Chapter 52
The PRINCOMP Procedure

Chapter Table of Contents

Chapter 52
The PRINCOMP Procedure

Overview

The PRINCOMP procedure performs principal component analysis. As input you can use raw data, a correlation matrix, a covariance matrix, or a sums of squares and crossproducts (SSCP) matrix. You can create output data sets containing eigenvalues, eigenvectors, and standardized or unstandardized principal component scores.

Principal component analysis is a multivariate technique for examining relationships among several quantitative variables. The choice between using factor analysis and principal component analysis depends in part upon your research objectives. You should use the PRINCOMP procedure if you are interested in summarizing data and detecting linear relationships. Plots of principal components are especially valuable tools in exploratory data analysis. You can use principal components to reduce the number of variables in regression, clustering, and so on. See Chapter 6, "Introduction to Multivariate Procedures," for a detailed comparison of the PRINCOMP and FACTOR procedures.

Principal component analysis was originated by Pearson (1901) and later developed by Hotelling (1933). The application of principal components is discussed by Rao (1964), Cooley and Lohnes (1971), and Gnanadesikan (1977). Excellent statistical treatments of principal components are found in Kshirsagar (1972), Morrison (1976), and Mardia, Kent, and Bibby (1979).

Given a data set with p numeric variables, you can compute p principal components. Each principal component is a linear combination of the original variables, with coefficients equal to the eigenvectors of the correlation or covariance matrix. The eigenvectors are customarily taken with unit length. The principal components are sorted by descending order of the eigenvalues, which are equal to the variances of the components.

Principal components have a variety of useful properties (Rao 1964; Kshirsagar 1972):

- The eigenvectors are orthogonal, so the principal components represent jointly perpendicular directions through the space of the original variables.
- The principal component scores are jointly uncorrelated. Note that this property is quite distinct from the previous one.
- The first principal component has the largest variance of any unit-length linear combination of the observed variables. The jth principal component has the largest variance of any unit-length linear combination orthogonal to the first $j - 1$ principal components. The last principal component has the smallest variance of any linear combination of the original variables.

- The scores on the first j principal components have the highest possible generalized variance of any set of unit-length linear combinations of the original variables.
- The first j principal components provide a least-squares solution to the model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

  where $\mathbf{Y}$ is an $n \times p$ matrix of the centered observed variables; $\mathbf{X}$ is the $n \times j$ matrix of scores on the first j principal components; $\mathbf{B}$ is the $j \times p$ matrix of eigenvectors; $\mathbf{E}$ is an $n \times p$ matrix of residuals; and you want to minimize trace$(\mathbf{E}'\mathbf{E})$, the sum of all the squared elements in $\mathbf{E}$. In other words, the first j principal components are the best linear predictors of the original variables among all possible sets of j variables, although any nonsingular linear transformation of the first j principal components would provide equally good prediction. The same result is obtained if you want to minimize the determinant or the Euclidean (Schur, Frobenious) norm of $\mathbf{E}'\mathbf{E}$ rather than the trace.
- In geometric terms, the j-dimensional linear subspace spanned by the first j principal components provides the best possible fit to the data points as measured by the sum of squared perpendicular distances from each data point to the subspace. This is in contrast to the geometric interpretation of least squares regression, which minimizes the sum of squared vertical distances. For example, suppose you have two variables. Then, the first principal component minimizes the sum of squared perpendicular distances from the points to the first principal axis. This is in contrast to least squares, which would minimize the sum of squared vertical distances from the points to the fitted line.

Principal component analysis can also be used for exploring polynomial relationships and for multivariate outlier detection (Gnanadesikan 1977), and it is related to factor analysis, correspondence analysis, allometry, and biased regression techniques (Mardia, Kent, and Bibby 1979).

Getting Started

The following example uses the PRINCOMP procedure to analyze mean daily temperatures in selected cities in January and July. Both the raw data and the principal components are plotted to illustrate how principal components are orthogonal rotations of the original variables.

The following statements create the Temperature data set:

```
data Temperature;
   title 'Mean Temperature in January and July for
          Selected Cities';
   input City $1-15 January July;
   datalines;
Mobile          51.2 81.6
Phoenix         51.2 91.2
Little Rock     39.5 81.4
```

```
Sacramento      45.1 75.2
Denver          29.9 73.0
Hartford        24.8 72.7
Wilmington      32.0 75.8
Washington DC   35.6 78.7
Jacksonville    54.6 81.0
Miami           67.2 82.3
Atlanta         42.4 78.0
Boise           29.0 74.5
Chicago         22.9 71.9
Peoria          23.8 75.1
Indianapolis    27.9 75.0
Des Moines      19.4 75.1
Wichita         31.3 80.7
Louisville      33.3 76.9
New Orleans     52.9 81.9
Portland, ME    21.5 68.0
Baltimore       33.4 76.6
Boston          29.2 73.3
Detroit         25.5 73.3
Sault Ste Marie 14.2 63.8
Duluth           8.5 65.6
Minneapolis     12.2 71.9
Jackson         47.1 81.7
Kansas City     27.8 78.8
St Louis        31.3 78.6
Great Falls     20.5 69.3
Omaha           22.6 77.2
Reno            31.9 69.3
Concord         20.6 69.7
Atlantic City   32.7 75.1
Albuquerque     35.2 78.7
Albany          21.5 72.0
Buffalo         23.7 70.1
New York        32.2 76.6
Charlotte       42.1 78.5
Raleigh         40.5 77.5
Bismarck         8.2 70.8
Cincinnati      31.1 75.6
Cleveland       26.9 71.4
Columbus        28.4 73.6
Oklahoma City   36.8 81.5
Portland, OR    38.1 67.1
Philadelphia    32.3 76.8
Pittsburgh      28.1 71.9
Providence      28.4 72.1
Columbia        45.4 81.2
Sioux Falls     14.2 73.3
Memphis         40.5 79.6
Nashville       38.3 79.6
Dallas          44.8 84.8
El Paso         43.6 82.3
Houston         52.1 83.3
Salt Lake City  28.0 76.7
```

```
Burlington      16.8 69.8
Norfolk         40.5 78.3
Richmond        37.5 77.9
Spokane         25.4 69.7
Charleston, WV  34.5 75.0
Milwaukee       19.4 69.9
Cheyenne        26.6 69.1
;
```

The following statements plot the temperature data set. For information on the %PLOTIT macro, see Appendix B, "Using the %PLOTIT Macro."

```
title2 'Plot of Raw Data';
%plotit(data=Temperature, labelvar=City,
        plotvars=July January, color=black, colors=blue);
run;
```

The results are displayed in Figure 52.1, which shows a scatter diagram of the 64 pairs of data points with July temperatures plotted against January temperatures.

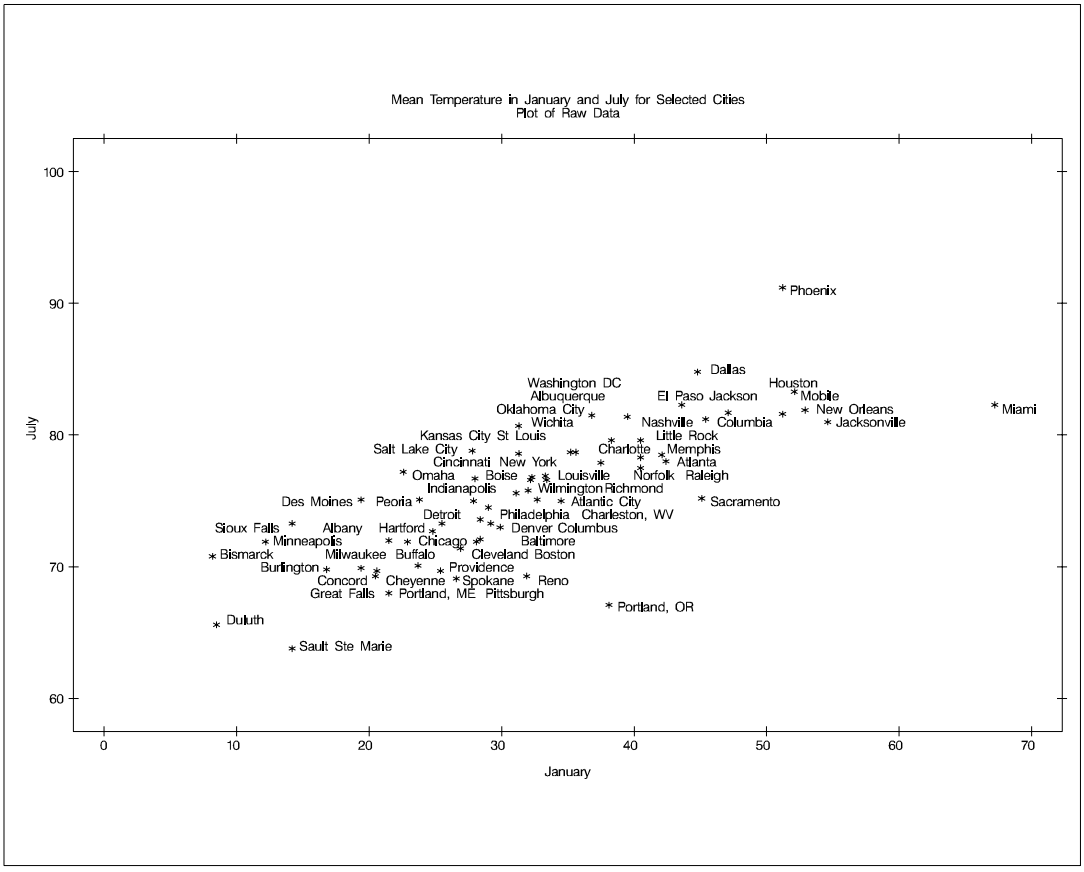


Figure 52.1. Plot of Raw Data

The following statement requests a principal component analysis on the Temperature data set and outputs the scores to the Prin data set (OUT= Prin):

```
proc princomp data=Temperature cov out=Prin;
title2;
var July January;
run;
```

Figure 52.2 displays the PROC PRINCOMP output, beginning with simple statistics. The standard deviation of January (11.712) is higher than the standard deviation of July (5.128). The COV option in the PROC PRINCOMP statement requests the principal components to be computed from the covariance matrix. The total variance is 163.474. The first principal component explains about 94 percent of the total variance, and the second principal component explains only about 6 percent. Note that the eigenvalues sum to the total variance.

From the Eigenvectors matrix, you can represent the first principal component Prin1 as a linear combination of the original variables

$$\text{Prin1} = 0.3435 \times (\text{July} - \overline{\text{July}}) + 0.9391 \times (\text{January} - \overline{\text{January}})$$

and, similarly, the second principal component Prin2 as

$$\text{Prin2} = 0.9391 \times (\text{July} - \overline{\text{July}}) - 0.3435 \times (\text{January} - \overline{\text{January}})$$

where $\overline{\text{July}}$ and $\overline{\text{January}}$ are the means of July temperatures and January temperatures, respectively. Note that January receives a higher loading on Prin1 because it has a higher standard deviation than July, and the PRINCOMP procedure calculates the scores using the centered variables rather than the standardized variables.

```
        Mean Temperature in January and July for Selected Cities

                        The PRINCOMP Procedure

                      Observations          64
                      Variables              2


                         Simple Statistics

                                July           January

                Mean       75.60781250       32.09531250
                StD         5.12761910       11.71243309


                         Covariance Matrix

                                July           January

              July           26.2924777        46.8282912
              January        46.8282912       137.1810888


                    Total Variance    163.47356647


                  Eigenvalues of the Covariance Matrix

              Eigenvalue    Difference    Proportion    Cumulative

        1    154.310607    145.147647        0.9439        0.9439
        2      9.162960                      0.0561        1.0000


                           Eigenvectors

                                Prin1         Prin2

                   July         0.343532      0.939141
                   January      0.939141      -.343532
```

Figure 52.2. Results of Principal Component Analysis

The following statement plots the Prin data set created from the previous PROC PRINCOMP statement:

```
title2 'Plot of Principal Components';
%plotit(data=Prin, labelvar=City,
       plotvars=Prin2 Prin1, color=black, colors=blue);
run;
```

Figure 52.3 displays a plot of the second principal component Prin2 against the first principal component Prin1. It is clear from this plot that the principal components are orthogonal rotations of the original variables and that the first principal component has a larger variance than the second principal component. In fact, Prin1 has a larger variance than either of the original variables July and January.

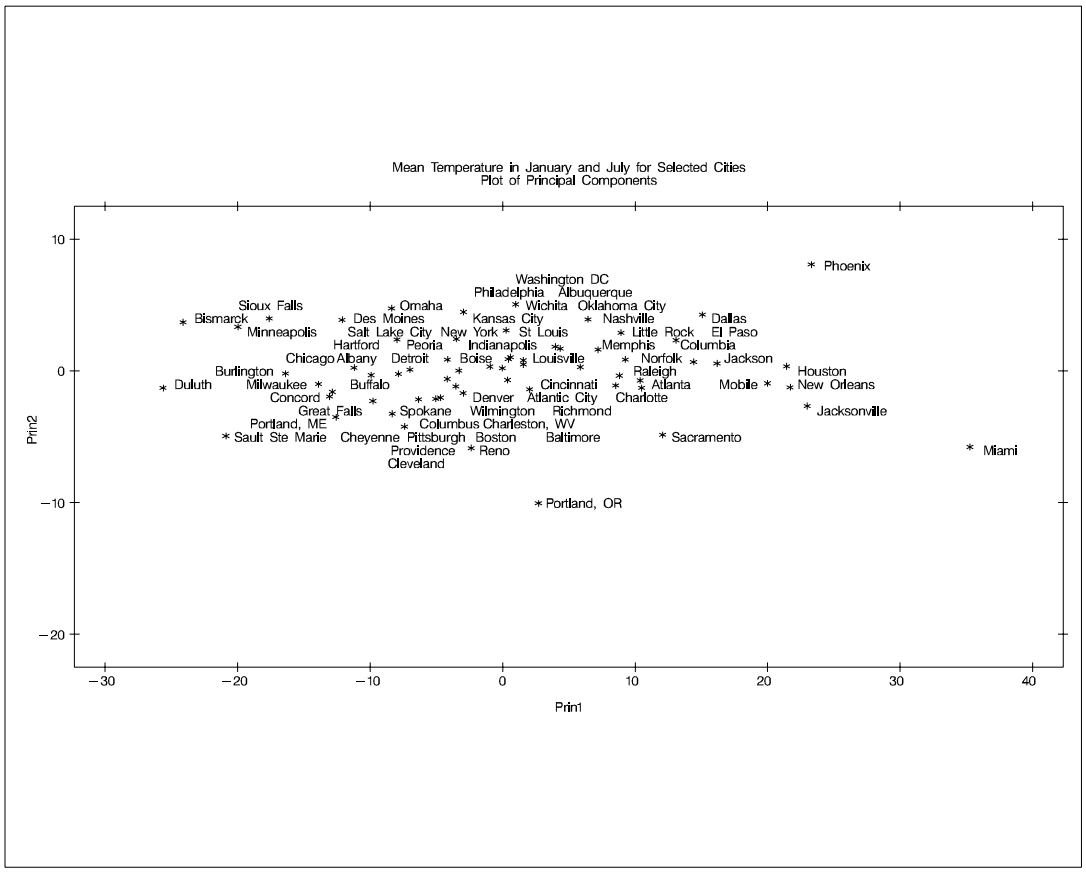


Figure 52.3. Plot of Principal Components

Syntax

The following statements are available in PROC PRINCOMP.

PROC PRINCOMP < *options* > **;**
BY *variables* **;**
FREQ *variable* **;**
PARTIAL *variables* **;**
VAR *variables* **;**
WEIGHT *variable* **;**

Usually only the VAR statement is used in addition to the PROC PRINCOMP statement. The rest of this section provides detailed syntax information for each of the preceding statements, beginning with the PROC PRINCOMP statement. The remaining statements are described in alphabetical order.

PROC PRINCOMP Statement

PROC PRINCOMP < *options* > **;**

The PROC PRINCOMP statement starts the PRINCOMP procedure and, optionally, identifies input and output data sets, specifies details of the analysis, or suppresses the display of output. You can specify the following options in the PROC PRINCOMP statement.

Task	Options
Specify data sets	DATA= OUT= OUTSTAT=
Specify details of analysis	COV N= NOINT PREFIX= SINGULAR= STD VARDEF=
Suppress the display of output	NOPRINT

The following list provides details on these options.

COVARIANCE
COV

computes the principal components from the covariance matrix. If you omit the COV option, the correlation matrix is analyzed. Use of the COV option causes variables with large variances to be more strongly associated with components with large eigenvalues and causes variables with small variances to be more strongly associated with components with small eigenvalues. You should not specify the COV option unless the units in which the variables are measured are comparable or the variables are standardized in some way. If you specify the COV option, the procedure calculates scores using the centered variables rather than the standardized variables.

DATA=*SAS-data-set*

specifies the SAS data set to be analyzed. The data set can be an ordinary SAS data set or a TYPE=ACE, TYPE=CORR, TYPE=COV, TYPE=FACTOR, TYPE=SSCP, TYPE=UCORR, or TYPE=UCOV data set (see Appendix A, "Special SAS Data Sets"). Also, the PRINCOMP procedure can read the _TYPE_='COVB' matrix from a TYPE=EST data set. If you omit the DATA= option, the procedure uses the most recently created SAS data set.

N=*number*

specifies the number of principal components to be computed. The default is the number of variables. The value of the N= option must be an integer greater than or equal to zero.

NOINT

omits the intercept from the model. In other words, the NOINT option requests that the covariance or correlation matrix not be corrected for the mean. When you use the PRINCOMP procedure with the NOINT option, the covariance matrix and, hence, the standard deviations are not corrected for the mean. If you are interested in the standard deviations corrected for the mean, you can get them by using a procedure such as the MEANS procedure.

If you use a TYPE=SSCP data set as input to the PRINCOMP procedure and list the variable Intercept in the VAR statement, the procedure acts as if you had also specified the NOINT option. If you use NOINT and also create an OUTSTAT= data set, the data set is TYPE=UCORR or TYPE=UCOV rather than TYPE=CORR or TYPE=COV.

NOPRINT

suppresses the display of all output. Note that this option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 15, "Using the Output Delivery System."

OUT=*SAS-data-set*

creates an output SAS data set that contains all the original data as well as the principal component scores. If you want to create a permanent SAS data set, you must specify a two-level name (refer to *SAS Language Reference: Concepts* for information on permanent SAS data sets).

OUTSTAT=*SAS-data-set*

creates an output SAS data set that contains means, standard deviations, number of observations, correlations or covariances, eigenvalues, and eigenvectors. If you specify the COV option, the data set is TYPE=COV or TYPE=UCOV, depending on the NOINT option, and it contains covariances; otherwise, the data set is TYPE=CORR or TYPE=UCORR, depending on the NOINT option, and it contains correlations. If you specify the PARTIAL statement, the OUTSTAT= data set contains R-squares as well. If you want to create a permanent SAS data set, you must specify a two-level name (refer to *SAS Language Reference: Concepts* for information on permanent SAS data sets).

PREFIX=*name*

specifies a prefix for naming the principal components. By default, the names are Prin1, Prin2, . . . , Prin$n$. If you specify PREFIX=ABC, the components are named ABC1, ABC2, ABC3, and so on. The number of characters in the prefix plus the number of digits required to designate the variables should not exceed the current name length defined by the VALIDVARNAME= system option.

SINGULAR=*p*
SING=*p*

specifies the singularity criterion, where $0 < p < 1$. If a variable in a PARTIAL statement has an R-square as large as $1 - p$ when predicted from the variables listed before it in the statement, the variable is assigned a standardized coefficient of 0. By default, SINGULAR=1E−8.

STANDARD
STD

standardizes the principal component scores in the OUT= data set to unit variance. If you omit the STANDARD option, the scores have variance equal to the corresponding eigenvalue. Note that STANDARD has no effect on the eigenvalues themselves.

VARDEF=DF | N | WDF | WEIGHT | WGT

specifies the divisor used in calculating variances and standard deviations. By default, VARDEF=DF. The following table displays the values and associated divisors.

Value	Divisor	Formula	
DF	error degrees of freedom	$n - i$	(before partialling)
		$n - p - i$	(after partialling)
N	number of observations	n	
WEIGHT | WGT	sum of weights	$\sum_{j=1}^{n} w_j$	
WDF	sum of weights minus one	$\left(\sum_{j=1}^{n} w_j\right) - i$	(before partialling)
		$\left(\sum_{j=1}^{n} w_j\right) - p - i$	(after partialling)

In the formulas for VARDEF=DF and VARDEF=WDF, p is the number of degrees of freedom of the variables in the PARTIAL statement, and i is 0 if the NOINT option is specified and 1 otherwise.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC PRINCOMP to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the PRINCOMP procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure.

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide.*

FREQ Statement

FREQ *variable* **;**

The FREQ statement specifies a variable that provides frequencies for each observation in the DATA= data set. Specifically, if n is the value of the FREQ variable for a given observation, then that observation is used n times.

The analysis produced using a FREQ statement reflects the expanded number of observations. The total number of observations is considered equal to the sum of the FREQ variable. You could produce the same analysis (without the FREQ statement) by first creating a new data set that contains the expanded number of observations. For example, if the value of the FREQ variable is 5 for the first observation, the first 5 observations in the new data set would be identical. Each observation in the old data set would be replicated n_j times in the new data set, where n_j is the value of the FREQ variable for that observation.

If the value of the FREQ variable is missing or is less than one, the observation is not used in the analysis. If the value is not an integer, only the integer portion is used.

PARTIAL Statement

PARTIAL *variables* **;**

If you want to analyze a partial correlation or covariance matrix, specify the names of the numeric variables to be partialled out in the PARTIAL statement. The PRINCOMP procedure computes the principal components of the residuals from the prediction of the VAR variables by the PARTIAL variables. If you request an OUT= or OUTSTAT= data set, the residual variables are named by prefixing the characters R_ to the VAR variables. Thus, the number of characters required to distinguish the VAR variables should be, at most, two characters fewer than the current name length defined by the VALIDVARNAME= system option.

VAR Statement

VAR *variables* **;**

The VAR statement lists the numeric variables to be analyzed. If you omit the VAR statement, all numeric variables not specified in other statements are analyzed. If, however, the DATA= data set is TYPE=SSCP, the default set of variables used as VAR variables does not include Intercept so that the correlation or covariance matrix is constructed correctly. If you want to analyze Intercept as a separate variable, you should specify it in the VAR statement.

WEIGHT Statement

WEIGHT *variable* **;**

If you want to use relative weights for each observation in the input data set, place the weights in a variable in the data set and specify the name in a WEIGHT statement. This is often done when the variance associated with each observation is different and the values of the weight variable are proportional to the reciprocals of the variances.

The observation is used in the analysis only if the value of the WEIGHT statement variable is nonmissing and is greater than zero.

Details

Missing Values

Observations with missing values for any variable in the VAR, PARTIAL, FREQ, or WEIGHT statement are omitted from the analysis and are given missing values for principal component scores in the OUT= data set. If a correlation, covariance, or SSCP matrix is read, it can contain missing values as long as every pair of variables has at least one nonmissing entry.

Output Data Sets

OUT= Data Set

The OUT= data set contains all the variables in the original data set plus new variables containing the principal component scores. The N= option determines the number of new variables. The names of the new variables are formed by concatenating the value given by the PREFIX= option (or Prin if PREFIX= is omitted) and the numbers 1, 2, 3, and so on. The new variables have mean 0 and variance equal to the corresponding eigenvalue, unless you specify the STANDARD option to standardize the scores to unit variance.

If you specify the COV option, the procedure calculates scores using the centered variables rather than the standardized variables.

If you use a PARTIAL statement, the OUT= data set also contains the residuals from predicting the VAR variables from the PARTIAL variables. The names of the residual variables are formed by prefixing R_ to the names of the VAR variables.

An OUT= data set cannot be created if the DATA= data set is TYPE=ACE, TYPE=CORR, TYPE=COV, TYPE=EST, TYPE=FACTOR, TYPE=SSCP, TYPE=UCORR, or TYPE=UCOV.

OUTSTAT= Data Set

The OUTSTAT= data set is similar to the TYPE=CORR data set produced by the CORR procedure. The following table relates the TYPE= value for the OUTSTAT= data set to the options specified in the PROC PRINCOMP statement.

Options	TYPE=
(default)	CORR
COV	COV
NOINT	UCORR
COV NOINT	UCOV

Notice that the default (neither the COV nor NOINT option) produces a TYPE=CORR data set.

The new data set contains the following variables:

- the BY variables, if any
- two new variables, _TYPE_ and _NAME_, both character variables
- the variables analyzed, that is, those in the VAR statement; or, if there is no VAR statement, all numeric variables not listed in any other statement; or, if there is a PARTIAL statement, the residual variables as described under the OUT= data set

Each observation in the new data set contains some type of statistic as indicated by the _TYPE_ variable. The values of the _TYPE_ variable are as follows:

TYPE	**Contents**
MEAN	mean of each variable. If you specify the PARTIAL statement, this observation is omitted.
STD	standard deviations. If you specify the COV option, this observation is omitted, so the SCORE procedure does not standardize the variables before computing scores. If you use the PARTIAL statement, the standard deviation of a variable is computed as its root mean squared error as predicted from the PARTIAL variables.
USTD	uncorrected standard deviations. When you specify the NOINT option in the PROC PRINCOMP statement, the OUTSTAT= data set contains standard deviations not corrected for the mean. However, if you also specify the COV option in the PROC PRINCOMP statement, this observation is omitted.
N	number of observations on which the analysis is based. This value is the same for each variable. If you specify the PARTIAL statement and the value of the VARDEF= option is DF or unspecified, then the number of observations is decremented by the degrees of freedom for the PARTIAL variables.
SUMWGT	the sum of the weights of the observations. This value is the same for each variable. If you specify the PARTIAL statement and VARDEF=WDF, then the sum of the weights is decremented by the degrees of freedom for the PARTIAL variables. This observation is output only if the value is different from that in the observation with _TYPE_='N'.

CORR correlations between each variable and the variable specified by the _NAME_ variable. The number of observations with _TYPE_='CORR' is equal to the number of variables being analyzed. If you specify the COV option, no _TYPE_='CORR' observations are produced. If you use the PARTIAL statement, the partial correlations, not the raw correlations, are output.

UCORR uncorrected correlation matrix. When you specify the NOINT option without the COV option in the PROC PRINCOMP statement, the OUTSTAT= data set contains a matrix of correlations not corrected for the means. However, if you also specify the COV option in the PROC PRINCOMP statement, this observation is omitted.

COV covariances between each variable and the variable specified by the _NAME_ variable. _TYPE_='COV' observations are produced only if you specify the COV option. If you use the PARTIAL statement, the partial covariances, not the raw covariances, are output.

UCOV uncorrected covariance matrix. When you specify the NOINT and COV options in the PROC PRINCOMP statement, the OUTSTAT= data set contains a matrix of covariances not corrected for the means.

EIGENVAL eigenvalues. If the N= option requested fewer than the maximum number of principal components, only the specified number of eigenvalues are produced, with missing values filling out the observation.

SCORE eigenvectors. The _NAME_ variable contains the name of the corresponding principal component as constructed from the PREFIX= option. The number of observations with _TYPE_='SCORE' equals the number of principal components computed. The eigenvectors have unit length unless you specify the STD option, in which case the unit-length eigenvectors are divided by the square roots of the eigenvalues to produce scores with unit standard deviations.

USCORE scoring coefficients to be applied without subtracting the mean from the raw variables. _TYPE_='USCORE' observations are produced when you specify the NOINT option in the PROC PRINCOMP statement.

RSQUARED R-squares for each VAR variable as predicted by the PARTIAL variables

B regression coefficients for each VAR variable as predicted by the PARTIAL variables. This observation is produced only if you specify the COV option.

STB standardized regression coefficients for each VAR variable as predicted by the PARTIAL variables. If you specify the COV option, this observation is omitted.

The data set can be used with the SCORE procedure to compute principal component scores, or it can be used as input to the FACTOR procedure specifying METHOD=SCORE to rotate the components. If you use the PARTIAL statement, the scoring coefficients should be applied to the residuals, not the original variables.

Computational Resources

Let

n = number of observations
v = number of VAR variables
p = number of PARTIAL variables
c = number of components

- The minimum allocated memory required is

$$232v + 120p + 48c + \max(8cv, 8vp + 4(v + p)(v + p + 1))$$

  bytes
- The time required to compute the correlation matrix is roughly proportional to

$$n(v + p)^2 + \frac{p}{2}(v + p)(v + p + 1)$$

- The time required to compute eigenvalues is roughly proportional to v^3.
- The time required to compute eigenvectors is roughly proportional to cv^2.

Displayed Output

The PRINCOMP procedure displays the following items if the DATA= data set is not TYPE=CORR, TYPE=COV, TYPE=SSCP, TYPE=UCORR, or TYPE=UCOV:

- Simple Statistics, including the Mean and Std (standard deviation) for each variable. If you specify the NOINT option, the uncorrected standard deviation (UStD) is displayed.
- the Correlation or, if you specify the COV option, the Covariance Matrix

The PRINCOMP procedure displays the following items if you use the PARTIAL statement.

- Regression Statistics, giving the R-square and RMSE (root mean square error) for each VAR variable as predicted by the PARTIAL variables (not shown)
- Standardized Regression Coefficients or, if you specify the COV option, Regression Coefficients for predicting the VAR variables from the PARTIAL variables (not shown)
- the Partial Correlation Matrix or, if you specify the COV option, the Partial Covariance Matrix (not shown)

The PRINCOMP procedure displays the following item if you specify the COV option:

- the Total Variance

The PRINCOMP procedure displays the following items unless you specify the NOPRINT option:

- Eigenvalues of the correlation or covariance matrix, as well as the Difference between successive eigenvalues, the Proportion of variance explained by each eigenvalue, and the Cumulative proportion of variance explained
- the Eigenvectors

ODS Table Names

PROC PRINCOMP assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table.

For more information on ODS, see Chapter 15, "Using the Output Delivery System."

Table 52.1. ODS Tables Produced in PROC PRINCOMP

ODS Table Name	Description	Statement / Option
NObsNVar	Number of Observations, Variables and (Partial) Variables	default
SimpleStatistics	Simple Statistics	default
Corr	Correlation Matrix	default unless COV is specified
Cov	Covariance Matrix	default if COV is specified
RSquareRMSE	Regression Statistics: R-Squares and RMSEs	PARTIAL statement
RegCoef	Regression Coefficients	PARTIAL statement COV
StdRegCoef	Standardized Regression Coefficients	PARTIAL statement
ParCorr	Partial Correlation Matrix	PARTIAL statement
ParCov	Uncorrected Partial Covariance Matrix	PARTIAL statement COV
TotalVariance	Total Variance	PROC PRINCOMP COV
Eigenvalues	Eigenvalues	default
Eigenvectors	Eigenvectors	default

Examples

Example 52.1. Crime Rates

The following data provide crime rates per 100,000 people in seven categories for each of the fifty states in 1977. Since there are seven numeric variables, it is impossible to plot all the variables simultaneously. Principal components can be used to summarize the data in two or three dimensions, and they help to visualize the data. The following statements produce Output 52.1.1:

```
data Crime;
   title 'Crime Rates per 100,000 Population by State';
   input State $1-15 Murder Rape Robbery Assault
         Burglary Larceny Auto_Theft;
   datalines;
Alabama        14.2 25.2  96.8 278.3 1135.5 1881.9 280.7
Alaska         10.8 51.6  96.8 284.0 1331.7 3369.8 753.3
Arizona         9.5 34.2 138.2 312.3 2346.1 4467.4 439.5
Arkansas        8.8 27.6  83.2 203.4  972.6 1862.1 183.4
California     11.5 49.4 287.0 358.0 2139.4 3499.8 663.5
Colorado        6.3 42.0 170.7 292.9 1935.2 3903.2 477.1
Connecticut     4.2 16.8 129.5 131.8 1346.0 2620.7 593.2
Delaware        6.0 24.9 157.0 194.2 1682.6 3678.4 467.0
Florida        10.2 39.6 187.9 449.1 1859.9 3840.5 351.4
Georgia        11.7 31.1 140.5 256.5 1351.1 2170.2 297.9
Hawaii          7.2 25.5 128.0  64.1 1911.5 3920.4 489.4
Idaho           5.5 19.4  39.6 172.5 1050.8 2599.6 237.6
Illinois        9.9 21.8 211.3 209.0 1085.0 2828.5 528.6
Indiana         7.4 26.5 123.2 153.5 1086.2 2498.7 377.4
Iowa            2.3 10.6  41.2  89.8  812.5 2685.1 219.9
Kansas          6.6 22.0 100.7 180.5 1270.4 2739.3 244.3
Kentucky       10.1 19.1  81.1 123.3  872.2 1662.1 245.4
Louisiana      15.5 30.9 142.9 335.5 1165.5 2469.9 337.7
Maine           2.4 13.5  38.7 170.0 1253.1 2350.7 246.9
Maryland        8.0 34.8 292.1 358.9 1400.0 3177.7 428.5
Massachusetts   3.1 20.8 169.1 231.6 1532.2 2311.3 1140.1
Michigan        9.3 38.9 261.9 274.6 1522.7 3159.0 545.5
Minnesota       2.7 19.5  85.9  85.8 1134.7 2559.3 343.1
Mississippi    14.3 19.6  65.7 189.1  915.6 1239.9 144.4
Missouri        9.6 28.3 189.0 233.5 1318.3 2424.2 378.4
Montana         5.4 16.7  39.2 156.8  804.9 2773.2 309.2
Nebraska        3.9 18.1  64.7 112.7  760.0 2316.1 249.1
Nevada         15.8 49.1 323.1 355.0 2453.1 4212.6 559.2
New Hampshire   3.2 10.7  23.2  76.0 1041.7 2343.9 293.4
New Jersey      5.6 21.0 180.4 185.1 1435.8 2774.5 511.5
New Mexico      8.8 39.1 109.6 343.4 1418.7 3008.6 259.5
New York       10.7 29.4 472.6 319.1 1728.0 2782.0 745.8
North Carolina 10.6 17.0  61.3 318.3 1154.1 2037.8 192.1
North Dakota    0.9  9.0  13.3  43.8  446.1 1843.0 144.7
Ohio            7.8 27.3 190.5 181.1 1216.0 2696.8 400.4
Oklahoma        8.6 29.2  73.8 205.0 1288.2 2228.1 326.8
Oregon          4.9 39.9 124.1 286.9 1636.4 3506.1 388.9
```

```
Pennsylvania    5.6 19.0 130.3 128.0  877.5 1624.1 333.2
Rhode Island    3.6 10.5  86.5 201.0 1489.5 2844.1 791.4
South Carolina 11.9 33.0 105.9 485.3 1613.6 2342.4 245.1
South Dakota    2.0 13.5  17.9 155.7  570.5 1704.4 147.5
Tennessee      10.1 29.7 145.8 203.9 1259.7 1776.5 314.0
Texas          13.3 33.8 152.4 208.2 1603.1 2988.7 397.6
Utah            3.5 20.3  68.8 147.3 1171.6 3004.6 334.5
Vermont         1.4 15.9  30.8 101.2 1348.2 2201.0 265.2
Virginia        9.0 23.3  92.1 165.7  986.2 2521.2 226.7
Washington      4.3 39.6 106.2 224.8 1605.6 3386.9 360.3
West Virginia   6.0 13.2  42.2  90.9  597.4 1341.7 163.3
Wisconsin       2.8 12.9  52.2  63.7  846.9 2614.2 220.7
Wyoming         5.4 21.9  39.7 173.9  811.6 2772.2 282.0
;

proc princomp out=Crime_Components;
run;
```

Output 52.1.1. Results of Principal Component Analysis: PROC PRINCOMP

Crime Rates per 100,000 Population by State

The PRINCOMP Procedure

Observations	50
Variables	7

Simple Statistics

	Murder	Rape	Robbery	Assault
Mean	7.444000000	25.73400000	124.0920000	211.3000000
StD	3.866768941	10.75962995	88.3485672	100.2530492

Simple Statistics

	Burglary	Larceny	Auto_Theft
Mean	1291.904000	2671.288000	377.5260000
StD	432.455711	725.908707	193.3944175

Correlation Matrix

	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto_ Theft
Murder	1.0000	0.6012	0.4837	0.6486	0.3858	0.1019	0.0688
Rape	0.6012	1.0000	0.5919	0.7403	0.7121	0.6140	0.3489
Robbery	0.4837	0.5919	1.0000	0.5571	0.6372	0.4467	0.5907
Assault	0.6486	0.7403	0.5571	1.0000	0.6229	0.4044	0.2758
Burglary	0.3858	0.7121	0.6372	0.6229	1.0000	0.7921	0.5580
Larceny	0.1019	0.6140	0.4467	0.4044	0.7921	1.0000	0.4442
Auto_Theft	0.0688	0.3489	0.5907	0.2758	0.5580	0.4442	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.11495951	2.87623768	0.5879	0.5879
2	1.23872183	0.51290521	0.1770	0.7648
3	0.72581663	0.40938458	0.1037	0.8685
4	0.31643205	0.05845759	0.0452	0.9137
5	0.25797446	0.03593499	0.0369	0.9506
6	0.22203947	0.09798342	0.0317	0.9823
7	0.12405606		0.0177	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
Murder	0.300279	-.629174	0.178245	-.232114	0.538123	0.259117	0.267593
Rape	0.431759	-.169435	-.244198	0.062216	0.188471	-.773271	-.296485
Robbery	0.396875	0.042247	0.495861	-.557989	-.519977	-.114385	-.003903
Assault	0.396652	-.343528	-.069510	0.629804	-.506651	0.172363	0.191745
Burglary	0.440157	0.203341	-.209895	-.057555	0.101033	0.535987	-.648117
Larceny	0.357360	0.402319	-.539231	-.234890	0.030099	0.039406	0.601690
Auto_Theft	0.295177	0.502421	0.568384	0.419238	0.369753	-.057298	0.147046

The eigenvalues indicate that two or three components provide a good summary of the data, two components accounting for 76 percent of the total variance and three components explaining 87 percent. Subsequent components contribute less than 5 percent each.

The first component is a measure of overall crime rate since the first eigenvector shows approximately equal loadings on all variables. The second eigenvector has high positive loadings on the variables Auto_Theft and Larceny and high negative loadings on the variables Murder and Assault. There is also a small positive loading on Burglary and a small negative loading on Rape. This component seems to measure the preponderance of property crime over violent crime. The interpretation of the third component is not obvious.

A simple way to examine the principal components in more detail is to display the output data set sorted by each of the large components. The following statements produce Output 52.1.2 through Output 52.1.3:

```
proc sort;
   by Prin1;
run;

proc print;
   id State;
   var Prin1 Prin2 Murder Rape Robbery
       Assault Burglary Larceny Auto_Theft;
   title2 'States Listed in Order of Overall Crime Rate';
   title3 'As Determined by the First Principal Component';
run;

proc sort;
   by Prin2;
run;

proc print;
   id State;
   var Prin1 Prin2 Murder Rape Robbery
       Assault Burglary Larceny Auto_Theft;
   title2 'States Listed in Order of Property Vs.
           Violent Crime';
   title3 'As Determined by the Second Principal Component';
run;
```

Output 52.1.2. OUT= Data Set Sorted by First Principal Component

Crime Rates per 100,000 Population by State
States Listed in Order of Overall Crime Rate
As Determined by the First Principal Component

State	Prin1	Prin2	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto_Theft
North Dakota	-3.96408	0.38767	0.9	9.0	13.3	43.8	446.1	1843.0	144.7
South Dakota	-3.17203	-0.25446	2.0	13.5	17.9	155.7	570.5	1704.4	147.5
West Virginia	-3.14772	-0.81425	6.0	13.2	42.2	90.9	597.4	1341.7	163.3
Iowa	-2.58156	0.82475	2.3	10.6	41.2	89.8	812.5	2685.1	219.9
Wisconsin	-2.50296	0.78083	2.8	12.9	52.2	63.7	846.9	2614.2	220.7
New Hampshire	-2.46562	0.82503	3.2	10.7	23.2	76.0	1041.7	2343.9	293.4
Nebraska	-2.15071	0.22574	3.9	18.1	64.7	112.7	760.0	2316.1	249.1
Vermont	-2.06433	0.94497	1.4	15.9	30.8	101.2	1348.2	2201.0	265.2
Maine	-1.82631	0.57878	2.4	13.5	38.7	170.0	1253.1	2350.7	246.9
Kentucky	-1.72691	-1.14663	10.1	19.1	81.1	123.3	872.2	1662.1	245.4
Pennsylvania	-1.72007	-0.19590	5.6	19.0	130.3	128.0	877.5	1624.1	333.2
Montana	-1.66801	0.27099	5.4	16.7	39.2	156.8	804.9	2773.2	309.2
Minnesota	-1.55434	1.05644	2.7	19.5	85.9	85.8	1134.7	2559.3	343.1
Mississippi	-1.50736	-2.54671	14.3	19.6	65.7	189.1	915.6	1239.9	144.4
Idaho	-1.43245	-0.00801	5.5	19.4	39.6	172.5	1050.8	2599.6	237.6
Wyoming	-1.42463	0.06268	5.4	21.9	39.7	173.9	811.6	2772.2	282.0
Arkansas	-1.05441	-1.34544	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
Utah	-1.04996	0.93656	3.5	20.3	68.8	147.3	1171.6	3004.6	334.5
Virginia	-0.91621	-0.69265	9.0	23.3	92.1	165.7	986.2	2521.2	226.7
North Carolina	-0.69925	-1.67027	10.6	17.0	61.3	318.3	1154.1	2037.8	192.1
Kansas	-0.63407	-0.02804	6.6	22.0	100.7	180.5	1270.4	2739.3	244.3
Connecticut	-0.54133	1.50123	4.2	16.8	129.5	131.8	1346.0	2620.7	593.2
Indiana	-0.49990	0.00003	7.4	26.5	123.2	153.5	1086.2	2498.7	377.4
Oklahoma	-0.32136	-0.62429	8.6	29.2	73.8	205.0	1288.2	2228.1	326.8
Rhode Island	-0.20156	2.14658	3.6	10.5	86.5	201.0	1489.5	2844.1	791.4
Tennessee	-0.13660	-1.13498	10.1	29.7	145.8	203.9	1259.7	1776.5	314.0
Alabama	-0.04988	-2.09610	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
New Jersey	0.21787	0.96421	5.6	21.0	180.4	185.1	1435.8	2774.5	511.5
Ohio	0.23953	0.09053	7.8	27.3	190.5	181.1	1216.0	2696.8	400.4
Georgia	0.49041	-1.38079	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9
Illinois	0.51290	0.09423	9.9	21.8	211.3	209.0	1085.0	2828.5	528.6
Missouri	0.55637	-0.55851	9.6	28.3	189.0	233.5	1318.3	2424.2	378.4
Hawaii	0.82313	1.82392	7.2	25.5	128.0	64.1	1911.5	3920.4	489.4
Washington	0.93058	0.73776	4.3	39.6	106.2	224.8	1605.6	3386.9	360.3
Delaware	0.96458	1.29674	6.0	24.9	157.0	194.2	1682.6	3678.4	467.0
Massachusetts	0.97844	2.63105	3.1	20.8	169.1	231.6	1532.2	2311.3	1140.1
Louisiana	1.12020	-2.08327	15.5	30.9	142.9	335.5	1165.5	2469.9	337.7
New Mexico	1.21417	-0.95076	8.8	39.1	109.6	343.4	1418.7	3008.6	259.5
Texas	1.39696	-0.68131	13.3	33.8	152.4	208.2	1603.1	2988.7	397.6
Oregon	1.44900	0.58603	4.9	39.9	124.1	286.9	1636.4	3506.1	388.9
South Carolina	1.60336	-2.16211	11.9	33.0	105.9	485.3	1613.6	2342.4	245.1
Maryland	2.18280	-0.19474	8.0	34.8	292.1	358.9	1400.0	3177.7	428.5
Michigan	2.27333	0.15487	9.3	38.9	261.9	274.6	1522.7	3159.0	545.5
Alaska	2.42151	0.16652	10.8	51.6	96.8	284.0	1331.7	3369.8	753.3
Colorado	2.50929	0.91660	6.3	42.0	170.7	292.9	1935.2	3903.2	477.1
Arizona	3.01414	0.84495	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
Florida	3.11175	-0.60392	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4
New York	3.45248	0.43289	10.7	29.4	472.6	319.1	1728.0	2782.0	745.8
California	4.28380	0.14319	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5
Nevada	5.26699	-0.25262	15.8	49.1	323.1	355.0	2453.1	4212.6	559.2

Output 52.1.3. OUT= Data Set Sorted by Second Principal Component

Crime Rates per 100,000 Population by State
States Listed in Order of Property Vs. Violent Crime
As Determined by the Second Principal Component

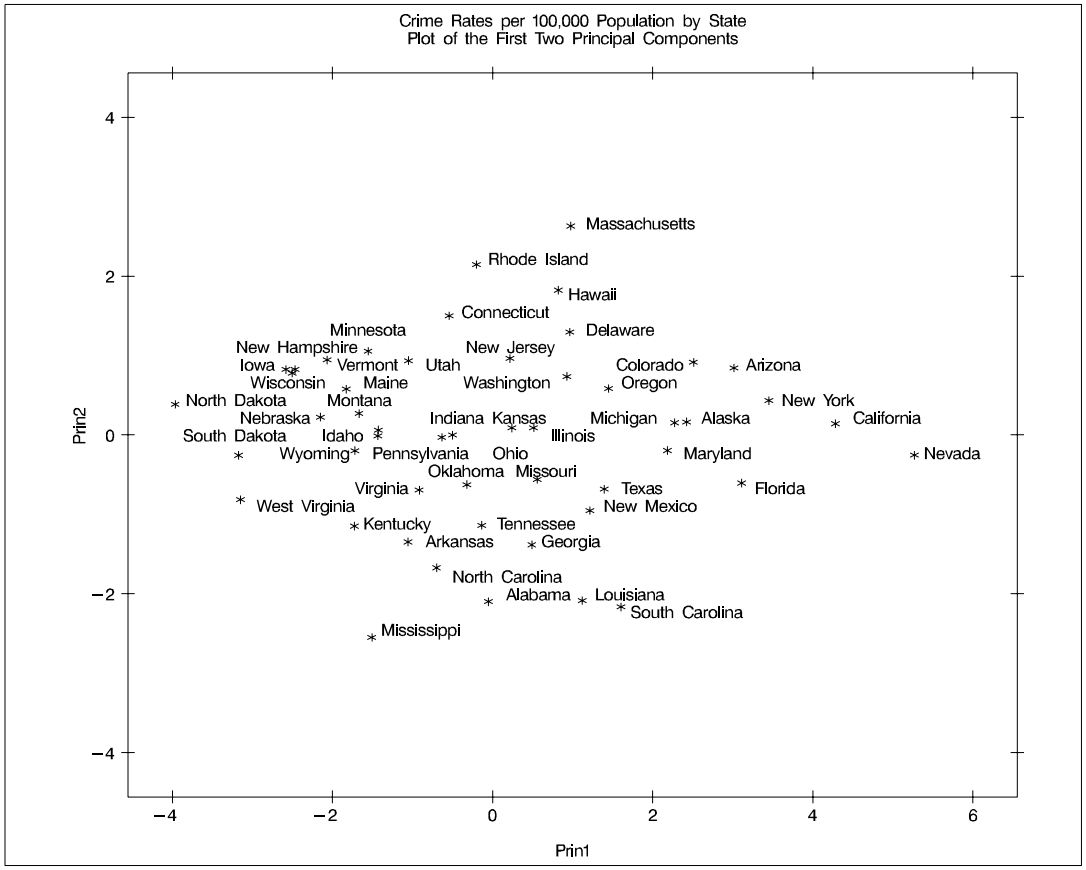
State	Prin1	Prin2	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto_Theft
Mississippi	-1.50736	-2.54671	14.3	19.6	65.7	189.1	915.6	1239.9	144.4
South Carolina	1.60336	-2.16211	11.9	33.0	105.9	485.3	1613.6	2342.4	245.1
Alabama	-0.04988	-2.09610	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
Louisiana	1.12020	-2.08327	15.5	30.9	142.9	335.5	1165.5	2469.9	337.7
North Carolina	-0.69925	-1.67027	10.6	17.0	61.3	318.3	1154.1	2037.8	192.1
Georgia	0.49041	-1.38079	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9
Arkansas	-1.05441	-1.34544	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
Kentucky	-1.72691	-1.14663	10.1	19.1	81.1	123.3	872.2	1662.1	245.4
Tennessee	-0.13660	-1.13498	10.1	29.7	145.8	203.9	1259.7	1776.5	314.0
New Mexico	1.21417	-0.95076	8.8	39.1	109.6	343.4	1418.7	3008.6	259.5
West Virginia	-3.14772	-0.81425	6.0	13.2	42.2	90.9	597.4	1341.7	163.3
Virginia	-0.91621	-0.69265	9.0	23.3	92.1	165.7	986.2	2521.2	226.7
Texas	1.39696	-0.68131	13.3	33.8	152.4	208.2	1603.1	2988.7	397.6
Oklahoma	-0.32136	-0.62429	8.6	29.2	73.8	205.0	1288.2	2228.1	326.8
Florida	3.11175	-0.60392	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4
Missouri	0.55637	-0.55851	9.6	28.3	189.0	233.5	1318.3	2424.2	378.4
South Dakota	-3.17203	-0.25446	2.0	13.5	17.9	155.7	570.5	1704.4	147.5
Nevada	5.26699	-0.25262	15.8	49.1	323.1	355.0	2453.1	4212.6	559.2
Pennsylvania	-1.72007	-0.19590	5.6	19.0	130.3	128.0	877.5	1624.1	333.2
Maryland	2.18280	-0.19474	8.0	34.8	292.1	358.9	1400.0	3177.7	428.5
Kansas	-0.63407	-0.02804	6.6	22.0	100.7	180.5	1270.4	2739.3	244.3
Idaho	-1.43245	-0.00801	5.5	19.4	39.6	172.5	1050.8	2599.6	237.6
Indiana	-0.49990	0.00003	7.4	26.5	123.2	153.5	1086.2	2498.7	377.4
Wyoming	-1.42463	0.06268	5.4	21.9	39.7	173.9	811.6	2772.2	282.0
Ohio	0.23953	0.09053	7.8	27.3	190.5	181.1	1216.0	2696.8	400.4
Illinois	0.51290	0.09423	9.9	21.8	211.3	209.0	1085.0	2828.5	528.6
California	4.28380	0.14319	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5
Michigan	2.27333	0.15487	9.3	38.9	261.9	274.6	1522.7	3159.0	545.5
Alaska	2.42151	0.16652	10.8	51.6	96.8	284.0	1331.7	3369.8	753.3
Nebraska	-2.15071	0.22574	3.9	18.1	64.7	112.7	760.0	2316.1	249.1
Montana	-1.66801	0.27099	5.4	16.7	39.2	156.8	804.9	2773.2	309.2
North Dakota	-3.96408	0.38767	0.9	9.0	13.3	43.8	446.1	1843.0	144.7
New York	3.45248	0.43289	10.7	29.4	472.6	319.1	1728.0	2782.0	745.8
Maine	-1.82631	0.57878	2.4	13.5	38.7	170.0	1253.1	2350.7	246.9
Oregon	1.44900	0.58603	4.9	39.9	124.1	286.9	1636.4	3506.1	388.9
Washington	0.93058	0.73776	4.3	39.6	106.2	224.8	1605.6	3386.9	360.3
Wisconsin	-2.50296	0.78083	2.8	12.9	52.2	63.7	846.9	2614.2	220.7
Iowa	-2.58156	0.82475	2.3	10.6	41.2	89.8	812.5	2685.1	219.9
New Hampshire	-2.46562	0.82503	3.2	10.7	23.2	76.0	1041.7	2343.9	293.4
Arizona	3.01414	0.84495	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
Colorado	2.50929	0.91660	6.3	42.0	170.7	292.9	1935.2	3903.2	477.1
Utah	-1.04996	0.93656	3.5	20.3	68.8	147.3	1171.6	3004.6	334.5
Vermont	-2.06433	0.94497	1.4	15.9	30.8	101.2	1348.2	2201.0	265.2
New Jersey	0.21787	0.96421	5.6	21.0	180.4	185.1	1435.8	2774.5	511.5
Minnesota	-1.55434	1.05644	2.7	19.5	85.9	85.8	1134.7	2559.3	343.1
Delaware	0.96458	1.29674	6.0	24.9	157.0	194.2	1682.6	3678.4	467.0
Connecticut	-0.54133	1.50123	4.2	16.8	129.5	131.8	1346.0	2620.7	593.2
Hawaii	0.82313	1.82392	7.2	25.5	128.0	64.1	1911.5	3920.4	489.4
Rhode Island	-0.20156	2.14658	3.6	10.5	86.5	201.0	1489.5	2844.1	791.4
Massachusetts	0.97844	2.63105	3.1	20.8	169.1	231.6	1532.2	2311.3	1140.1

Another recommended procedure is to make scatter plots of the first few components. The sorted listings help to identify observations on the plots. The following statements produce Output 52.1.4 through Output 52.1.5:

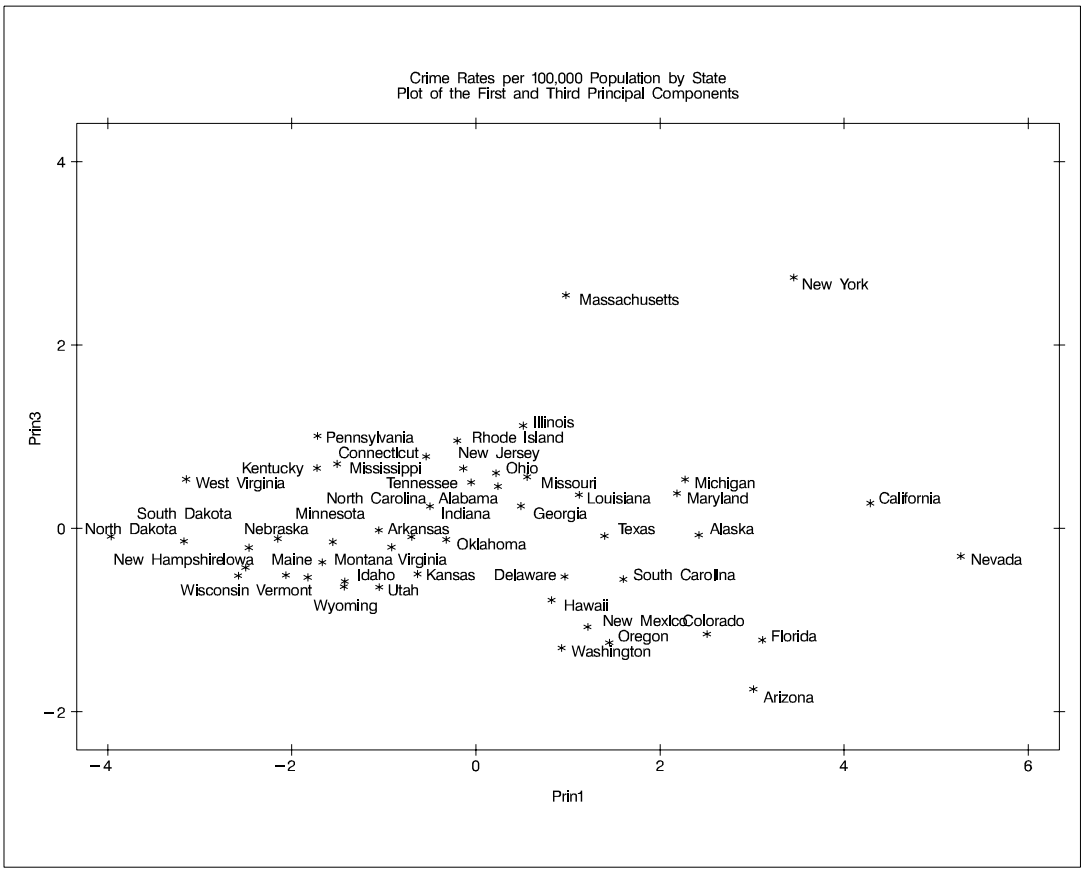
```
title2 'Plot of the First Two Principal Components';
%plotit(data=Crime_Components, labelvar=State,
        plotvars=Prin2 Prin1, color=black, colors=blue);
run;

title2 'Plot of the First and Third Principal Components';
%plotit(data=Crime_Components, labelvar=State,
        plotvars=Prin3 Prin1, color=black, colors=blue);
run;
```

Output 52.1.4. Plot of the First Two Principal Components



Output 52.1.5. Plot of the First and Third Principal Components



It is possible to identify regional trends on the plot of the first two components. Nevada and California are at the extreme right, with high overall crime rates but an average ratio of property crime to violent crime. North and South Dakota are on the extreme left with low overall crime rates. Southeastern states tend to be in the bottom of the plot, with a higher-than-average ratio of violent crime to property crime. New England states tend to be in the upper part of the plot, with a greater-than-average ratio of property crime to violent crime.

The most striking feature of the plot of the first and third principal components is that Massachusetts and New York are outliers on the third component.

Example 52.2. Basketball Data

The data in this example are rankings of 35 college basketball teams. The rankings were made before the start of the 1985–86 season by 10 news services.

The purpose of the principal component analysis is to compute a single variable that best summarizes all 10 of the preseason rankings.

Note that the various news services rank different numbers of teams, varying from 20 through 30 (there is a missing rank in one of the variables, WashPost). And, of course, each service does not rank the same teams, so there are missing values in these data. Each of the 35 teams is ranked by at least one news service.

The PRINCOMP procedure omits observations with missing values. To obtain principal component scores for all of the teams, it is necessary to replace the missing values. Since it is the best teams that are ranked, it is not appropriate to replace missing values with the mean of the nonmissing values. Instead, an ad hoc method is used that replaces missing values by the mean of the unassigned ranks. For example, if 20 teams are ranked by a news service, then ranks 21 through 35 are unassigned. The mean of ranks 21 through 35 is 28, so missing values for that variable are replaced by the value 28. To prevent the method of missing-value replacement from having an undue effect on the analysis, each observation is weighted according to the number of nonmissing values it has. See Example 53.2 in Chapter 53, “The PRINQUAL Procedure,” for an alternative analysis of these data.

Since the first principal component accounts for 78 percent of the variance, there is substantial agreement among the rankings. The eigenvector shows that all the news services are about equally weighted, so a simple average would work almost as well as the first principal component. The following statements produce Output 52.2.1 through Output 52.2.3:

```
/*-------------------------------------------------------------*/
/*                                                             */
/* Preseason 1985 College Basketball Rankings                  */
/* (rankings of 35 teams by 10 news services)                  */
/*                                                             */
/* Note: (a) news services rank varying numbers of teams;      */
/*       (b) not all teams are ranked by all news services;    */
/*       (c) each team is ranked by at least one service;      */
/*       (d) rank 20 is missing for UPI.                       */
/*                                                             */
/*-------------------------------------------------------------*/
title1 'Preseason 1985 College Basketball Rankings';
data HoopsRanks;
   input School $13. CSN DurSun DurHer WashPost USAToday
         Sport InSports UPI AP SI;
   label CSN      = 'Community Sports News (Chapel Hill, NC)'
         DurSun   = 'Durham Sun'
         DurHer   = 'Durham Morning Herald'
         WashPost = 'Washington Post'
         USAToday = 'USA Today'
         Sport    = 'Sport Magazine'
```

```
          InSports = 'Inside Sports'
          UPI      = 'United Press International'
          AP       = 'Associated Press'
          SI       = 'Sports Illustrated'
          ;
   format CSN--SI 5.1;
   datalines;
Louisville     1  8  1  9  8  9  6 10  9  9
Georgia Tech   2  2  4  3  1  1  1  2  1  1
Kansas         3  4  5  1  5 11  8  4  5  7
Michigan       4  5  9  4  2  5  3  1  3  2
Duke           5  6  7  5  4 10  4  5  6  5
UNC            6  1  2  2  3  4  2  3  2  3
Syracuse       7 10  6 11  6  6  5  6  4 10
Notre Dame     8 14 15 13 11 20 18 13 12  .
Kentucky       9 15 16 14 14 19 11 12 11 13
LSU           10  9 13  . 13 15 16  9 14  8
DePaul        11  . 21 15 20  . 19  .  . 19
Georgetown    12  7  8  6  9  2  9  8  8  4
Navy          13 20 23 10 18 13 15  . 20  .
Illinois      14  3  3  7  7  3 10  7  7  6
Iowa          15 16  .  . 23  .  . 14  . 20
Arkansas      16  .  .  . 25  .  .  .  . 16
Memphis State 17  . 11  . 16  8 20  . 15 12
Washington    18  .  .  .  .  .  . 17  .  .
UAB           19 13 10  . 12 17  . 16 16 15
UNLV          20 18 18 19 22  . 14 18 18  .
NC State      21 17 14 16 15  . 12 15 17 18
Maryland      22  .  .  . 19  .  .  . 19 14
Pittsburgh    23  .  .  .  .  .  .  .  .  .
Oklahoma      24 19 17 17 17 12 17  . 13 17
Indiana       25 12 20 18 21  .  .  .  .  .
Virginia      26  . 22  .  . 18  .  .  .  .
Old Dominion  27  .  .  .  .  .  .  .  .  .
Auburn        28 11 12  8 10  7  7 11 10 11
St. Johns     29  .  .  .  . 14  .  .  .  .
UCLA          30  .  .  .  .  .  . 19  .  .
St. Joseph's   .  . 19  .  .  .  .  .  .  .
Tennessee      .  . 24  .  . 16  .  .  .  .
Montana        .  .  . 20  .  .  .  .  .  .
Houston        .  .  .  . 24  .  .  .  .  .
Virginia Tech  .  .  .  .  .  . 13  .  .  .
;

/* PROC MEANS is used to output a data set containing the   */
/* maximum value of each of the newspaper and magazine      */
/* rankings.  The output data set, maxrank, is then used    */
/* to set the missing values to the next highest rank plus  */
/* thirty-six, divided by two (that is, the mean of the     */
/* missing ranks).  This ad hoc method of replacing missing */
/* values is based more on intuition than on rigorous       */
/* statistical theory.  Observations are weighted by the    */
/* number of nonmissing values.                             */
```

```
proc means data=HoopsRanks;
   output out=MaxRank
          max=CSNMax DurSunMax DurHerMax
              WashPostMax USATodayMax SportMax
              InSportsMax UPIMax APMax SIMax;
run;




/* The following method of filling in missing values is a   */
/* reasonable method for this specific example.  It would   */
/* be inappropriate to use this method for other data sets. */
/* sets.  In addition, any method of filling in missing     */
/* values can result in incorrect statistics. The choice    */
/* of whether to fill in missing values, and what method    */
/* to use to do so, is the responsibility of the person     */
/* performing the analysis.                                 */

data Basketball;
   set HoopsRanks;
   if _n_=1 then set MaxRank;
   array Services{10} CSN--SI;
   array MaxRanks{10} CSNMax--SIMax;
   keep School CSN--SI Weight;
   Weight=0;
   do i=1 to 10;
      if Services{i}=. then Services{i}=(MaxRanks{i}+36)/2;
      else Weight=Weight+1;
   end;
run;


/* Use the PRINCOMP procedure to transform the observed */
/* ranks. Use n=1 because the data should be related to */
/* a single underlying variable. Sort the data and      */
/* display the resulting component.                     */          */


proc princomp data=Basketball n=1 out=PCBasketball
              standard;
   var CSN--SI;
   weight Weight;
run;

proc sort data=PCBasketball;
   by Prin1;
run;

proc print;
   var School Prin1;
   title2 'College Teams as Ordered by PROC PRINCOMP';
run;
```

Output 52.2.1. Summary Statistics for Basketball Rankings Using PROC MEANS

Pre-Season 1985 College Basketball Rankings

The MEANS Procedure

Variable	Label	N	Mean
CSN	Community Sports News (Chapel Hill, NC)	30	15.5000000
DurSun	Durham Sun	20	10.5000000
DurHer	Durham Morning Herald	24	12.5000000
WashPost	Washington Post	19	10.4210526
USAToday	USA Today	25	13.0000000
Sport	Sport Magazine	20	10.5000000
InSports	Inside Sports	20	10.5000000
UPI	United Press International	19	10.0000000
AP	Associated Press	20	10.5000000
SI	Sports Illustrated	20	10.5000000

Variable	Label	Std Dev	Minimum
CSN	Community Sports News (Chapel Hill, NC)	8.8034084	1.0000000
DurSun	Durham Sun	5.9160798	1.0000000
DurHer	Durham Morning Herald	7.0710678	1.0000000
WashPost	Washington Post	6.0673607	1.0000000
USAToday	USA Today	7.3598007	1.0000000
Sport	Sport Magazine	5.9160798	1.0000000
InSports	Inside Sports	5.9160798	1.0000000
UPI	United Press International	5.6273143	1.0000000
AP	Associated Press	5.9160798	1.0000000
SI	Sports Illustrated	5.9160798	1.0000000

Variable	Label	Maximum
CSN	Community Sports News (Chapel Hill, NC)	30.0000000
DurSun	Durham Sun	20.0000000
DurHer	Durham Morning Herald	24.0000000
WashPost	Washington Post	20.0000000
USAToday	USA Today	25.0000000
Sport	Sport Magazine	20.0000000
InSports	Inside Sports	20.0000000
UPI	United Press International	19.0000000
AP	Associated Press	20.0000000
SI	Sports Illustrated	20.0000000

Output 52.2.2. Principal Components Analysis of Basketball Rankings Using PROC PRINCOMP

The PRINCOMP Procedure

Observations	35
Variables	10

Simple Statistics

	CSN	DurSun	DurHer	WashPost	USAToday
Mean	13.33640553	13.06451613	12.88018433	13.83410138	12.55760369
StD	22.08036285	21.66394183	21.38091837	23.47841791	20.48207965

Simple Statistics

	Sport	InSports	UPI	AP	SI
Mean	13.83870968	13.24423963	13.59216590	12.83410138	13.52534562
StD	23.37756267	22.20231526	23.25602811	21.40782406	22.93219584

The PRINCOMP Procedure

Correlation Matrix

		CSN	DurSun	DurHer
CSN	Community Sports News (Chapel Hill, NC)	1.0000	0.6505	0.6415
DurSun	Durham Sun	0.6505	1.0000	0.8341
DurHer	Durham Morning Herald	0.6415	0.8341	1.0000
WashPost	Washington Post	0.6121	0.7667	0.7035
USAToday	USA Today	0.7456	0.8860	0.8877
Sport	Sport Magazine	0.4806	0.6940	0.7788
InSports	Inside Sports	0.6558	0.7702	0.7900
UPI	United Press International	0.7007	0.9015	0.7676
AP	Associated Press	0.6779	0.8437	0.8788
SI	Sports Illustrated	0.6135	0.7518	0.7761

Correlation Matrix

	WashPost	USAToday	Sport	InSports	UPI	AP	SI
CSN	0.6121	0.7456	0.4806	0.6558	0.7007	0.6779	0.6135
DurSun	0.7667	0.8860	0.6940	0.7702	0.9015	0.8437	0.7518
DurHer	0.7035	0.8877	0.7788	0.7900	0.7676	0.8788	0.7761
WashPost	1.0000	0.7984	0.6598	0.8717	0.6953	0.7809	0.5952
USAToday	0.7984	1.0000	0.7716	0.8475	0.8539	0.9479	0.8426
Sport	0.6598	0.7716	1.0000	0.7176	0.6220	0.8217	0.7701
InSports	0.8717	0.8475	0.7176	1.0000	0.7920	0.8830	0.7332
UPI	0.6953	0.8539	0.6220	0.7920	1.0000	0.8436	0.7738
AP	0.7809	0.9479	0.8217	0.8830	0.8436	1.0000	0.8212
SI	0.5952	0.8426	0.7701	0.7332	0.7738	0.8212	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	7.88601647		0.7886	0.7886

Eigenvectors

		Prin1
CSN	Community Sports News (Chapel Hill, NC)	0.270205
DurSun	Durham Sun	0.326048
DurHer	Durham Morning Herald	0.324392
WashPost	Washington Post	0.300449
USAToday	USA Today	0.345200
Sport	Sport Magazine	0.293881
InSports	Inside Sports	0.324088
UPI	United Press International	0.319902
AP	Associated Press	0.342151
SI	Sports Illustrated	0.308570

Output 52.2.3. Basketball Rankings Using PROC PRINCOMP

Pre-Season 1985 College Basketball Rankings
College Teams as Ordered by PROC PRINCOMP

Obs	School	Prin1
1	Georgia Tech	-0.58068
2	UNC	-0.53317
3	Michigan	-0.47874
4	Kansas	-0.40285
5	Duke	-0.38464
6	Illinois	-0.33586
7	Syracuse	-0.31578
8	Louisville	-0.31489
9	Georgetown	-0.29735
10	Auburn	-0.09785
11	Kentucky	0.00843
12	LSU	0.00872
13	Notre Dame	0.09407
14	NC State	0.19404
15	UAB	0.19771
16	Oklahoma	0.23864
17	Memphis State	0.25319
18	Navy	0.28921
19	UNLV	0.35103
20	DePaul	0.43770
21	Iowa	0.50213
22	Indiana	0.51713
23	Maryland	0.55910
24	Arkansas	0.62977
25	Virginia	0.67586
26	Washington	0.67756
27	Tennessee	0.70822
28	St. Johns	0.71425
29	Virginia Tech	0.71638
30	St. Joseph’s	0.73492
31	UCLA	0.73965
32	Pittsburgh	0.75078
33	Houston	0.75534
34	Montana	0.75790
35	Old Dominion	0.76821

References

Cooley, W.W. and Lohnes, P.R. (1971), *Multivariate Data Analysis*, New York: John Wiley & Sons, Inc.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations*, New York: John Wiley & Sons, Inc.

Hotelling, H. (1933), "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, 24, 417–441, 498–520.

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

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis*, London: Academic Press.

Morrison, D.F. (1976), *Multivariate Statistical Methods*, Second Edition, New York: McGraw-Hill Book Co.

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

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

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999