Prediction of Principal Components
by Variable Subsets

by

George P. McCabe
Purdue University

Technical Report#86-19

Department of Statistics
Purdue University

June 1986
Prediction of Principal Components 
by Variable Subsets 

by 

George P. McCabe 
Purdue University 

ABSTRACT 

For multivariate normal data from a single population, principal component analysis is a useful dimensionality reduction technique. Prediction of principal components by variable subsets is considered and the relationship between this problem and principal variables is established. An application to multivariate quality control is discussed and the results are illustrated with an example. 

Key words: Principal variables, regression, quality control, multivariate prediction.
1. INTRODUCTION

In McCabe (1984) principal variables are introduced as a variable selection alternative to principal components. In Section 2, principal variables are briefly described.

The basic notation and results are given in Section 3 with a theorem explaining the relation between principal variables and the prediction of principal components. In Section 4 shift models are examined. The application of these models to multivariate quality control is described in Section 5.

It is shown that sample cost savings can be achieved by using variable subsets to detect a shift in the principal components in some situations. The ideas are illustrated with an example in Section 6.

2. PRINCIPAL VARIABLES

Let $X$ be a $p$-dimensional normally distributed random vector with known positive definite covariance matrix $\Psi$. Without loss of generality, we assume that the mean is zero. We denote this by

$$X \sim N(0, \Psi).$$

(2.1)

We consider partitioning $X$ into $(X'_1, X'_2)'$ where $X'_1$ is a $t$-dimensional vector of retained variables and $X'_2$ is an $s$-dimensional vector of discarded variables. Note that $p = t + s$ and the elements of the vector $X$ are permuted so that the selected variables are the first $t$ variables.

Let $\Psi$ be partitioned correspondingly, i.e.

$$\Psi = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix}$$

(2.2)
where $\Xi$ is the $t \times t$ covariance matrix of $X_1$, etc. Selection of a subset of variables is equivalent to selection of a partition of $\Xi$. Note that there are $\binom{t}{r}$ choices for given $t$ and $2^p - 1$ choices for all $t = 1, \ldots, p$.

McCabe (1984) gives four criteria for selecting principal variables. In this paper, we focus on the second of these. For given $t$, the principal variables are the components of $X_1$, where $X_1$ is such that the trace of $\Xi_{22.1}$ is minimum. Here, $\Xi_{22.1} = \Xi_{22} - \Xi_{21} \Xi_{11}^{-1} \Xi_{12}$.

3. ESTIMATION OF PRINCIPAL COMPONENTS

We consider estimation of the first $u$ principal components by $X_1$. Let the principal component vector be denoted by $Y$. Then,

$$Y = G'X$$  \hspace{1cm} (3.1)

where the columns of $G$ are the eigenvectors of $\Xi$. Let

$$\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p),$$

where $\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$ are the eigenvalues. We partition $G$ as follows

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}$$  \hspace{1cm} (3.2)

where $G_{11}$ is $t \times u$, $G_{12}$ is $t \times v$, $G_{21}$ is $s \times u$ and $G_{22}$ is $s \times v$. Here, $u + v = t + s = p$. Thus,

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} G_{11}'X_1 + G_{21}'X_2 \\ G_{12}'X_1 + G_{22}'X_2 \end{pmatrix}.$$  \hspace{1cm} (3.3)

Let

$$\Gamma = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix},$$  \hspace{1cm} (3.4)

where $\Lambda_1 = \text{diag}(\lambda_1, \ldots, \lambda_u)$ and $\Lambda_2 = \text{diag}(\lambda_{u+1}, \ldots, \lambda_p)$.

To study the estimation of $Y_1$ from $X_1$, we need the conditional distribution of $Y_1$ given $X_1$. In Appendix A it is demonstrated that

$$Y_1|X_1 \sim N(\Lambda_1 G_{11}' \Xi_{11}^{-1} X_1; \Lambda_1 - \Lambda_1 G_{11}' \Xi_{11}^{-1} G_{11} \Lambda_1).$$  \hspace{1cm} (3.5)
It will be useful in what follows to consider the normalized version of $Y_1$ which we denote by $Z_1$. We let

$$Z_1 = \Lambda_1^{-\frac{1}{2}} Y_1.$$ \hfill (3.6)

It follows from (3.5) that

$$Z_1 | X_1 \sim N(\Lambda_1^{\frac{1}{2}} G'_{11} \Sigma_{11}^{-1} X_1; I - \Lambda_1^{\frac{1}{2}} G'_{11} \Sigma_{11}^{-1} G_{11} \Lambda_1^{\frac{1}{2}}).$$ \hfill (3.7)

Let

$$R^2 = \Lambda_1^{\frac{1}{2}} G'_{11} \Sigma_{11}^{-1} G_{11} \Lambda_1^{\frac{1}{2}}.$$ \hfill (3.8)

Since $I - R^2$ is the covariance matrix of the standardized variable $Z_1$ given $X_1$, this matrix is a multivariate analog of the squared multiple correlation coefficient. The diagonal elements of $R^2$ are the squared multiple correlation coefficients of the elements of $Y_1$ with $X_1$. Note that $R^2$ is singular if $t < u$.

The following theorem establishes the relationship between principal variables and the prediction of principal components by subsets of variables. In this theorem, $u = p, \Lambda_1 = \Lambda, G = (G'_1, G'_2)'$, and $R^2$ is a $p \times p$ matrix. The expression $\text{tr}(M)$ denotes the trace of the matrix $M$.

**Theorem.** Let

$$R^2 = \Lambda_1^{\frac{1}{2}} G'_1 \Sigma_{11}^{-1} G_1 \Lambda_1^{\frac{1}{2}},$$

where

$$G_1 = (G_{11}, G_{12}).$$

Then,

$$\text{tr}(\Lambda_1^{\frac{1}{2}} R^2 \Lambda_1^{\frac{1}{2}}) = \text{tr} \Sigma - \text{tr} \Sigma_{22.1}.$$ \hfill (3.9)
The proof is given in Appendix B.

Recall that the principal variable criterion mentioned in the previous section is the trace of $\Psi_{22.1}$. Since $tr(\Psi)$ is fixed, optimization corresponds to maximizing $tr(\Lambda^{\frac{1}{2}} R^2 \Lambda^{\frac{1}{2}})$.

If we normalize the quantity $tr(\Psi - tr(\Psi_{22.1}) by tr(\Psi)$, we obtain the proportion of variation in $X$ explained by $X_1$. Let

$$\lambda^*_i = \lambda_i / \Sigma \lambda_i,$$  \hspace{1cm} (3.10)

and let $R_i^2$ denote the squared multiple correlation coefficient between the $i$-th element of $Y$ and $X_1$. We then have the following corollary.

COROLLARY. The proportion of variation in $Y$ explained by $X_1$ is

$$\sum_{i=1} \lambda^*_i R_i^2.$$ \hspace{1cm} (3.11)

Thus, the principal variables maximize the weighted average of the $R^2$'s for predicting the principal components with weights proportional to the eigenvalues.

4. SHIFT MODELS

We consider models in which the covariance structure of the problem remains as above but there is a shift in the mean. Given that the shift occurs in $Y_1$, we investigate the suitability of $X_1$ for detecting the shift.

Specifically, we assume $EZ$ has changed from zero to $\Delta$ where $\Delta = (\Delta_1, 0)'$ and $\Delta_1$ is $(u \times 1)$. It follows that

$$EY_1 = \Lambda_1^{\frac{1}{2}} \Delta_1$$ \hspace{1cm} (4.1)
and
\[ EY_2 = 0. \] (4.2)

Since \( X = GY \), we have
\[ EX = \begin{pmatrix} G_{11}A_1^{1/2} \Delta_1 \\ G_{21}A_1^{1/2} \Delta_1 \end{pmatrix} \] (4.3)

Let \( \hat{Z}_1 \) denote the conditional expectation of \( Z \), given \( X_1 \). From (3.7) it follows that
\[ \hat{Z}_1 = A_1^{1/2} G_{11}^{1/2} \Sigma_{11}^{-1} X_1. \] (4.4)

Under the shift model,
\[ X_1 \sim N(G_{11}A_1^{1/2} \Delta_1, \Sigma_{11}) \] (4.5)

and therefore,
\[ \hat{Z}_1 \sim N(\lambda_1^{1/2} G_{11}^{1/2} \Sigma_{11}^{-1} G_{11} \lambda_1^{1/2} \Delta_1, \lambda_1^{1/2} G_{11}^{1/2} \Sigma_{11}^{-1} G_{11} \lambda_1^{1/2} \Delta_1). \] (4.6)

From the definition of \( R^2 \) in (3.8) we see that
\[ \hat{Z}_1 \sim N(R^2 \Delta_1, R^2). \]

To compare the efficiency of using \( X_1 \) (through \( \hat{Z}_1 \)) versus \( Z_1 \) (equivalently, \( Y_1 \)) to detect the shift \( \Delta \), we consider the noncentrality parameters for the distributions of \( Z_1'Z_1 \) and \( \hat{Z}_1' (R^2)^{-1} \hat{Z}_1 \). Note that if \( R^2 \) is nonsingular we use any generalized inverse in the quadratic form for \( \hat{Z}_1 \).

Let \( EFF \) denote the ratio of the noncentrality parameters corresponding to \( \hat{Z}_1 \) and \( Z_1 \), respectively. Then, it is easy to show that
\[ EFF = \frac{\Delta_1' R^2 \Delta_1}{\Delta_1' \Delta_1}. \]
Some special cases are worthy of note. If \( \Delta_1 = k(1,1,\ldots,1)' \), corresponding to a shift of \( k \) standard deviations in each of the first \( u \) principal components, then
\[
EFF = \frac{\sum_{i=1}^{u} \sum_{j=1}^{u} r_{ij}}{u}
\]
where \( R^2 = (r_{ij}) \). If \( u = 1 \) and \( \Delta_1 = (k) \), then
\[
EFF = R_1^2,
\]
the squared multiple correlation of the first principal component with \( X_1 \). In this case, regression subset algorithms can be used to find the vector \( X_1 \) which maximizes \( R_1^2 \) and thereby maximizes the efficiency as long as \( p \) is not too large.

5. APPLICATION TO MULTIVARIATE QUALITY CONTROL

Suppose \( X \) is measured for a process and it is believed that if the process goes out of control that there will be a shift in the first principal component. If a subset \( X_1 \) has been chosen for monitoring purposes, then some sample size comparisons can be made using the value of \( R_1^2 \). Specifically, \( N \) observations on \( X \) give the same information as \( N/R_1^2 \) observations on \( X_1 \).

If all observations were equally costly then \( N \) measurements on \( X \) would require \( Np \) units while \( N/R_1^2 \) observations on \( X_1 \) would require \( Nt/R_1^2 \) units. In many applications, there is overhead in obtaining the sample to measure so that not all costs would be equal. If some of the components of \( X \) are expensive, however, it may be possible to find a relatively inexpensive \( X_1 \) with an adequate but suboptimal (given \( t \)) \( R_1^2 \).

The full analysis of a given problem would require the complete cost structure. However, as long as data is not free, the above analysis
suggests that savings can be made by considering subsets of variables as predictors of principal components.

6. EXAMPLE

To illustrate the results described in the previous sections the Fisher Iris data are analyzed. As in McCabe (1984), we use the 50 samples on Iris versicolor. There are four size measurements on each sample. The covariance matrix rather than the correlation matrix is analyzed.

Table 1 gives $R^2$ values for predicting the principal components for all possible subsets. The last column gives the value of the principal variables criterion. Observe that this value is obtained for each row by summing the products of the correlations and the normed eigenvalues ($\lambda_i^*$).

For each subset size the principal variables are optimal for predicting the first principal component. This observation is a consequence of the dominance of the first principal component ($\lambda_1^* = .781$).

The efficiency of the first variable relative to $Y_1$ for detecting a shift in $Y_1$ is .864. Suppose the cost of measuring this variable is $c$. Then, since $Y_1$ requires measurement of $X = (X_1, X_2, X_3, X_4)$; $X_1$ would be preferred whenever the cost of measuring $X$ is greater than $c / .864 = 1.157c$. 
Table 1.

Values for $R^2$ for Predicting Principal Components from Variable Subsets

<table>
<thead>
<tr>
<th>Subsets</th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_3$</th>
<th>$Y_4$</th>
<th>Proportion of Variation Explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.864</td>
<td>.122</td>
<td>.014</td>
<td>.000</td>
<td>.690</td>
</tr>
<tr>
<td>2</td>
<td>.462</td>
<td>.237</td>
<td>.296</td>
<td>.005</td>
<td>.414</td>
</tr>
<tr>
<td>3</td>
<td>.859</td>
<td>.039</td>
<td>.098</td>
<td>.004</td>
<td>.685</td>
</tr>
<tr>
<td>1 2</td>
<td>.914</td>
<td>.742</td>
<td>.334</td>
<td>.010</td>
<td>.829</td>
</tr>
<tr>
<td>1 3</td>
<td>.982</td>
<td>.611</td>
<td>.391</td>
<td>.016</td>
<td>.873</td>
</tr>
<tr>
<td>1 4</td>
<td>.954</td>
<td>.718</td>
<td>.043</td>
<td>.285</td>
<td>.836</td>
</tr>
<tr>
<td>2 3</td>
<td>.897</td>
<td>.245</td>
<td>.852</td>
<td>.006</td>
<td>.803</td>
</tr>
<tr>
<td>2 4</td>
<td>.632</td>
<td>.269</td>
<td>.637</td>
<td>.463</td>
<td>.587</td>
</tr>
<tr>
<td>3 4</td>
<td>.862</td>
<td>.277</td>
<td>.174</td>
<td>.687</td>
<td>.731</td>
</tr>
<tr>
<td>1 2 3</td>
<td>.999</td>
<td>.982</td>
<td>.999</td>
<td>.020</td>
<td>.982</td>
</tr>
<tr>
<td>1 2 4</td>
<td>.960</td>
<td>.918</td>
<td>.637</td>
<td>.485</td>
<td>.919</td>
</tr>
<tr>
<td>1 3 4</td>
<td>.990</td>
<td>.774</td>
<td>.507</td>
<td>.728</td>
<td>.919</td>
</tr>
<tr>
<td>2 3 4</td>
<td>.898</td>
<td>.349</td>
<td>.865</td>
<td>.888</td>
<td>.832</td>
</tr>
<tr>
<td>1 2 3 4</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>.781</td>
<td>.116</td>
<td>.087</td>
<td>.016</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX A

Proof that

\[ Y_1|X_1 \sim N(\Lambda_1 G'_{11} \Sigma^{-1}_{11} X_1; \Lambda_1 - \Lambda_1 G'_{11} \Sigma^{-1}_{11} G_{11} \Lambda_1). \]

Let \( A \) and \( B \) denote \( p \times t \) and \( p \times u \) matrices. From the assumption

\[ X \sim N(0, \Sigma), \]

it follows that

\[ B'X|A'X \sim N(B' \Sigma A(A' \Sigma A)^{-1} A'X; B' \Sigma B - B' \Sigma A(A' \Sigma A)^{-1} A' \Sigma B). \]

We let

\[ B = \begin{pmatrix} G_{11} \\ G_{21} \end{pmatrix} \]

and

\[ A = \begin{pmatrix} I \\ 0 \end{pmatrix}, \]

where \( G_{ij} \) is given by (3.2) and \( I \) is the \( t \times t \) identity matrix. Thus, \( A'X = X_1 \) and \( B'X = Y_1 \).

We first note that

\[ B' \Sigma A = (G'_{11} \Sigma_{11} + G'_{21} \Sigma_{21}) \]

\[ = \Lambda_1 G'_{11} \]

where the second equality follows from the relation

\[ G' \Sigma = \Lambda G'. \]

Also note that

\[ G' \Sigma G = \Lambda \]
implies

\[ B' \otimes B = \Lambda_1. \]

Combining the above with the facts that

\[ A' \otimes A = \Psi_{11} \]

and

\[ A'X = X_1 \]

gives the desired result.

**APPENDIX B**

*Proof of the theorem in Section 3.*

From the definition of \( R^2 \) it follows that

\[ \Lambda^\frac{1}{2} R^2 \Lambda^\frac{1}{2} = \Lambda G_1^\prime \Psi^{-1} G_1 \Lambda \]

where

\[ G_1 = (G_{11}, G_{12}). \]

Since

\[ tr \Psi = tr \Lambda, \]

it is sufficient to show that

\[ tr(\Lambda - \Lambda G_1^\prime \Psi_{11}^{-1} G_{11} \Lambda) = tr \Psi_{22.1} \]

First, we note that

\[ \Lambda = G' \otimes G \]

\[ = G_1' \Psi_{11} G_1 + G_2' \Psi_{21} G_1 + G_1' \Psi_{12} G_2' + G_2' \Psi_{22} G. \]
Second, since
\[ G' \Sigma = \Lambda G', \]

it follows that
\[ \Lambda G'_1 = G'_1 \Sigma_{11} + G'_2 \Sigma_{21}. \]

Therefore,
\[ \Lambda G'_1 \Sigma_{11}^{-1} G'_1 \Lambda = (G'_1 \Sigma_{11} + G'_2 \Sigma_{21}) \Sigma_{11}^{-1} (\Sigma_{11} G_1 + \Sigma_{21} G_2) \]
\[ = G'_1 \Sigma_{11} G_1 + G'_2 \Sigma_{21} G_1 + G'_1 \Sigma_{21} G_2 + G'_2 \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} G_2, \]

and
\[ \Lambda - \Lambda G'_{11} \Sigma_{11}^{-1} G_{11} \Lambda = G'_2 (\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}) G_2 \]
\[ = G'_2 \Sigma_{22.1} G_2. \]

The result follows from
\[ trG'_2 \Sigma_{22.1} G_2 = tr \Sigma_{22.1} G_2 G'_2 \]
\[ = tr \Sigma_{22.1} \]

since
\[ G_2 G'_2 = I. \]

REFERENCE