PRINCIPAL COMPONENTS

PURPOSE

Find the principal components of a matrix.

DESCRIPTION

Given a data matrix X with n cases and p variables (i.e., variables X_1, X_2, ... , X_p) a linear transformation to a new set of variables Y_1, Y_2, ... , Y_p can be calculated as:

$$Y_1 = a_{11}X_1 + a_{21}X_2 + ... + a_{p1}X_p$$
$$Y_2 = a_{12}X_1 + a_{22}X_2 + ... + a_{p2}X_p$$
$$....$$
$$Y_p = a_{1p}X_1 + a_{2p}X_2 + ... + a_{pp}X_p$$

The principal components are a specific linear combination that are chosen so that the Y_i (called the principal components) have the following characteristics:

1. The p principal components are uncorrelated.
2. The first principal component explains the largest percentage of the variation in the original p-dimensional data set (and the second principal component explains the second largest percentage and so on). Typically the first few principal components account for most of the variation while the remaining principal components make a negligible contribution.

Principal components are used to reduce large dimensional data sets to data sets with a few dimensions that still retain most of the information in the original data matrix. That is, typically only the first few principal components are used. If the first few principal components do not account for most of the variation, there is little advantage to using them. By reducing the dimensionality of the original data, principal components can often simplify many analyses. The primary disadvantage of principal components is that interpretation can be more difficult since you are no longer working with the original variables and the principal components are heavily affected by the scaling of the variables.

Principal components reduce to the problem of finding the eigenvalues and eigenvectors of the covariance (or correlation matrix) of the data matrix. The ith row of the A matrix is the eigenvector corresponding to the ith eigenvalue of the covariance (or correlation) matrix. Also, the proportion of the variance in the original data matrix explained by the ith principal component is the ith eigenvalue divided by the sum of all p eigenvalues. Using the covariance matrix is preferred when the original data have reasonably comparable scales. If this is not the case, the correlation matrix is preferred.

DATAPLOT returns the principal components (i.e., the Y matrix), the eigenvectors (i.e., the A matrix), and the eigenvalues in separate steps.

SYNTAX 1

LET <mat2> = PRINCIPAL COMPONENTS <mat1> <SUBSET/EXCEPT/FOR qualification>

where <mat1> is the matrix for which the principal components are computed (it can be the original data matrix, a covariance matrix, or a correlation matrix);

<mat2> is a p by p matrix where the principal components (i.e., the Y matrix) are saved (column 1 is the first principal component, column p is the pth principal component);

and where the <SUBSET/EXCEPT/FOR qualification> is optional and rarely used in this context.

SYNTAX 2

LET <var> = <id> PRINCIPAL COMPONENTS <mat1> <SUBSET/EXCEPT/FOR qualification>

where <mat1> is the matrix for which the principal components are computed (it can be the original data matrix, a covariance matrix, or a correlation matrix);

<id> identifies a specific principal component to be returned (FIRST, SECOND, THIRD, FOURTH, FIFTH, SIXTH, SEVENTH, EIGHTH, NINTH, TENTH);

<var> is a variable of length p where the specific principal component is saved;

and where the <SUBSET/EXCEPT/FOR qualification> is optional and rarely used in this context.

SYNTAX 3

LET <var> = PRINCIPAL COMPONENTS EIGENVALUES <mat1> <SUBSET/EXCEPT/FOR qualification>

where <mat1> is the matrix for which the principal components are computed (it can be the original data matrix, a covariance matrix, or a correlation matrix);

<var> is a variable of length p where the eigenvalues of the covariance (or correlation) matrix are saved;

and where the <SUBSET/EXCEPT/FOR qualification> is optional and rarely used in this context.

SYNTAX 4

LET <par> = <id> PRINCIPAL COMPONENTS EIGENVALUES <mat1> <SUBSET/EXCEPT/FOR qualification>
where <mat1> is the matrix for which the principal components are computed (it can be the original data matrix, a covariance matrix, or a correlation matrix);
<id> identifies a specific eigenvalue to be returned (FIRST, SECOND, THIRD, FOURTH, FIFTH, SIXTH, SEVENTH, EIGHTH, NINTH, TENTH);
<par> is a parameter where the specific eigenvalue is saved;
and where the <SUBSET/EXCEPT/FOR qualification> is optional and rarely used in this context.

SYNTAX 5

LET <mat2> = PRINCIPAL COMPONENTS EIGENVECTORS <mat1> <SUBSET/EXCEPT/FOR qualification>
where <mat1> is the matrix for which the principal components are computed (it can be the original data matrix, a covariance matrix, or a correlation matrix);
<mat2> is a p by p matrix where the eigenvectors of the covariance (or correlation) matrix (i.e., the A matrix) are saved;
and where the <SUBSET/EXCEPT/FOR qualification> is optional and rarely used in this context.

SYNTAX 6

LET <var> = <id> PRINCIPAL COMPONENTS EIGENVECTORS <mat1> <SUBSET/EXCEPT/FOR qualification>
where <mat1> is the matrix for which the principal components are computed (it can be the original data matrix, a covariance matrix, or a correlation matrix);
<id> identifies a specific eigenvector to be returned (FIRST, SECOND, THIRD, FOURTH, FIFTH, SIXTH, SEVENTH, EIGHTH, NINTH, TENTH);
<var> is a variable where the specific eigenvector is saved;
and where the <SUBSET/EXCEPT/FOR qualification> is optional and rarely used in this context.

EXAMPLES

```
LET Y = PRINCIPALECOMPONENTS X
LET E = PRINCIPAL COMPONENTS EIGENVALUES X
LET A = PRINCIPAL COMPONENTS EIGENVECTORS X
LET Y1 = FIRST PRINCIPAL COMPONENT X
LET E1 = FIRST PRINCIPAL COMPONENT EIGENVALUE X
LET A1 = FIRST PRINCIPAL COMPONENT EIGENVECTOR X
```

NOTE 1

If the principal components are derived from the original data and the covariance matrix, the data matrix is scaled by subtracting the column mean before it is multiplied by the eigenvector matrix. That is, $Y = A*(X-\overline{X})$.

NOTE 2

If you have more than 750 rows in your input data, you can use the LOOP command to generate a covariance matrix and compute the principal components from this covariance matrix.

NOTE 3

If the principal components are derived from the covariance matrix, then:

$$COV(X_i,Y_j) = l_j a_{ij}$$
$$CORR(X_i,Y_j) = a_{ij} SQRT(l_j)/S_i$$

where the above equations represent the covariance and the correlation between the ith response and the jth principal component. The l_j is the jth principal component eigenvalue and S_i is the standard deviation of the ith response. If the principal components are derived from the correlation matrix, the correlation equation becomes:

$$CORR(X_i,Y_j) = SQRT(l_j) a_{ij}$$

The first program example uses this to show a graphical representation of the principal components. It plots the correlation with the first component against the correlation with the second component. This is a form of scaling in that it can be used to place the original variables in homogeneous groups.

NOTE 4

The PRINCIPAL COMPONENT TYPE command (documented in the SUPPORT chapter of Volume 1) is used to specify whether the original matrix is a data matrix, a covariance matrix, or a correlation matrix. It also specifies whether the principal components are derived from the covariance or the correlation matrix. The default is an original data matrix with the principal components derived from the correlation matrix.

NOTE 5

Principal components are sometimes used in regression problems to avoid multicolinearity problems while still maintaining most of the information in a large number of variables. This is demonstrated in the second program example. For the sake of simplicity, this is demonstrated with a small number of starting variables. In practice, this technique is normally used when there are a large number of independent variables.

DEFAULT

None

SYNONYMS

None

RELATED COMMANDS

MATRIX EIGENVALUES	=	Compute the matrix eigenvalues.
MATRIX EIGENVECTORS	=	Compute the matrix eigenvectors.
MATRIX SUBTRACTION	=	Perform a matrix subtraction.
CORRELATION MATRIX	=	Compute the correlation matrix of a matrix.
VARIANCE-COVA MATRIX	=	Compute the variance-covariance matrix of a matrix.
SINGULAR VALUES	=	Compute the singular values of a matrix.
SINGULAR VALUE DECOM	=	Compute the singular value decomposition of a matrix.
SINGULAR VALUE FACT	=	Compute the singular value factorization of a matrix.

REFERENCE

“Principal Components and Factor Analysis: Part I - Principal Components,” Jackson, Journal of Quality Technology, October, 1980.

“Multivariate Statistical Methods,” Morrison, McGraw-Hill, 1976.

“Graphical Exploratory Data Analysis,” du Toit, Steyn, and Stumpf, Springer-Verlang, 1986.

“Applied Regression Analysis,” Draper and Smith, John Wiley, 1981.

APPLICATIONS

Multivariate Analysis, Regression

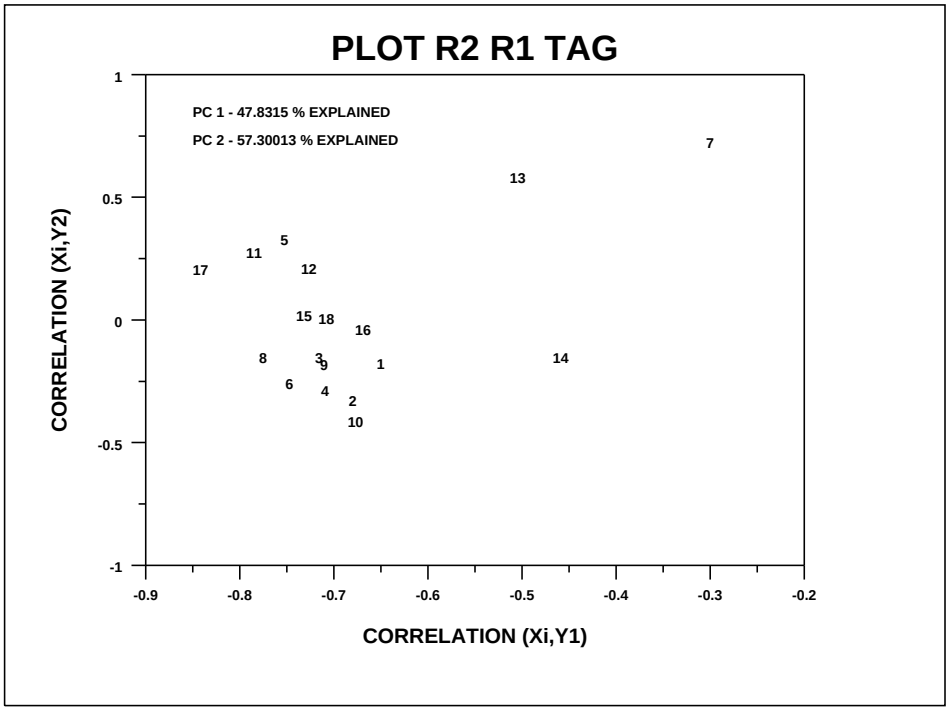
IMPLEMENTATION DATE

87/10

PROGRAM 1

```
. SOURCE: "GRAPHICAL EXPLORATORY DATA ANALYSIS", DU TOIT, ET AL
. 28 STUDENTS FROM ABILITY TEST DATA SET (PAGE 6)
.
DIMENSION 100 COLUMNS
READ MATRIX X
19 21 21 18 20 21 15 14 15 13 15 16 19 19 19 20 17 17
21 20 15 24 22 18 11 18 16 19 14 17 21 15 17 18 18 19
18 19 16 18 18 23 11 13 13 15 11 11 15 18 13 15 18 13
18 23 10 18 16 16 11  9  8 15  6  9 12 16  8 13  9 15
24 24 19 20 23 24 22 18 16 19 16 19 19 21 21 20 18 20
19 19 23 21 23 23  9  8 13 15 20 15 17 12 20 16 16 21
21 20 19 21 21 23 11 16 11 18 18 14 21 17 14 19 18 16
21 20 21 20 16 22  7 11 17 16  8 10 13 17 16 17 15 11
19 20 19 22 18 21 11 12  7 15  9 11 13 12 13 17 12 12
19 23 22 22 16 25 12 15 16 19 15 10 15 20 18 18 17 13
17 13  8 18 13 18 12  8  9 12 12 11  9 14 15 12 13  9
21 22 22 15 23 23 16 12 16 15 13 14 19 17 16 18 19 18
18 18 17 16 15 22  8 11 10 16  8 14 10 13 10 14  9 14
13 18 21 16 17 15 11 12 11  9 11 11 16 18 14 13 15 18
17 13 17 20 22 19 15 11 11 12 11 13 15 15 15 13 16 12
18 12  9  9 15 17  9  5  3 12  7  7 12 10 10 13 10 12
22 15 24 17 15 20 10 12 12 11  9 12 19 16 16  8 11 17
18 17 18 18 13 18 14 12 15 11 10  9 21 14 12 15 11 13
17 15 14 14 12 13  9 10 11  9  7 11 13 15 11 13 10 13
16 20 17 13 15 16 10 16 12 10  7 13 12 18 13 18 10 15
24 21 22 21 21 25 11 17 17 21 11 15 15 18 16 16 17 17
23 23 21 22 16 21 10 18 16 14 14 13 17 21 19 16 17 19
22 22 21 24 18 24  6 16 14 20 16 18 12 12 13 18 18 21
22 17 19 19 21 20 17 15  9 13 16 17 18 11 13 16 19 14
20 23 23 22 22 24 11 18 16 16 16 20 13 16 18 18 20 20
22 17 21 17 17 22 10 14 16 16 13  8 13 18 21 12 13 15
21 18 20 23 21 22  8 15  9 17 11 13 13 20 20 21 15 20
21 22 19 20 18 17 11 15 12 14 11 10 11 13 14 14 15 14
END OF DATA
LET P = MATRIX NUMBER OF COLUMNS X
PRINCIPAL COMPONENTS TYPE DATA COVARIANCE
LET Y = PRINCIPAL COMPONENTS X
LET A = PRINCIPAL COMPONENTS EIGENVECTORS X
LET E = PRINCIPAL COMPONENTS EIGENVALUES X
PRINT E
LET ESUM = CUMULATIVE SUM E
LET TEMP = ESUM(P)
LET RATIO = 100*ESUM/TEMP
. FORM CORR(Xi,Y1) = Ai1*SQRT(L1)/Si
LOOP FOR K = 1 1 P
    LET TEMP = STANDARD DEVIATION X^K
    LET SD(K) = TEMP
END OF LOOP
LET TEMP = E(1)
LET R1 = A1*SQRT(TEMP)/SD
. FORM CORR(Xi,Y2) = Ai2*SQRT(L2)/Si
LET TEMP = E(2)
LET R2 = A2*SQRT(TEMP)/SD
LINE BLANK ALL
CHARACTER 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
LET TAG = SEQUENCE 1 1 P
```

```
X1LABEL CORRELATION (XLC()I,UC()Y1)
Y1LABEL CORRELATION (XLC()I,UC()Y2)
LET TEMP = RATIO(1)
LEGEND 1 PC 1 - ^TEMP % EXPLAINED
LET TEMP = RATIO(2)
LEGEND 2 PC 2 - ^TEMP % EXPLAINED
PLOT R2 R1 TAG
```



PROGRAM 2

```
. SOURCE: "APPLIED REGRESSION ANALYSIS", DRAPER AND SMITH
. HALD'S DATA SET (APPENDIX B)
.
FEEDBACK OFF
DIMENSION 100 COLUMNS
READ X1 X2 X3 X4 Y
7 26 6 60 78.5
1 29 15 52 74.3
11 56 8 20 104.3
11 31 8 47 87.6
7 52 6 33 95.9
11 55 9 22 109.2
3 71 17 6 102.7
1 31 22 44 72.5
2 54 18 22 93.1
21 47 4 26 115.9
1 40 23 34 83.8
11 66 9 12 113.3
10 68 8 12 109.4
END OF DATA
LET X = MATRIX DEFINITION X1 13 4
LET CORR = CORRELATION MATRIX X
PRINT "CORRELATION MATRIX"; PRINT CORR
.
PRINCIPAL COMPONENT TYPE DATA CORRELATION
LET W = PRINCIPAL COMPONENTS X
LET E = PRINCIPAL COMPONENT EIGENVALUES X
LET EVECT = PRINCIPAL COMPONENT EIGENVECTORS X
LET RATIO = CUMULATIVE SUM E
LET RATIO = RATIO/4.
.
PRINT; PRINT "EIGENVALUES (AND PROPORTION OF VARIANCE OF X VARIABLES):"
PRINT E RATIO
PRINT; PRINT "EIGENVECTORS:"; PRINT EVECT
PRINT; PRINT "PRINCIPAL COMPONENTS:"; PRINT W
FEEDBACK ON
FIT Y W1 W2
```

The following output is generated.

```
Correlation matrix

VARIABLES--CORR1          CORR2          CORR3          CORR4

  0.1000000E+01  0.2285795E+00 -0.8241338E+00 -0.2454451E+00
  0.2285795E+00  0.1000000E+01 -0.1392424E+00 -0.9729550E+00
 -0.8241338E+00 -0.1392424E+00  0.1000000E+01  0.2953700E-01
 -0.2454451E+00 -0.9729550E+00  0.2953700E-01  0.1000000E+01

eigenvalues (and proportion of variance of X variables):

VARIABLES--E              RATIO

  0.2235704E+01  0.5589260E+00
  0.1576066E+01  0.9529425E+00
  0.1866062E+00  0.9995940E+00
```

```
  0.1623492E-02  0.9999999E+00

eigenvectors:

VARIABLES--EVECT1         EVECT2         EVECT3         EVECT4

  0.4759549E+00  0.5089797E+00  0.6755005E+00  0.2410518E+00
  0.5638706E+00 -0.4139312E+00 -0.3144200E+00  0.6417563E+00
 -0.3940663E+00 -0.6049693E+00  0.6376914E+00  0.2684656E+00
 -0.5479314E+00  0.4512348E+00 -0.1954205E+00  0.6767342E+00

Principal components:

VARIABLES--W1             W2             W3             W4

 -0.2687606E+02  0.2596251E+02 -0.2887760E+01  0.4424562E+01
 -0.2720332E+02  0.1261223E+02 -0.5814366E+00  0.1905836E+01
  0.1307300E+02 -0.3678842E+01 -0.5261565E+00 -0.1890980E+01
 -0.1581791E+02  0.1885278E+02  0.2057991E+01  0.3369364E+00
  0.2578727E+01  0.3016956E+01 -0.5786327E+01  0.2838402E+01
  0.1101921E+02 -0.2967411E+01  0.3511390E-01 -0.9108021E+00
  0.2184787E+02 -0.2572166E+02 -0.2171351E+01 -0.1251138E+01
 -0.2445059E+02  0.3939706E+01  0.4816927E+01 -0.3452657E+00
  0.2625144E+01 -0.1257902E+02  0.9252250E-02 -0.1305834E+01
  0.1104639E+02  0.1026362E+02  0.5335340E+01 -0.2269726E+01
 -0.1429051E+02 -0.4902992E+01  0.4579043E+01 -0.1068336E+01
  0.2270110E+02 -0.1203300E+02 -0.1469301E+01 -0.6188249E+00
  0.2374695E+02 -0.1276487E+02 -0.3411333E+01  0.1551703E+00

 LEAST SQUARES MULTILINEAR FIT
       SAMPLE SIZE N       =       13
       NUMBER OF VARIABLES =        2
       NO REPLICATION CASE


               PARAMETER ESTIMATES           (APPROX. ST. DEV.)    T VALUE
        1  A0                    95.4231       (0.9145    )         0.10E+03
        2  A1       W1          0.982327       (0.7375E-01)          13.
        3  A2       W2          0.462248       (0.9908E-01)          4.7

       RESIDUAL    STANDARD DEVIATION =         3.2972862720
       RESIDUAL    DEGREES OF FREEDOM =          10
       COEF AND SD(COEF) WRITTEN TO FILE DPST1F.DAT
       VARIOUS DIAGNOSTIC STATISTICS WRITTEN TO FILE DPST2F.DAT
```