Analysis: Part III—What Is Factor Analysis?" *Journal of Quality Technology*, 13, 125–130.

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.

——— (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: Marcel Dekker.

McCABE, GEORGE P., JR. (1975), "Computations for Variable Selection in Discriminant Analysis," *Technometrics*, 17, 103–109.

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

ORHEIM, ALV (1981), Personal communication.

SCHÖNEMANN, P. H. (1983), "Do IQ Tests Really Measure Intelligence?" *The Behavioral and Brain Sciences*, 6, 311–315.

SCHÖNEMANN, P. H., and STEIGER, J. H. (1978), "On the Validity of Indeterminate Factor Scores," *Bulletin of the Psychonomic Society*, 12, 287–290.