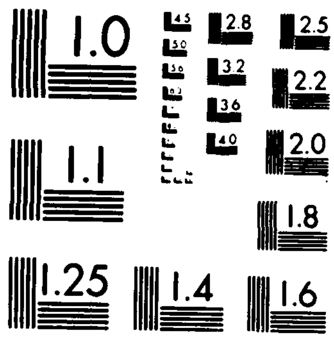
AD-A184 920 ILLUSTRATIVE EXAMPLES OF PRINCIPAL COMPONENT ANALYSIS USING SYSTAT/FACTOR (U) CORNELL UNIV ITHACA NY MATHEMATICAL SCIENCES INST W T FEDERER ET AL 1/1

UNCLASSIFIED 20 MAY 87 TR-87-50 ARO-23306 87-MA F/G 12/3 NL

END
10 87
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

UNCLASSIFIED DTIC FILE COPY MASTER COPY - FOR REPRODUCTION PURPOSES (2)

AD-A184 920

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S): ARO 23306.87-MA
6a. NAME OF PERFORMING ORGANIZATION: Mathematical Sciences Inst. — 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION: U. S. Army Research Office
6c. ADDRESS (City, State, and ZIP Code): 294 Caldwell Hall; Cornell University, Ithaca, New York 14853	7b. ADDRESS (City, State, and ZIP Code): P. O. Box 12211, Research Triangle Park, NC 27709-2211
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: U. S. Army Research Office — 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: DAAG-29-85-C-0018
8c. ADDRESS (City, State, and ZIP Code): P. O. Box 12211, Research Triangle Park, NC 27709-2211	10. SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)

Illustrative Examples of Principal Component Analysis using Systat/Factor* (U)

12. PERSONAL AUTHOR(S)

W.T. Federer, C.E. McCulloch and N.J. Miles-McDermott

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim Technical	FROM ____ TO ____	May 20, 1987	31

16. SUPPLEMENTARY NOTATION

The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	data sets; principal component analysis; linear combinations multivariate statistical analysis; correlation matrix;

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

In order to provide a deeper understanding of the workings of principal components, four data sets were constructed by taking linear combinations of values of two uncorrelated variables to form the X-variates for the principal component analysis. The examples highlight some of the properties and limitations of principal component analysis.

This is part of a continuing project that produces annotated computer output for principal component analysis. The complete project will involve processing four examples on SAS/PRINCOMP, BMDP/4M, SPSS-X/ FACTOR, GENSTAT/PCP, and SYSTAT/FACTOR. We show here the results from SYSTAT/FACTOR, Version 3.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: ☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION: Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code) / 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE: UNCLASSIFIED

18. (con't.)

coefficients; variance-covariance; control language

ARO 23306.87-MA

CORNELL UNIVERSITY

MATHEMATICAL SCIENCES INSTITUTE

TECHNICAL REPORT '87-50

*ILLUSTRATIVE EXAMPLES OF PRINCIPAL COMPONENT ANALYSIS USING SYSTAT/FACTOR**

BY

W.T. Federer, C.E. McCulloch and N.J. Miles-McDermott

MAY 1987

294 Caldwell Hall ■ Ithaca, New York 14853-2602

(607) 255-8005

ILLUSTRATIVE EXAMPLES OF PRINCIPAL COMPONENT ANALYSIS USING SYSTAT/FACTOR*

W. T. Federer, C. E. McCulloch and N. J. Miles-McDermott

BU-901-M November 1986

ABSTRACT

In order to provide a deeper understanding of the workings of principal components, four data sets were constructed by taking linear combinations of values of two uncorrelated variables to form the X-variates for the principal component analysis. The examples highlight some of the properties and limitations of principal component analysis.

This is part of a continuing project that produces annotated computer output for principal component analysis. The complete project will involve processing four examples on SAS/PRINCOMP, BMDP/4M, SPSS-X/FACTOR, GENSTAT / PCP, and SYSTAT / FACTOR. We show here the results from SYSTAT/FACTOR, Version 3.

* Supported by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University.

1. INTRODUCTION

Principal components is a form of multivariate statistical analysis and is one method of studying the correlation or covariance structure in a set of measurements on m variables for n observations. For example, a data set may consist of $n = 260$ samples and $m = 15$ different fatty acid variables. It may be advantageous to study the structure of the 15 fatty acid variables since some or all of the variables may be measuring the same response. One simple method of studying the correlation structure is to compute the $m(m-1)/2$ pairwise correlations and note which correlations are close to unity. When a group of variables are all highly inter-correlated, one may be selected for use and the others discarded or the sum of all the variables may be used. When the structure is more complex, the method of principal components analysis (PCA) becomes useful.

In order to use and interpret a principal components analysis, there needs to be some practical meaning associated with the various principal components. In Section 2 we describe the basic features of principal components and in Section 3 we examine some constructed examples using SYSTAT/FACTOR to illustrate the interpretations that are possible. In Section 4 we summarize our results.

2. BASIC FEATURES OF PRINCIPAL COMPONENT ANALYSIS

PCA can be performed on either the variances and covariances among the m variables or their correlations. One should always check which is being used in a particular computer package program. SYSTAT can use either the variances and covariances or the correlations but uses the correlations by default. First we will consider analyses using the matrix of variances and covariances. A PCA generates m new variables, the principal components (PCs), by forming linear combinations of the original variables, $X = (X_1, X_2, \ldots, X_m)$, as follows:

$$PC_1 = b_{11}X_1 + b_{12}X_2 + \ldots + b_{1m}X_m = Xb_1$$

$$PC_2 = b_{21}X_1 + b_{22}X_2 + \ldots + b_{2m}X_m = Xb_2$$

$$\vdots$$

$$PC_m = b_{m1}X_1 + b_{m2}X_2 + \ldots + b_{mm}X_m = Xb_m \quad ,$$

In matrix notation,

$$P = (PC_1, PC_2, \ldots, PC_m) = X\ (b_1, b_2, \ldots, b_m) = XB,$$

and conversely $X = P\ B^{-1}$.

The rationale in the selection of the coefficients, b_{ij}, that define the linear combinations that are the PC_i is to try to capture as much of the variation in the original variables with as few PCs as possible. Since the variance of a linear combination of the Xs can be made arbitrarily large by selecting very large coefficients, the b_{ij} are constrained by convention so that the sum of squares of the coefficients for any PC is unity:

$$\Sigma_{j=1}^{m} b_{ij}^2 = 1 \qquad i = 1,2,\ldots,m \quad .$$

Under this constraint, the b_{1j} in PC_1 are chosen so that PC_1 has maximal variance.

If we denote the variance of X_i by s_i^2 and if we define the total variance, $\Sigma_{i=1}^{m} s_i^2$, as T, then the proportion of the variance in the original variables that is captured in PC_1 can be quantified as $var(PC_1)/T$. In selecting the coefficients for PC_2, they are further constrained by the requirement that PC_2 be uncorrelated with PC_1. Subject to this constraint and the constraint that the squared coefficients sum to one, the coefficients b_{2j} are selected so as to maximize $var(PC_2)$. Further coefficients and PCs are selected in a similar manner, by requiring that a PC be uncorrelated with all PCs previously selected and then selecting the coefficients to maximize variance. In this manner, all the PCs are constructed so that they are uncorrelated and so that the first few PCs capture as much variance as possible. The coefficients also have the following interpretation which helps to relate the PCs back to the original variables. The correlation between the i^{th} PC and the j^{th} variable is

$$b_{ij}\sqrt{var(PC_i)}/s_j \quad .$$

After all m PCs have been constructed, the following identity holds:

$$var(PC_1) + var(PC_2) + \ldots + var(PC_m) = T = \Sigma_{i=1}^{m} s_i^2 \quad .$$

This equation has the interpretation that the PCs divide up the total variance of the Xs completely. It may happen that one or more of the last few PCs have variance zero. In such a case, all the variation in the data can be captured by fewer than m variables. Actually, a much stronger result is also true; the PCs can also be used to reproduce the actual values of the Xs, not just their variance. We will demonstrate this more explicitly later.

The above properties of PCA are related to a matrix analysis of the variance-covariance matrix of the Xs, S_x. Let D be a diagonal matrix with entries being the eigenvalues, λ_i, of S_x arranged in order from largest to smallest. Then the following properties hold:

(i) $\lambda_i = \text{var}(PC_i)$

(ii) $\text{trace}(S_x) = \Sigma_{i=1}^{m} s_i^2 = T = \Sigma_{i=1}^{m} \lambda_i = \Sigma_{i=1}^{m} \text{var}(PC_i)$

(iii) $\text{corr}(PC_i, X_j) = \dfrac{b_{ij}\sqrt{\lambda_i}}{s_j}$

(iv) $S_x = B'DB$.

The statements made above are for the case when the analysis is performed on the variance-covariance matrix of the Xs. The correlation matrix could also be used, which is equivalent to performing a PCA on the variance-covariance matrix of the standardized variables,

$$Y_i = \frac{X_i - \bar{X}_i}{s_i}$$

PCA using the correlation martrix is different in these respects:

(i) The total "variance" is m, the number of variables. (It is not truly variance anymore.)

(ii) The correlation between PC_i and X_j is given by

$b_{ij}\sqrt{var(PC_i)} = b_{ij}\sqrt{\lambda_i} = \Lambda_i$ (called component loading in SYSTAT). Thus PC_i is most highly correlated with the X_j having the largest coefficient in PC_i in absolute value.

The experimenter must choose whether to use standardized (PCA on a correlation matrix) or unstandardized coefficients (PCA on a variance-covariance matrix). The latter is used when the variables are measured on a comparable basis. This usually means that the variables must be in the same units and have roughly comparable variances. If the variables are measured in different units, then the analysis will usually be performed on the standardized scale, otherwise the analysis may only reflect the different scales of measurement. For example, if a number of fatty acid analyses are made, but the variances, s_i^2, and means, $\bar{X}_i$, are obtained on different bases and by different methods, then standardized variables would be used (PCA on the correlation matrix).

To illustrate some of the above ideas, a number of examples have been constructed and these are described in Section 3. In each case two variables, Z_1 and Z_2, which are uncorrelated, are used to construct X_i. Thus, all the variance can be captured with two variables and hence only two of the PCs will have nonzero variances. In matrix analysis terms, only two eigenvalues will be nonzero. An important thing to note is that in general, **PCA will not recover the original variables Z_1 and Z_2.** Both standardized and nonstandardized computations will be made.

3. EXAMPLES

Throughout the examples we will use the variables Z_1 and Z_2 (with n = 11) from which we will construct $X_1, X_2, \ldots, X_m$. We will perform PCA on the Xs. Thus, in our constructed examples, there will only really be two underlying variables.

Values of Z_1 and Z_2

Z_1	-5	-4	-3	-2	-1	0	1	2	3	4	5
Z_2	15	6	-1	-6	-9	-10	-9	-6	-1	6	15

Notice that Z_1 exhibits a linear trend through the 11 samples and Z_2 exhibits a quadratic trend. They are also chosen to have mean zero and be uncorrelated. Z_1 and Z_2 have the following variance-covariance matrix (a variance-covariance matrix has the variance for the i^{th} variable in the i^{th} row and i^{th} column and the covariance between the i^{th} variable and the j^{th} variable in the i^{th} row and j^{th} column).

Variance-covariance matrix of Z_1 and Z_2

$$\begin{bmatrix} 11 & 0 \\ 0 & 85.8 \end{bmatrix}$$

Thus the variance of Z_1 is 11 and the covariance between Z_1 and Z_2 is zero. Also the total variance is 11 + 85.8 = 96.8.

Example 1: In this first example we analyze Z_1 and Z_2 as if they were the data. Thus $X_1 = Z_1$ and $X_2 = Z_2$ and m = 2. If PCA is perfomed on the variance-covariance matrix, then the SYSTAT output is as follows (SYSTAT control language for this example and all subsequent examples is in the appendix and the bold face print was typed on the computer output to explain the calculation performed):

MATRIX TO BE FACTORED = **Covariance Matrix (s_{ij})**

	X1	X2
X1	$s_{11} = 11.000$	
X2	$s_{12} = s_{21} = -0.000$	$s_{22} = 85.800$

LATENT ROOTS (EIGENVALUES) $\lambda_i = s_i^2$

1	2
$\lambda_1 = 85.800$	$\lambda_2 = 11.000$

COMPONENT LOADINGS = $\underset{\sim}{b}_i \sqrt{\lambda_i} = \underset{\sim}{A}_i$

Note: SYSTAT does not print out $\underset{\sim}{b}_i$ (eigenvectors). To obtain eigenvectors, divide the component loadings by $\sqrt{\lambda_i}$.

	$1 = \underset{\sim}{A}_1$	$2 = \underset{\sim}{A}_2$
X1	0.000	3.317
X2	9.263	0.000

i.e. $\underset{\sim}{b}_1' = [0 \quad 9.26]/\sqrt{85.8}$
$= [0 \quad 1]$

VARIANCE EXPLAINED BY COMPONENTS

1	2
85.800	11.000

PERCENT OF TOTAL VARIANCE EXPLAINED = **proportion of variance explained** by PC_i

1	2
88.636	11.364

FACTOR SCORE COEFFICIENTS = $\underset{\sim}{b}_i / \sqrt{\lambda_i} = \underset{\sim}{y}_i$

	$1 = \underset{\sim}{y}_1$	$2 = \underset{\sim}{y}_2$
X1	0.000	0.302
X2	0.108	0.000

		X1	X2	FACTOR(1) $=PC_1$	FACTOR(2) $=PC_2$
CASE	1	-5.000	15.000	15.000	-5.000
CASE	2	-4.000	6.000	6.000	-4.000
CASE	3	-3.000	-1.000	-1.000	-3.000
CASE	4	-2.000	-6.000	-6.000	-2.000
CASE	5	-1.000	-9.000	-9.000	-1.000
CASE	6	0.000	-10.000	-10.000	0.000
CASE	7	1.000	-9.000	-9.000	1.000
CASE	8	2.000	-6.000	-6.000	2.000
CASE	9	3.000	-1.000	-1.000	3.000
CASE	10	4.000	6.000	6.000	4.000
CASE	11	5.000	15.000	15.000	5.000

$$PC_i = (y_{i1}X_1 + y_{i2}X_2)\sqrt{\lambda_i}$$

$$= b_{i1}X_1 + b_{i2}X_2$$

$$PC_1 = 0X_1 + 1X_2$$

for case 1, $PC_1 = 0(-5) + 1(15) = 15$

We can interpret the results as follows:

1) The first principal component is

$$PC_1 = 0 \cdot X_1 + 1 \cdot X_2 = X_2$$

2) $PC_2 = 1 \cdot X_1 + 0 \cdot X_2 = X_1$

3) $Var(PC_1)$ = eigenvalue = 85.8 = $Var(X_2)$

4) $Var(PC_2)$ = eigenvalue = 11.0 = $Var(X_1)$

The PCs will be the same as the Xs whenever the Xs are uncorrelated. Since X_2 has the larger variance, it becomes the first principal component.

If PCA is performed on the correlation matrix, we get slightly different results.

Correlation Matrix of Z_1 and Z_2

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

A correlation matrix always has unities along its diagonal and the correlation between the i^{th} variable and the j^{th} variable in the i^{th} row and j^{th} column. PCA in SYSTAT would yield the following output:

MATRIX TO BE FACTORED = **Correlation Matrix** (r_{ij})

	X1	X2
X1	r_{11} = 1.000	
X2	$r_{12} = r_{21}$ = -0.000	r_{22} = 1.000

LATENT ROOTS (EIGENVALUES) = λ_i

1	2	
λ_1 = 1.000	λ_2 = 1.000	$\Sigma_{i=1}^{m} \lambda_i = m$

COMPONENT LOADINGS = $\underset{\sim}{b}_i \sqrt{\lambda_i} = \underset{\sim}{\Lambda}_i$

	$1 = \underset{\sim}{\Lambda}_1$	$2 = \underset{\sim}{\Lambda}_2$
X1	1.000	0.000
X2	0.000	1.000

$\underset{\sim}{b}_1' = [1 \quad 0] \sqrt{1}$

$= [1 \quad 0]$

VARIANCE EXPLAINED BY COMPONENTS

1	2
1.000	1.000

PERCENT OF TOTAL VARIANCE EXPLAINED = **proportion of variance explained by** PC_i

1	2
50.000	50.000

FACTOR SCORE COEFFICIENTS = $\underset{\sim}{b}_i / \sqrt{\lambda_i} = \underset{\sim}{y}_i$

	$1 = \underset{\sim}{y}_1$	$2 = \underset{\sim}{y}_2$
X1	1.000	0.000
X2	0.000	1.000

	X1	X2	FACTOR(1) $=PC_1$	FACTOR(2) $=PC_2$
CASE 1	-5.000	15.000	-1.508	1.619
CASE 2	-4.000	6.000	-1.206	0.648
CASE 3	-3.000	-1.000	-0.905	-0.108
CASE 4	-2.000	-6.000	-0.603	-0.648
CASE 5	-1.000	-9.000	-0.302	-0.972
CASE 6	0.000	-10.000	0.000	-1.080
CASE 7	1.000	-9.000	0.302	-0.972
CASE 8	2.000	-6.000	0.603	-0.648
CASE 9	3.000	-1.000	0.905	-0.108
CASE 10	4.000	6.000	1.206	0.648
CASE 11	5.000	15.000	1.508	1.619

$$PC_i = y_{i1}X_1/S_1 + y_{i2}X_2/S_2$$

$$= b_{i1}X_1/S_1 + b_{i2}X_2/S_2$$

$$PC_1 = 1X_1/3.32 + 0X_2/9.26$$

for case 1,

$$= -5/3.32$$

$$= -1.508$$

The principal components are again the Xs (standardized Zs) themselves, but the eigenvalues (var(PCs)) are unity since the variables have been standardized first.

Example 2: Let $X_1 = Z_1$, $X_2 = 2Z_1$ and $X_3 = Z_2$. The summary statistics are given below.

	X1	X2	X3
MEAN	0.000000	0.000000	0.000000
ST DEV	3.316625	6.63325	9.262829

If the analysis is performed on the variance-covariance matrix using SYSTAT the results are:

MATRIX TO BE FACTORED = **Covariance Matrix (s_{ij})**

	X1	X2	X3
X1	11.000		
X2	22.000	44.000	
X3	-0.000	-0.000	85.800

Note: SYSTAT does not give covariances above the diagonal

LATENT ROOTS (EIGENVALUES) = λ_i

1	2	3
85.800	55.000	0.000

Note: $\Sigma_{i=1}^{m} s_i^2 = \Sigma_{i=1}^{m} \lambda_i$

COMPONENT LOADINGS = $\underset{\sim}{b}_i \sqrt{\lambda_i} = \underset{\sim}{\Lambda}_i$

	1	2
X1	0.000	3.317
X2	0.000	6.633
X3	9.263	0.000

$\underset{\sim}{b}_2' = [3.317 \quad 6.633 \quad 0]/\sqrt{55}$

$= [.447 \quad .894 \quad 0]$

Note: The 3rd component loadings were 0's and are not printed by SYSTAT.

VARIANCE EXPLAINED BY COMPONENTS

1	2
85.800	55.000

PERCENT OF TOTAL VARIANCE EXPLAINED

1	2
60.938	39.063

FACTOR SCORE COEFFICIENTS = $\underset{\sim}{b}_i \,/\, \sqrt{\lambda_i} = \underset{\sim}{y}_i$

	$1 = \underset{\sim}{y}_1$	$2 = \underset{\sim}{y}_2$
X1	0.000	0.060
X2	0.000	0.121
X3	0.108	0.000

		X1	X2	X3	FACTOR(1) $=PC_1$	FACTOR(2) $=PC_2$
CASE	1	-5.000	-10.000	15.000	15.000	-11.180
CASE	2	-4.000	-8.000	6.000	6.000	-8.944
CASE	3	-3.000	-6.000	-1.000	-1.000	-6.708
CASE	4	-2.000	-4.000	-6.000	-6.000	-4.472
CASE	5	-1.000	-2.000	-9.000	-9.000	-2.236
CASE	6	0.000	0.000	-10.000	-10.000	0.000
CASE	7	1.000	2.000	-9.000	-9.000	2.236
CASE	8	2.000	4.000	-6.000	-6.000	4.472
CASE	9	3.000	6.000	-1.000	-1.000	6.708
CASE	10	4.000	8.000	6.000	6.000	8.944
CASE	11	5.000	10.000	15.000	15.000	11.180

$$PC_i = (y_{i1}X_1 + y_{i2}X_2 + y_{i3}X_3)\sqrt{\lambda_i}$$

$$= b_{i1}X_1 + b_{i2}X_2 + b_{i3}X_3$$

$$PC_2 = .447X_1 + .894X_2 + 0X_3$$

for case 1,

$$= .447(-5) + .894(-10) + 0(15)$$

$$= -11.18$$

Analyzing the correlation matrix gives the following results:

MATRIX TO BE FACTORED = Correlation Matrix (r_{ij})

	X1	X2	X3
X1	1.000		
X2	1.000	1.000	
X3	-0.000	-0.000	1.000

LATENT ROOTS (EIGENVALUES) = λ_i

1	2	3
2.000	1.000	0.000

COMPONENT LOADINGS = $b_i\sqrt{\lambda_i} = \underset{\sim}{\Lambda}_i$

	1	2
X1	1.000	0.000
X2	1.000	0.000
X3	0.000	1.000

$\underset{\sim}{b}_1' = [1 \;\; 1 \;\; 0] / \sqrt{2}$

$= [.707 \;\; .707 \;\; 0]$

VARIANCE EXPLAINED BY COMPONENTS

1	2
2.000	1.000

PERCENT OF TOTAL VARIANCE EXPLAINED

1	2
66.667	33.333

FACTOR SCORE COEFFICIENTS = $b_i / \sqrt{\lambda_i} = \underset{\sim}{y}_i$

	$1 = \underset{\sim}{y}_1$	$2 = \underset{\sim}{y}_2$
X1	0.500	0.000
X2	0.500	0.000
X3	0.000	1.000

		X1	X2	X3	FACTOR(1) $=PC_1$	FACTOR(2) $=PC_2$
CASE	1	-5.000	-10.000	15.000	-1.508	1.619
CASE	2	-4.000	-8.000	6.000	-1.206	0.648
CASE	3	-3.000	-6.000	-1.000	-0.905	-0.108
CASE	4	-2.000	-4.000	-6.000	-0.603	-0.648
CASE	5	-1.000	-2.000	-9.000	-0.302	-0.972
CASE	6	0.000	0.000	-10.000	0.000	-1.080
CASE	7	1.000	2.000	-9.000	0.302	-0.972
CASE	8	2.000	4.000	-6.000	0.603	-0.648
CASE	9	3.000	6.000	-1.000	0.905	-0.108
CASE	10	4.000	8.000	6.000	1.206	0.648
CASE	11	5.000	10.000	15.000	1.508	1.619

$$PC_i = y_{i1}X_1/S_1 + y_{i2}X_2/S_2 + y_{i3}X_3/S_3$$

$$= (b_{i1}X_1/S_1 + b_{i2}X_2/S_2 + b_{i3}X_3/S_3) / \sqrt{\lambda_i}$$

$$PC_1 = (.707\ X_1/3.317 + .707\ X_2/6.633 + 0\ X. \quad 263) / \sqrt{2}$$

for case 1,

$$= .707(-5)/3.317 + .707(-10)/6.633) / \sqrt{2}$$

$$= -1.508$$

There are several items to note in these analyses:

i) There are only two nonzero eigenvalues since X_2 can be computed from X_1.

ii) X_3 is its own principal component since it is uncorrelated with all the other variables.

iii) The sum of the eigenvalues is the sum of the variances, i.e.,

$$11 + 44 + 85.8 = 140.8$$

and

$$1 + 1 + 1 = 3 \ .$$

iv) For the variance-covariance analysis, the ratio of the coefficients of X_1 and X_2 in PC_2 is the same as the ratio of the variables themselves (since $X_2 = 2X_1$).

v) Since there are only two nonzero eigenvalues, only two of the PCs have nonzero variances (are nonconstant).

vi) The coefficients help to relate the variables and the PCs. In the variance-covariance analysis,

$$\text{Corr}(PC_2, X_1) = \frac{(\text{coefficient of } X_1 \text{ in } PC_2)\sqrt{\text{var}(PC_2)}}{\sqrt{\text{var}(X_1)}} = \frac{\Lambda_{12}}{\sqrt{\text{var}(X_1)}}$$

$$= \frac{b_{21}\sqrt{\lambda_2}}{s_1}$$

$$= \frac{.447214\sqrt{55}}{3.16625}$$

$$= 1 \ .$$

In the correlation analysis,

$$\text{Corr}(PC_1, X_1) = b_{11}\sqrt{\lambda_1} = \Lambda_{11} = \text{Component loading for } PC_1,\ X_1$$

$$= .707107\sqrt{2}$$

$$= 1 \ .$$

Thus, in both these cases, the variable is perfectly correlated with the PC.

vii) The Xs can be reconstructed exactly from the PCs with nonzero eigenvalues. For example, in the variance-covariance analysis, X_3 is clearly given by PC_1 . X_1 and X_2 can be recovered via the formulas

$$X_1 = PC_2/\sqrt{5}$$

$$X_2 = 2 \cdot PC_2/\sqrt{5} \quad .$$

As a numerical example,

$$-5 = -11.180/\sqrt{5} \ .$$

Example 3: For Example 3 we use $X_1 = Z_1$, $X_2 = 2(Z_1+5)$, $X_3 = 3(Z_1+5)$ and $X_4 = Z_2$. Thus X_1, X_2 and X_3 are all created from Z_1. The data and summary statistics are:

OBS	X1	X2	X3	X4
1	-5	0	0	15
2	-4	2	3	6
3	-3	4	6	-1
4	-2	6	9	-6
5	-1	8	12	-9
6	0	10	15	-10
7	1	12	18	-9
8	2	14	21	-6
9	3	16	24	-1
10	4	18	27	6
11	5	20	30	15

	X1	X2	X3	X4
MEAN	0.000000	10.00000	15.00000	0.00000
ST DEV	3.316625	6.63325	9.94987	9.62823

The analyses for the variance-covariance matrix (unstandardized analysis) and correlation matrix (standardized analysis) are given below.

MATRIX TO BE FACTORED = Covariance Matrix (s_{ij})

	X1	X2	X4	X3
X1	11.000			
X2	22.000	44.000		
X4	-0.000	-0.000	85.800	
X3	33.000	66.000	-0.000	99.000

Note the order that SYSTAT prints variable information. (Order is set by SYSTAT based on order variables were created).

LATENT ROOTS (EIGENVALUES) = λ_i

1	2	3	4
154.000	85.800	0.000	-0.000

COMPONENT LOADINGS = $\underset{\sim}{b}_i \sqrt{\lambda_i} = \underset{\sim}{\Lambda}_i$

$\underset{\sim}{b}_1' = [3.317 \quad 6.633 \quad 0 \quad 9.950]/\sqrt{154}$

$= [.267 \quad .535 \quad 0 \quad .802]$

	1	2
X1	3.317	0.000
X2	6.633	0.000
X4	0.000	-9.263
X3	9.950	0.000

Note: The 3rd and 4th component loadings were 0

VARIANCE EXPLAINED BY COMPONENTS

1	2
154.000	85.800

PERCENT OF TOTAL VARIANCE EXPLAINED

1	2
64.220	35.780

FACTOR SCORE COEFFICIENTS = $\underset{\sim}{b}_i / \sqrt{\lambda_i} = \underset{\sim}{y}_i$

	$1 = \underset{\sim}{y}_1$	$2 = \underset{\sim}{y}_2$
X1	0.022	0.000
X2	0.043	0.000
X4	0.000	-0.108
X3	0.065	0.000

		FACTOR(1) = PC_1	FACTOR(2) = PC_2
CASE	1	-18.708	-15.000
CASE	2	-14.967	-6.000
CASE	3	-11.225	1.000
CASE	4	-7.483	6.000
CASE	5	-3.742	9.000
CASE	6	-0.000	10.000
CASE	7	3.742	9.000
CASE	8	7.483	6.000
CASE	9	11.225	1.000
CASE	10	14.967	-6.000
CASE	11	18.708	-15.000

$$PC_i = b_{i1}\ (X_1-\bar{X}_1) + b_{i2}\ (X_2-\bar{X}_2) + b_{i3}\ (X_3-\bar{X}_3) + b_{i4}\ (X_4-\bar{X}_4)$$

$$PC_1 = 0.267\ (X_1-0) + 0.535\ (X_2-10) + 0.802\ (X_3-15) + 0\ (X_4-0)$$

for case 1,

$$= 0.267(-5) + 0.535(0-10) + 0.802(0-15)$$

$$= -18.71$$

MATRIX TO BE FACTORED = **Correlation Matrix** (r_{ij})

	X1	X2	X4	X3
X1	1.000			
X2	1.000	1.000		
X4	-0.000	-0.000	1.000	
X3	1.000	1.000	-0.000	1.000

LATENT ROOTS (EIGENVALUES) = λ_i

1	2	3	4
3.000	1.000	0.000	-0.000

COMPONENT LOADINGS = $\underset{\sim}{b}_i \sqrt{\lambda_i} = \underset{\sim}{\Lambda}_i$

	1	2
X1	1.000	0.000
X2	1.000	0.000
X4	-0.000	-1.000
X3	1.000	0.000

$\underset{\sim}{b}_1' = [1 \quad 1 \quad 0 \quad 1] / \sqrt{3}$

$= [.577 \quad .577 \quad 0 \quad .577]$

VARIANCE EXPLAINED BY COMPONENTS

1	2
3.000	1.000

PERCENT OF TOTAL VARIANCE EXPLAINED

1	2
75.000	25.000

FACTOR SCORE COEFFICIENTS = $\underset{\sim}{b}_i / \sqrt{\lambda_i} = \underset{\sim}{y}_i$

	$1 = \underset{\sim}{y}_1$	$2 = \underset{\sim}{y}_2$
X1	0.333	-0.000
X2	0.333	-0.000
X4	-0.000	-1.000
X3	0.333	-0.000

	FACTOR(1) $=PC_1$	FACTOR(2) $=PC_2$
CASE 1	-1.508	-1.619
CASE 2	-1.206	-0.648
CASE 3	-0.905	0.108
CASE 4	-0.603	0.648
CASE 5	-0.302	0.972
CASE 6	0.000	1.080
CASE 7	0.302	0.972
CASE 8	0.603	0.648
CASE 9	0.905	0.108
CASE 10	1.206	-0.648
CASE 11	1.508	-1.619

$$PC_i = y_{i1}(X_1-\bar{X}_1)/s_1 + y_{i2}(X_2-\bar{X}_2)/s_2 + y_{i3}(X_3-\bar{X}_3)/s_3 + y_{i4}(X_4-\bar{X}_4)/s_4$$

$$PC_1 = .333(X_1-0)/3.317 + .333(X_2-10)/6.633 + .333(X_3-15)/9.950 + 0(X_4-0)/9.628$$

for case 1

$$= .333(-5)/3.317 + .333(-10)/6.633 + .333(-15)/9.950$$

$$= -1.508$$

For the variance-covariance analysis, the coefficients in PC_1 are in the same ratio as their relationship to Z_1. In the correlation analysis X_1, X_2 and X_3 have equal coefficients. In both analyses, as expected, the total variance is equal to the sum of the variances for the PCs. In both cases two PCs, PC_3 and PC_4, have zero variance and are identically zero.

Example 4. In this example we take more complicated combinations of Z_1 and Z_2.

$$X_1 = Z_1$$

$$X_2 = 2Z_1$$

$$X_3 = 3Z_1$$

$$X_4 = Z_1/2 + Z_2$$

$$X_5 = Z_1/4 + Z_2$$

$$X_6 = Z_1/8 + Z_2$$

$$X_7 = Z_2$$

Note that X_1, X_2 and X_3 are colinear (they all have correlation unity) and X_4, X_5, X_6 and X_7 have steadily decreasing correlations with X_1. The data and data summaries are below.

OBS	X1	X2	X3	X4	X5	X6	X7
1	-5.000	-10.000	-15.000	12.500	13.750	14.375	15.000
2	-4.000	-8.000	-12.000	4.000	5.000	5.500	6.000
3	-3.000	-6.000	-9.000	-2.500	-1.750	-1.375	-1.000
4	-2.000	-4.000	-6.000	-7.000	-6.500	-6.250	-6.000
5	-1.000	-2.000	-3.000	-9.500	-9.250	-9.125	-9.000
6	0.000	0.000	0.000	-10.000	-10.000	-10.000	-10.000
7	1.000	2.000	3.000	-8.500	-8.755	-8.875	-9.000
8	2.000	4.000	6.000	-5.000	-5.500	-5.750	-6.000
9	3.000	6.000	9.000	0.500	-0.250	-0.625	-1.000
10	4.000	8.000	12.000	8.000	7.000	6.500	6.000
11	5.000	10.000	15.000	17.500	16.250	15.625	15.000

	X1	X2	X3	X4	X5	X6	X7
Mean	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
ST DEV	3.31662	6.63325	9.94987	9.41010	9.29987	9.27210	9.26283

The PCAs for the variance-covariance and correlation matrices are given below.

MATRIX TO BE FACTORED = Covariance Matrix

	X1	X2	X4	X5	X6
X1	11.000				
X2	22.000	44.000			
X4	5.500	11.000	88.550		
X5	2.750	5.500	87.175	86.488	
X6	1.375	2.750	86.488	86.144	85.972
X7	-0.000	-0.000	85.800	85.800	85.800
X3	33.000	66.000	16.500	8.250	4.125

	X7	X3
X7	85.800	
X3	-0.000	99.000

LATENT ROOTS (EIGENVALUES) = λ_i

1	2	3	4	5
347.015	153.794	0.000	0.000	-0.000

6	7
-0.000	-0.000

COMPONENT LOADINGS = $\underset{\sim}{b}_i \sqrt{\lambda_i} = \underset{\sim}{\Lambda}_i$

	1	2
X1	0.466	3.284
X2	0.932	6.567
X4	9.404	0.340
X5	9.287	-0.481
X6	9.229	-0.891
X7	9.171	-1.302
X3	1.398	9.851

$\underset{\sim}{b}_1' = [.466 \quad .932 \quad 9.404 \quad 9.287 \quad 9.229 \quad 9.171 \quad 1.398] / \sqrt{347.015}$

$= [.025 \quad .050 \quad .505 \quad .499 \quad .495 \quad .492 \quad .075]$

VARIANCE EXPLAINED BY COMPONENTS

	1	2
	347.015	153.794

PERCENT OF TOTAL VARIANCE EXPLAINED

	1	2
	69.291	30.709

FACTOR SCORE COEFFICIENTS = $\tilde{b}_i/\sqrt{\lambda_i} = \tilde{y}_i$

	$1 = \tilde{y}_1$	$2 = \tilde{y}_2$
X1	0.001	0.021
X2	0.003	0.043
X4	0.027	0.002
X5	0.027	-0.003
X6	0.027	-0.006
X7	0.026	-0.008
X3	0.004	0.064

	FACTOR(1)	FACTOR(2)
CASE 1	25.921	-21.332
CASE 2	8.790	-15.937
CASE 3	-4.359	-10.918
CASE 4	-13.525	-6.275
CASE 5	-18.709	-2.009
CASE 6	-19.911	1.881
CASE 7	-17.131	5.395
CASE 8	-10.368	8.533
CASE 9	0.377	11.294
CASE 10	15.104	13.679
CASE 11	33.813	15.688

$$PC_i = b_{i1}X_1 + b_{i2}X_2 + b_{i3}X_3 + b_{i4}X_4 + b_{i5}X_5 + b_{i6}X_6 + b_{i7}X_7$$

$$PC_1 = .025X_1 + .050X_2 + .075X_3 + .505X_4 + .499X_5 + .495X_6 + .492X_7$$

for case 1

$$25.921 = .025(-5) + .050(-10) + .075(-15) + .505(12.4) + .499(13.75) + .495(14.375) +. 492(15)$$

MATRIX TO BE FACTORED = **Correlation Matrix**

	X1	X2	X4	X5	X6
X1	1.000				
X2	1.000	1.000			
X4	0.176	0.176	1.000		
X5	0.089	0.089	0.996	1.000	
X6	0.045	0.045	0.991	0.999	1.000
X7	-0.000	-0.000	0.984	0.996	0.999
X3	1.000	1.000	0.176	0.089	0.045

	X7	X3
X7	1.000	
X3	-0.000	1.000

LATENT ROOTS (EIGENVALUES) = λ_i

1	2	3	4	5
4.052	2.948	0.000	0.000	0.000

6	7
-0.000	-0.000

COMPONENT LOADINGS = $\underset{\sim}{b}_i \sqrt{\lambda_i} = \underset{\sim}{\Lambda}_i$

	1	2
X1	0.290	-0.957
X2	0.290	-0.957
X4	0.993	0.117
X5	0.979	0.204
X6	0.969	0.247
X7	0.957	0.290
X3	0.290	-0.957

VARIANCE EXPLAINED BY COMPONENTS

1	2
4.052	2.948

PERCENT OF TOTAL VARIANCE EXPLAINED

1	2
57.888	42.112

FACTOR SCORE COEFFICIENTS = $\underset{\sim}{b}_i / \sqrt{\lambda_i} = \underset{\sim}{y}_i$

	$1 = \underset{\sim}{y}_1$	$2 = \underset{\sim}{y}_2$
X1	0.072	-0.325
X2	0.072	-0.325
X4	0.245	0.040
X5	0.242	0.069
X6	0.239	0.084
X7	0.236	0.099
X3	0.072	-0.325

	FACTOR(1)	FACTOR(2)
CASE 1	1.112	1.913
CASE 2	0.270	1.342
CASE 3	-0.366	0.834
CASE 4	-0.795	0.389
CASE 5	-1.017	0.006
CASE 6	-1.033	-0.314
CASE 7	-0.842	-0.571
CASE 8	-0.445	-0.765
CASE 9	0.159	-0.897
CASE 10	0.970	-0.966
CASE 11	1.987	-0.972

We note several things:

i) In both analyses there are only two eigenvalues that are nonzero indicating that only two variables are needed. This is not readily apparent from the correlation or variance-covariance matrix.

ii) In PC_1, PC_2 and PC_3 where the standardizrd X_1, X_2 and X_3 are the same, they have the same coefficients.

iii) Neither PCA recovers Z_1 and Z_2. The PCAs with nonzero variances have elements of both Z_1 and Z_2 in them, i.e., neither PC_1 or PC_2 is perfectly correlated with one of the Zs.

4. SUMMARY

PCA provides a method of extracting structure from the variance-covariance or correlation matrix. If a multivariate data set is actually constructed in a linear fashion from fewer variables, then PCA will discover that structure. PCA constructs linear combinations of the original data, $\tilde{X}$, with maximal variance:

$$\tilde{P} = \tilde{X}B \ .$$

This relationship can be inverted to recover the Xs from the PCs (actually only those PCs with nonzero eigenvalues are needed - see example 2). Though PCA will often help discover structure in a data set, it does have limitations. It will not necessarily recover the exact underlying variables, even if they were uncorrelated (Example 4). Also, by its construction, PCA is limited to searching for linear structures in the Xs.

APPENDIX

Control Language

Control language is typed in upper case and comments are in lower case. Refer to SYSTAT, Version 3, 1986, for program documentation.

FACTOR	→	typed from DOS
USE PCA1	→	instructs SYSTAT to perform the analysis on the previously saved data file PCA1.SYS
SAVE PCACOR1	→	instructs SYSTAT to save the PC scores in order that they may be printed later with the DATA module
NUMBER = 2	→	indicates the number of components to print
FACTOR	→	instructs SYSTAT to perform the PCA on all variables in PCA1

* SYSTAT will compute the PCA on the correlation matrix unless otherwise directed. To request PCA on a variance-covariance matrix add the following command somewhere before the FACTOR command:

TYPE = COVARIANCE

END

10-81

DTIC