THE POWER TO KNOW.

SAS/STAT® 9.22 User's Guide

The PRINCOMP Procedure

(Book Excerpt)

SAS® Documentation

This document is an individual chapter from *SAS/STAT® 9.22 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2010. *SAS/STAT® 9.22 User's Guide*. Cary, NC: SAS Institute Inc.

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

1st electronic book, May 2010

Chapter 70

The PRINCOMP Procedure

Contents

Overview: PRINCOMP Procedure

The PRINCOMP procedure performs principal component analysis. As input you can use raw data, a correlation matrix, a covariance matrix, or a sum-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 using principal component analysis depends in part on your research objectives. You should use the PRINCOMP procedure if you are interested in summarizing data and detecting linear relationships. You can use principal components to reduce the number of variables in regression, clustering, and so on. See Chapter 9, "Introduction to Multivariate Procedures," for a detailed comparison of the PRINCOMP and FACTOR procedures.

You can use ODS Graphics to display the scree plot, component pattern plot, component pattern profile plot, matrix plot of component scores, and component score plots. These plots are especially valuable tools in exploratory data analysis.

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, Frobenius) 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: PRINCOMP Procedure

The following data provide crime rates per 100,000 people in seven categories for each of the 50 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 Figure 70.1 through Figure 70.5.

```
data Crime;
   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
;

ods graphics on;
title 'Crime Rates per 100,000 Population by State';
proc princomp out=Crime_Components plots= score(ellipse ncomp=3);
   id State;
run;
ods graphics off;
```

Figure 70.1 displays the PROC PRINCOMP output, beginning with simple statistics followed by the correlation matrix. The PROC PRINCOMP statement requests by default principal components computed from the correlation matrix, so the total variance is equal to the number of variables, 7.

Figure 70.1 Number of Observations and Simple Statistics from the PRINCOMP Procedure

```
                Crime Rates per 100,000 Population by State

                          The PRINCOMP Procedure

                         Observations         50
                         Variables             7
```

Figure 70.1 *continued*

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

Figure 70.2 displays the eigenvalues. The first principal component explains about 58.8% of the total variance, the second principal component explains about 17.7%, and the third principal component explains about 10.4%. Note that the eigenvalues sum to the total variance.

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

Figure 70.2 Results of Principal Component Analysis: PROC PRINCOMP

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

Figure 70.3 displays the eigenvectors. From the eigenvectors matrix, you can represent the first principal component Prin1 as a linear combination of the original variables:

$$\begin{aligned} \text{Prin1} = {} & 0.300279 \times (\text{Murder}) \\ & + 0.431759 \times (\text{Rape}) \\ & + 0.396875 \times (\text{Robbery}) \\ & . \\ & . \\ & . \\ & + 0.295177 \times (\text{Auto\_Theft}) \end{aligned}$$

Similarly, the second principal component Prin2 is

$$\begin{aligned} \text{Prin2} = {} & -0.629174 \times (\text{Murder}) \\ & - 0.169435 \times (\text{Rape}) \\ & + 0.042247 \times (\text{Robbery}) \\ & . \\ & . \\ & . \\ & - 0.502421 \times (\text{Auto\_Theft}) \end{aligned}$$

where the variables are standardized.

Figure 70.3 Results of Principal Component Analysis: PROC PRINCOMP

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 first component is a measure of the overall crime rate since the first eigenvector shows approximately equal loadings on all variables. The second eigenvector has high positive loadings on variables Auto_Theft and Larceny and high negative loadings on 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.

The ODS GRAPHICS statement enables the PRINCOMP procedure to produce statistical graphs by using ODS Graphics. See Chapter 21, "Statistical Graphics Using ODS," for more information. PLOTS=SCORE(ELLIPSE NCOMP=3) in the PROC PRINCOMP statement requests the pairwise component score plots for the first three components with a 95% prediction ellipse overlaid on each of the scatter plot. Figure 70.4 shows the plot of the first two 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 at the extreme left, with low overall crime rates. Southeastern states tend to be at 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 higherr-than-average ratio of property crime to violent crime. Assuming the first two components are from a bivariate normal distribution, the ellipse identifies Nevada as a possible outlier.

Figure 70.4 Plot of the First Two Component Scores

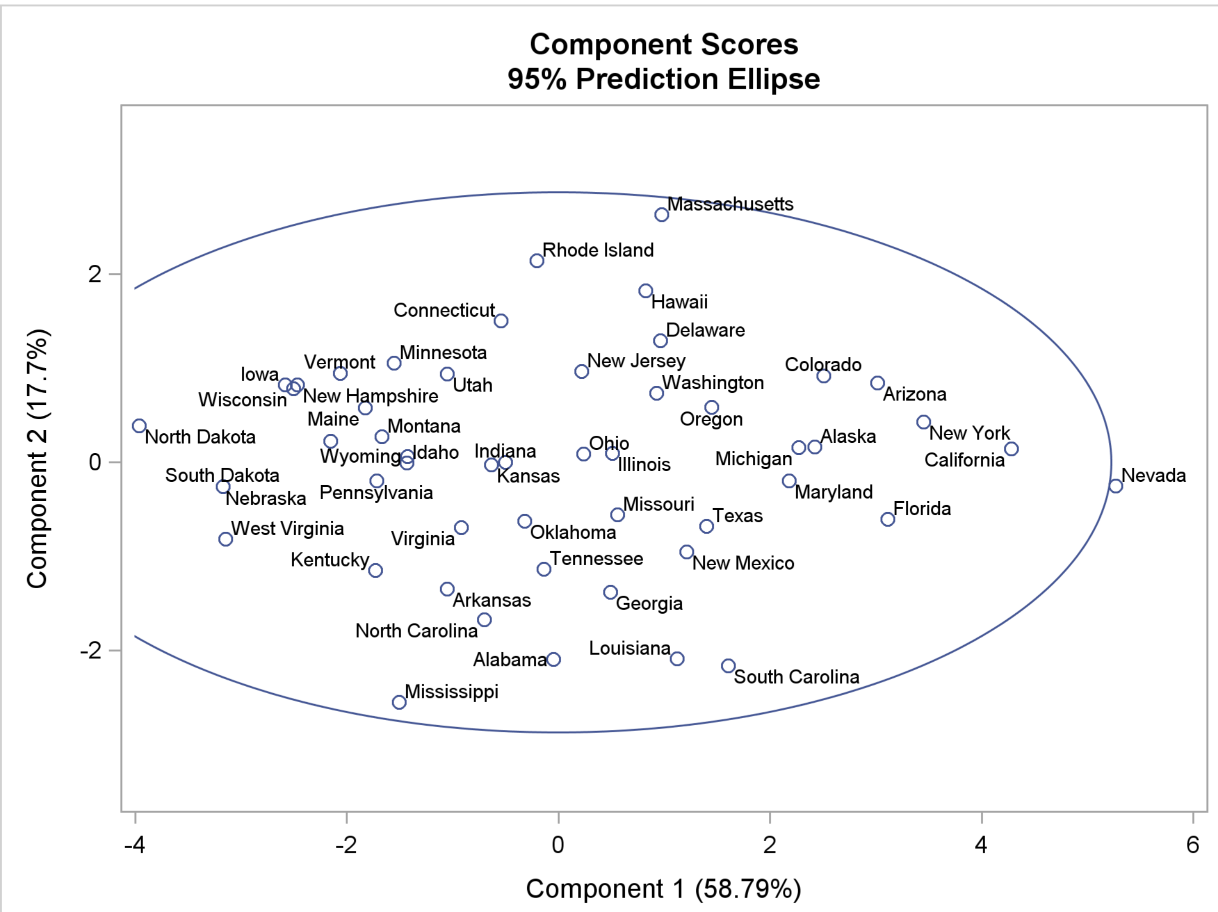
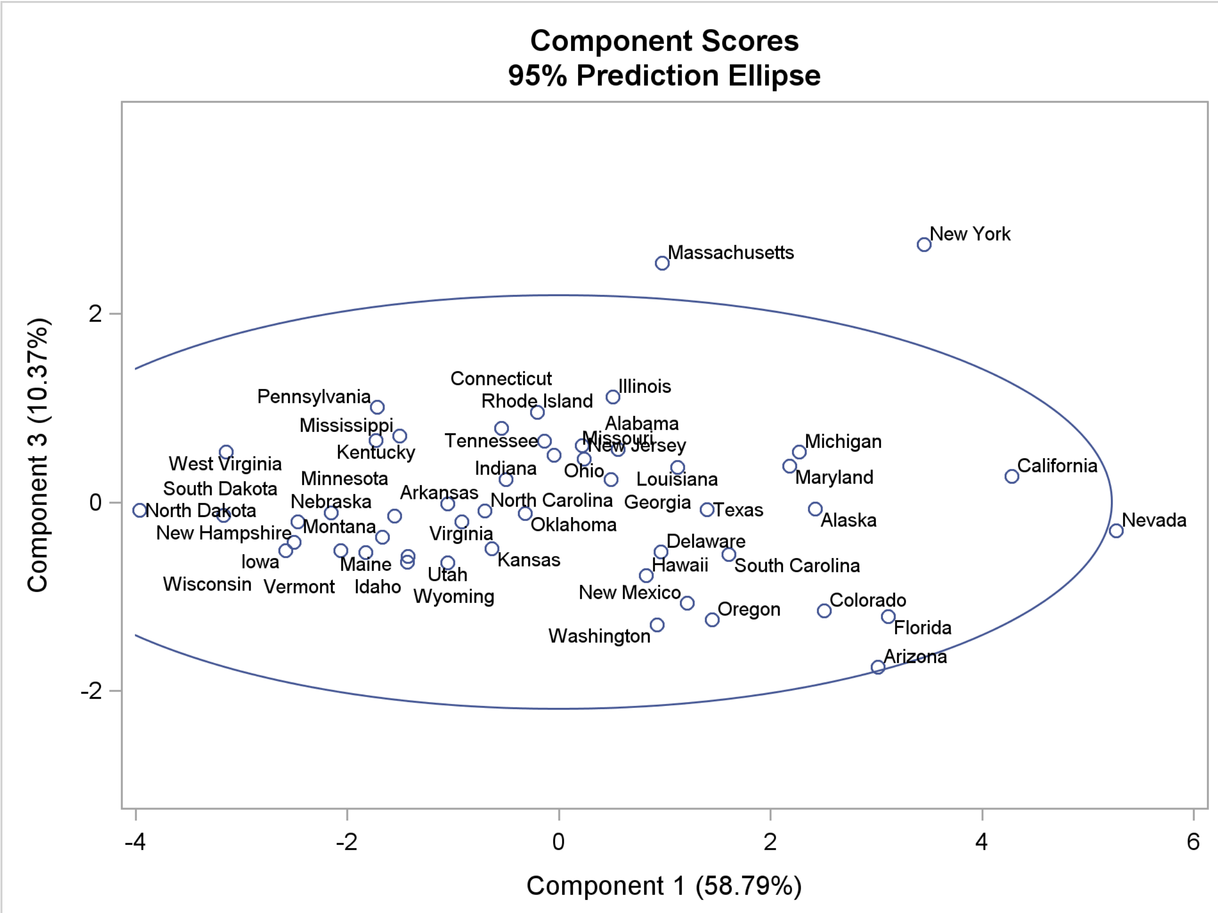


Figure 70.5 shows the plot of the first and third components. Assuming the first and the third components are from a bivariate normal distribution, the ellipse identifies Nevada, Massachusetts, and New York as possible outliers.

Figure 70.5 Plot of the First and Third Component Scores



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.

Syntax: PRINCOMP Procedure

The following statements are available in PROC PRINCOMP:

> **PROC PRINCOMP** < *options* > ;
> **BY** *variables* ;
> **FREQ** *variable* ;
> **ID** *variables* ;
> **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 the analyses performed, and controls displayed output. Table 70.1 summarizes the options.

Table 70.1 Summary of PROC PRINCOMP Statement Options

Option	Description
Specify data sets	
DATA=	specifies input data set name
OUT=	specifies output data set name
OUTSTAT=	specifies output data set name containing various statistics
Specify details of analysis	
COV	computes the principal components from the covariance matrix
N=	specifies the number of principal components to be computed
NOINT	omits the intercept from the model
PREFIX=	specifies a prefix for naming the principal components
RPREFIX=	specifies a prefix for naming the residual variables
SINGULAR=	specifies the singularity criterion
STD	standardizes the principal component scores
VARDEF=	specifies the divisor used in calculating variances and standard deviations
Suppress the display of output	
NOPRINT	suppresses the display of all output
Specify ODS Graphics details	
PLOTS=	specifies options that control the details of the plots

The following list provides details about 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.

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 20, "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. For details about OUT= data sets, see the section "Output Data Sets" on page 5869. See *SAS Language Reference: Concepts* for more information about 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. For details about OUTSTAT= data sets, see the section "Output Data Sets" on page 5869. See *SAS Language Reference: Concepts* for more information about permanent SAS data sets.

PLOTS < **(***global-plot-options***)** > < **=** *plot-request* < *(options)* > >

PLOTS < **(***global-plot-options***)** > < **=** **(***plot-request* < *(options)* > < **...** *plot-request* < *(options)* > > **)** >

controls the plots produced through ODS Graphics. When you specify only one plot request, you can omit the parentheses around the plot request. Here are some examples:

```
plots=none
plots=(scatter pattern)
plots(unpack)=scree
plots(ncomp=3 flip)=(pattern(circles=0.5 1.0) score)
```

You must enable ODS Graphics before requesting plots—for example, like this:

```
ods graphics on;
proc princomp plots=all;
   var x1--x10;
run;
ods graphics off;
```

For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS." If you have enabled ODS Graphics but do not specify the PLOTS= option, PROC PRINCOMP produces the scree plot by default.

The global plot options include the following:

FLIP

flips or interchanges the X-axis and Y-axis dimension for the component score plots and the component pattern plots. For example, if there are three components, the default plots ($y * x$) are Component 2 * Component 1, Component 3 * Component 1, and Component 3 * Component 2. When you specify PLOTS(FLIP), the plots are Component 1 * Component 2, Component 1 * Component 3, and Component 2 * Component 3.

NCOMP=*n*

specifies the number of components $n(\geq 2)$ to be plotted for the component pattern plots and the component score plots. If the NCOMP= option is again specified in an individual plot, such as PLOTS=SCORE(NCOMP= m), the value m will determine the number of components to be plotted in the component score plots. Be aware that the number of plots ($\frac{n\times(n-1)}{2}$) produced grows quadratically when n increases. The default is 5 or the total number of components $m(\geq 2)$, whichever is smaller. If $n > m$, NCOMP= m will be used.

ONLY

suppresses the default plots. Only plots specifically requested are displayed.

UNPACKPANEL

UNPACK

suppresses paneling in the scree plot. By default, multiple plots can appear in an output panel. Specify UNPACKPANEL to get each plot in a separate panel. You can specify PLOTS(UNPACKPANEL) to unpack the default plots. You can also specify UNPACK-PANEL as a suboption with SCREE (such as PLOTS=SCREE(UNPACKPANEL)).

The plot requests include the following:

ALL
produces all appropriate plots. You can specify other options with ALL; for example, to request all plots and unpack only the scree plot, specify PLOTS=(ALL SCREE(UNPACKPANEL)).

EIGEN | EIGENVALUE | SCREE < (UNPACKPANEL) >
produces the scree plot of eigenvalues and proportion variance explained. By default, both plots are output in a panel. Specify PLOTS= SCREE(UNPACKPANEL) to get each plot in a separate panel.

MATRIX
produces the matrix plot of principal component scores.

NONE
suppresses the display of all graphics output.

PATTERN < (*pattern-options* **) >**
produces the pairwise component pattern plots. Each variable is plotted as an observation whose coordinates are correlations between the variable and the two corresponding components on the plot. Use the NCOMP= option (for instance, PLOTS=PATTERN(NCOMP=3)) described in the following to control the number of plots to be displayed.

The available *pattern-options* are as follows:

CIRCLES < = *number list* **>**
plots the variance percentage circles. Each number in the list must be greater than 0. If the number is greater than or equal to 1, it is interpreted as a percentage and divided by 100; CIRCLES=0.05 and CIRCLES=5 are equivalent. For each number (c) specified, a ($c \times 100\%$) variance circle is created.

By default, there is no circle for the scatter pattern plot (PLOTS=PATTERN) and a unit circle with a 100% variance circle is plotted for the vector pattern plot (PLOTS=PATTERN (VECTOR)). You can display multiple circles by specifying PLOTS=PATTERN(CIRCLES=). For example, specifying PLOTS=PATTERN(CIRCLES= .3 .6 1.0) will display the 30%, 60%, and 100% variance circles in the pattern plots.

FLIP
flips or interchanges the X-axis and Y-axis dimensions for the component pattern plots. Specify PLOTS=PATTERN(FLIP) to flip the X-axis and Y-axis dimensions.

NCOMP=*n*
specifies the number of components $n (\geq 2)$ to be plotted. The default is 5 or the total number of components $m (\geq 2)$, whichever is smaller. If $n > m$, NCOMP= m will be used. Be aware that the number of plots ($\frac{n \times (n-1)}{2}$) produced grows quadratically when n increases.

VECTOR

plots pattern in a vector form.

PATTERNPROFILE | PROFILE

produces the pattern profile plot. There is a profile for each component. The Y-axis value represents the correlation between the variable (corresponding to the X-axis value) and the profiled principal component.

SCORE < (*score-options*) >

produces the pairwise component score plots. Use the NCOMP= option (for instance, PLOTS=SCORE(NCOMP=3)) described in the following to control the number of plots to be displayed.

The available *score-options* are as follows:

ALPHA=*number list*

specifies a list of numbers for the prediction ellipses to be displayed in the score plots. Each value (α) in the list must be greater than 0. If α is greater than or equal to 1, it is interpreted as a percentage and divided by 100; ALPHA=0.05 and ALPHA=5 are equivalent.

ELLIPSE

requests prediction ellipses for the principal component scores of a new observation to be created in the principal component score plots. See the section "Confidence and Prediction Ellipses" in "The CORR Procedure" (*Base SAS Procedures Guide: Statistical Procedures*), for details about the computation of a prediction ellipse.

FLIP

flips or interchanges the X-axis and Y-axis dimensions for the component score plots. Specify PLOTS=SCORE(FLIP) to flip the X-axis and Y-axis dimensions.

NCOMP=*n*

specifies the number of components $n(\geq 2)$ to be plotted. The default is 5 or the total number of components $m(\geq 2)$, whichever is smaller. If $n > m$, NCOMP= m will be used. Be aware that the number of plots ($\frac{n \times (n-1)}{2}$) produced grows quadratically when n increases.

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.

PARPREFIX=*name*

PPREFIX=*name*

RPREFIX=*name*

specifies a prefix for naming the residual variables in the OUT= data set and the OUTSTAT= data set. By default, the prefix is R_. The number of characters in the prefix plus the maximum length of the variable names 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 partialing)
		$n - p - i$	(after partialing)
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 partialing)
		$\left(\sum_{j=1}^{n} w_j\right) - p - i$	(after partialing)

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 that are 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 you specify more than one BY statement, only the last one specified is used.

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

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option 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 by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base 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.

ID Statement

ID *variables* ;

The ID statement labels observations with values from the first ID variable in the principal component score plot. If one or more ID variables are specified, their values are displayed in tooltips of the component score plot and the matrix plot of component scores.

PARTIAL Statement

PARTIAL *variables* ;

If you want to analyze a partial correlation or covariance matrix, specify the names of the numeric variables to be partialed 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_ by default or the string specified in the RPREFIX= option to the VAR variables.

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: PRINCOMP Procedure

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. Also, if you specify the COV option, the procedure computes the principal component scores from the corrected or the uncorrected (if the NOINT option is specified) 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.

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

Note 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:

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.

  To obtain the principal component scores, if the COV option is not specified, these coefficients should be multiplied by the standardized data. With the COV option, these coefficients should be multiplied by the centered data. Means obtained from the observation with _TYPE_='MEAN' and standard deviations obtained from the observation with _TYPE_='STD' should be used for centering and standardizing the data.

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.

  To obtain the principal component scores, these coefficients should be multiplied by the data that are standardized by the uncorrected standard deviations obtained from the observation with _TYPE_='USTD'.

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

$$\begin{aligned} n &= \text{number of observations} \\ v &= \text{number of VAR variables} \\ p &= \text{number of PARTIAL variables} \\ c &= \text{number of components} \end{aligned}$$

- The minimum allocated memory required (in bytes) is

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

- 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 squared 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 Table 70.2. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

All of the tables are created with the specification of the PROC PRINCOMP statement; a few tables need an additional PARTIAL statement.

Table 70.2 ODS Tables Produced by PROC PRINCOMP

ODS Table Name	Description	Statement / Option
Corr	Correlation matrix	default
Cov	Covariance matrix	COV
Eigenvalues	Eigenvalues	default
Eigenvectors	Eigenvectors	default
NObsNVar	Number of observations, variables, and partial variables	default
ParCorr	Partial correlation matrix	PARTIAL statement
ParCov	Uncorrected partial covariance matrix	PARTIAL statement and COV
RegCoef	Regression coefficients	PARTIAL statement and COV
RSquareRMSE	Regression statistics: R squares and RMSEs	PARTIAL statement
SimpleStatistics	Simple statistics	default
StdRegCoef	Standardized regression coefficients	PARTIAL statement
TotalVariance	Total variance	COV

ODS Graphics

To request graphics with PROC PRINCOMP, you must first enable ODS Graphics by specifying the **ods graphics on** statement. See Chapter 21, "Statistical Graphics Using ODS," for more information. Some graphs are produced by default; other graphs are produced by using statements and options. You can reference every graph produced through ODS Graphics with a name. The names of the graphs that PROC PRINCOMP generates are listed in Table 70.3, along with the required statements and options.

Table 70.3 ODS Graphics Produced by PROC PRINCOMP

ODS Graph Name	Plot Description	Statement and Option
PaintedScorePlot	Score plot of component 3 versus component 2, painted by component 1	PLOTS=SCORE when number of variables ≥ 3
PatternPlot	Component pattern plot	PLOTS=PATTERN
PatternProfilePlot	Component pattern profile plot	PLOTS=PATTERNPROFILE
ScoreMatrixPlot	Matrix plot of component scores	PLOTS=MATRIX
ScorePlot	Component score plot	PLOTS=SCORE
ScreePlot	Scree and variance plots	default and PLOTS=SCREE
VariancePlot	Variance proportion explained plot	PLOTS=SCREE(UNPACKPANEL)

Examples: PRINCOMP Procedure

Example 70.1: Temperatures

This example analyzes 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;
   length Cityid $ 2;
   title 'Mean Temperature in January and July for Selected Cities ';
   input City $1-15 January July;
   Cityid = substr(City,1,2);
   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
```

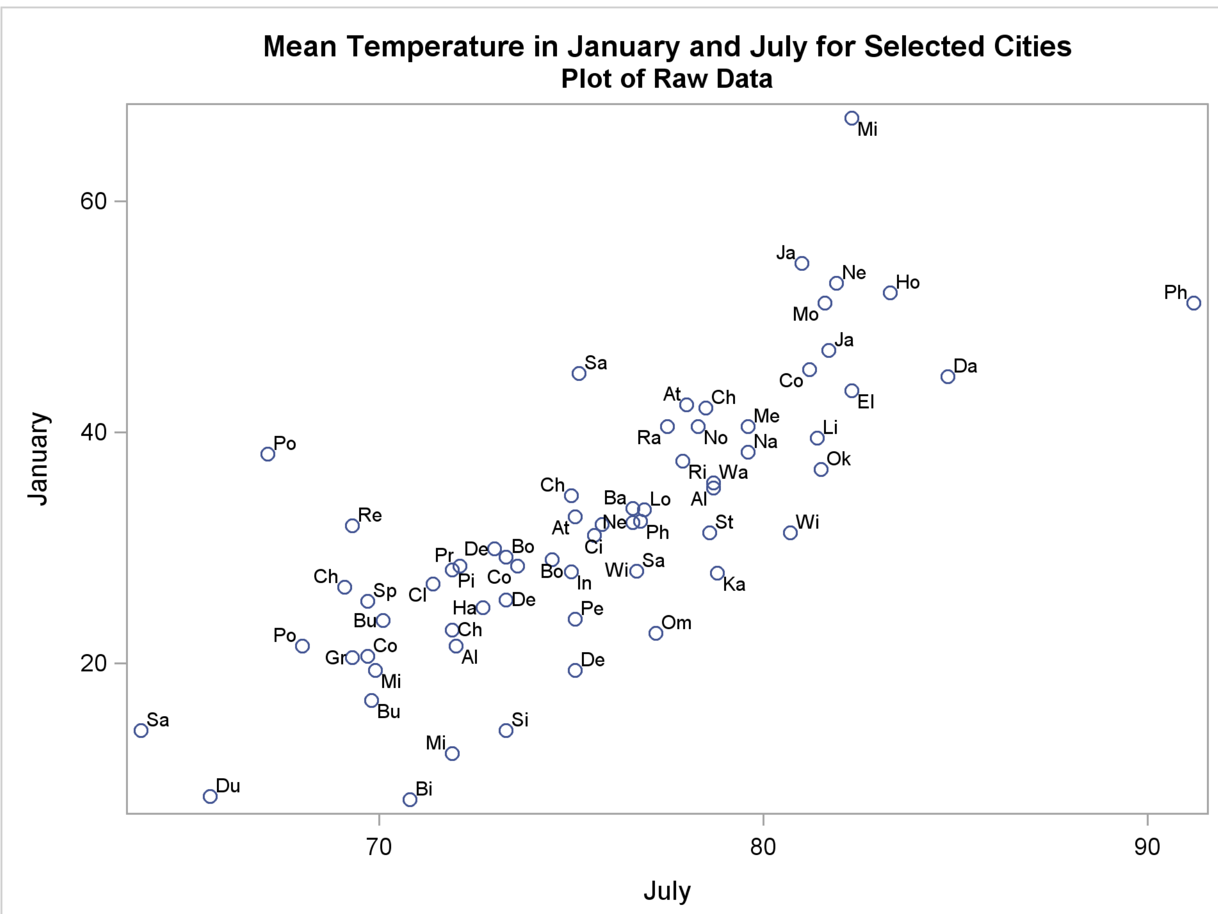
```
Richmond         37.5 77.9
Spokane          25.4 69.7
Charleston, WV   34.5 75.0
Milwaukee        19.4 69.9
Cheyenne         26.6 69.1
;
```

The following statements plot the temperature data set. The Cityid variable instead of City is used as a data label in the scatter plot for possible label clashing.

```
title 'Mean Temperature in January and July for Selected Cities';
title2 'Plot of Raw Data';
proc sgplot data=Temperature;
   scatter x=July y=January / datalabel=Cityid;
run;
```

The results are displayed in Output 70.1.1, which shows a scatter diagram of the 64 pairs of data points with July temperatures plotted against January temperatures.

Output 70.1.1 Plot of Raw Data



The following statement requests a principal component analysis on the Temperature data set:

```
ods graphics on;
title 'Mean Temperature in January and July for Selected Cities';
proc princomp data=Temperature cov plots=score(ellipse);
   var July January;
   id Cityid;
run;
ods graphics off;
```

Output 70.1.2 displays the PROC PRINCOMP output. 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% of the total variance, and the second principal component explains only about 6%. The eigenvalues sum to the total variance.

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 by using the centered variables rather than the standardized variables.

Output 70.1.2 Results of Principal Component Analysis

```
       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
```

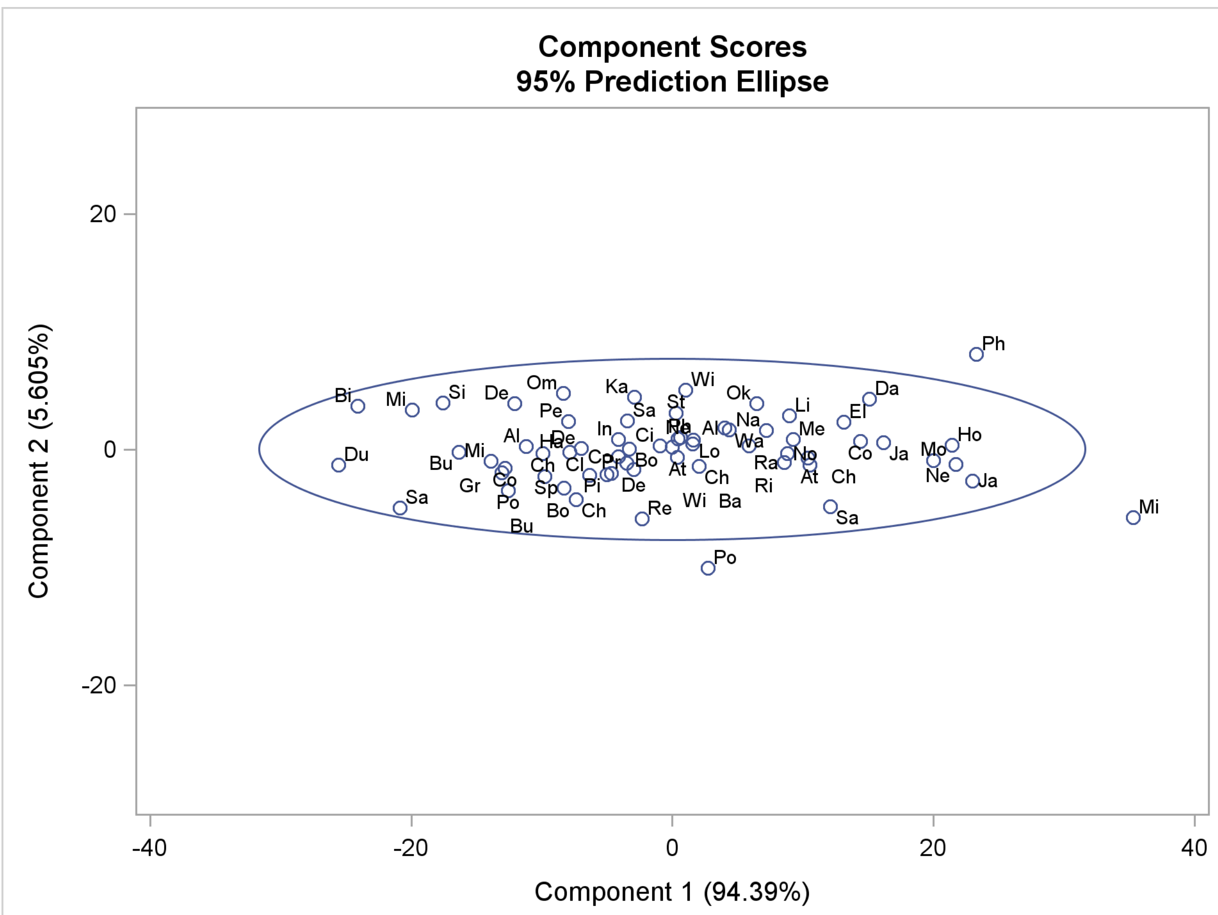
Output 70.1.2 *continued*

Eigenvectors

	Prin1	Prin2
July	0.343532	0.939141
January	0.939141	-.343532

PLOTS=SCORE in the PROC PRINCOMP statement requests a plot of the second principal component against the first principal component as shown in Output 70.1.3. 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, the first component has a larger variance than either of the original variables July and January. The ellipse indicates that Miami, Phoenix, and Portland are possible outliers.

Output 70.1.3 Plot of Component 2 by Component 1



Example 70.2: 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, not all services 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 with 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 71.2 in Chapter 71, "The PRINQUAL Procedure," for an alternative analysis of these data.

Since the first principal component accounts for 78% of the variance, there is substantial agreement among the rankings. The eigenvector shows that all the news services are about equally weighted; this is also suggested by the nearly horizontal line of the pattern profile plot in Output 70.2.3. So a simple average would work almost as well as the first principal component. The following statements produce Output 70.2.1.

```
/*-----------------------------------------------------------*/
/*                                                           */
/* Pre-season 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.                     */
/*                                                           */
/*-----------------------------------------------------------*/

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.                                 */
/*                                                              */
```

```
title 'Pre-Season 1985 College Basketball Rankings';
proc means data=HoopsRanks;
   output out=MaxRank
          max=CSNMax DurSunMax DurHerMax
              WashPostMax USATodayMax SportMax
              InSportsMax UPIMax APMax SIMax;
run;
```

Output 70.2.1 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
       ------------------------------------------------------------------
```

The following statements produce Output 70.2.2 and Output 70.2.3:

```
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;

ods graphics on;
proc princomp data=Basketball n=1 out=PCBasketball standard
              plots=patternprofile;
   var CSN--SI;
   weight Weight;
run;
ods graphics off;
```

Output 70.2.2 Principal Components Analysis of Basketball Rankings Using PROC PRINCOMP

Pre-Season 1985 College Basketball Rankings

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

Output 70.2.2 *continued*

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

	Wash Post	USAToday	Sport	In Sports	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

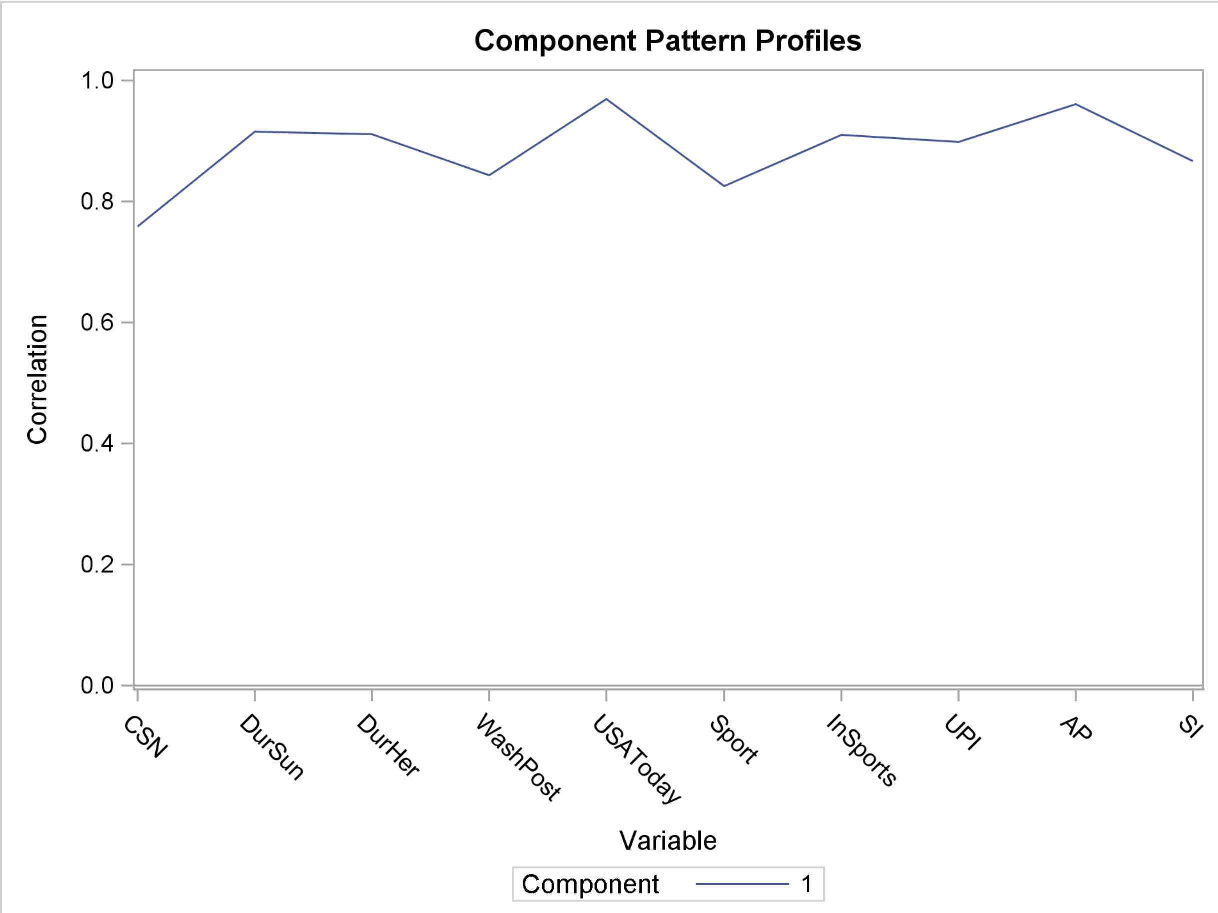
Eigenvalues of the Correlation Matrix

	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 70.2.3 Pattern Profile Plot



The following statements produce Output 70.2.4:

```
proc sort data=PCBasketball;
   by Prin1;
run;

proc print;
   var School Prin1;
   title 'Pre-Season 1985 College Basketball Rankings';
   title2 'College Teams as Ordered by PROC PRINCOMP';
run;
```

Output 70.2.4 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

Example 70.3: Job Ratings

This example uses the PRINCOMP procedure to analyze job performance. Police officers were rated by their supervisors in 14 categories as part of standard police departmental administrative procedure.

The following statements create the Jobratings data set:

```
options validvarname=any;
data Jobratings;
   input ('Communication Skills'n
          'Problem Solving'n
          'Learning Ability'n
          'Judgment Under Pressure'n
          'Observational Skills'n
          'Willingness to Confront Problems'n
          'Interest in People'n
          'Interpersonal Sensitivity'n
          'Desire for Self-Improvement'n
          'Appearance'n
          'Dependability'n
          'Physical Ability'n
          'Integrity'n
          'Overall Rating'n) (1.);
   datalines;
26838853879867
74758876857667
56757863775875
67869777988997
99997798878888
89897899888799
89999889899798
87794798468886
35652335143113
89888879576867
76557899446397
97889998898989
76766677598888

   ... more lines ...

99899899899899
76656399567486
;
```

The data set Jobratings contains 14 variables. Each variable contains the job ratings, using a scale measurement from 1 to 10 (1=fail to comply, 10=exceptional). The last variable Overall Rating contains a score as an overall index on how each officer performs.

The following statements request a principal component analysis on the Jobratings data set, output the scores to the Scores data set (OUT= Scores), and produce default plots. Note that variable Overall Rating is excluded from the analysis.

```
ods graphics on;
proc princomp data=Jobratings(drop='Overall Rating'n);
run;
```

Figure 70.3.1 and Figure 70.3.2 display the PROC PRINCOMP output, beginning with simple statistics followed by the correlation matrix. By default, the PROC PRINCOMP statement requests principal components computed from the correlation matrix, so the total variance is equal to the number of variables, 13. In this example, it would also be reasonable to use the COV option, which would cause variables with a high variance (such as Dependability) to have more influence on the results than variables with a low variance (such as Learning Ability). If you used the COV option, scores would be computed from centered rather than standardized variables.

Output 70.3.1 Simple Statistics and Correlation Matrix from the PRINCOMP Procedure

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

The PRINCOMP Procedure

Observations	103
Variables	13

Simple Statistics

	Communication Skills	Problem Solving	Learning Ability	Judgment Under Pressure	Observational Skills
Mean	6.650485437	6.631067961	6.990291262	6.737864078	6.932038835
StD	1.764068036	1.590352602	1.339411238	1.731830976	1.761584269

Simple Statistics

	Willingness to Confront Problems	Interest in People	Interpersonal Sensitivity	Desire for Self-Improvement	Appearance
Mean	7.291262136	6.708737864	6.621359223	6.572815534	7.000000000
StD	1.525155524	1.892353385	1.760773587	1.729796212	1.798692335

Simple Statistics

	Dependability	Physical Ability	Integrity
Mean	6.825242718	7.203883495	7.213592233
StD	1.917040123	1.555251845	1.845240223

Output 70.3.1 *continued*

Correlation Matrix

	Communication Skills	Problem Solving	Learning Ability	Judgment Under Pressure
Communication Skills	1.0000	0.6280	0.5546	0.5538
Problem Solving	0.6280	1.0000	0.5690	0.6195
Learning Ability	0.5546	0.5690	1.0000	0.4892
Judgment Under Pressure	0.5538	0.6195	0.4892	1.0000
Observational Skills	0.5381	0.4284	0.6230	0.3733
Willingness to Confront Problems	0.5265	0.5015	0.5245	0.4004
Interest in People	0.4391	0.3972	0.2735	0.6226
Interpersonal Sensitivity	0.5030	0.4398	0.1855	0.6134
Desire for Self-Improvement	0.5642	0.4090	0.5737	0.4826
Appearance	0.4913	0.3873	0.3988	0.2266
Dependability	0.5471	0.4546	0.5110	0.5471
Physical Ability	0.2192	0.3201	0.2269	0.3476
Integrity	0.5081	0.3846	0.3142	0.5883

Correlation Matrix

	Observational Skills	Willingness to Confront Problems	Interest in People
Communication Skills	0.5381	0.5265	0.4391
Problem Solving	0.4284	0.5015	0.3972
Learning Ability	0.6230	0.5245	0.2735
Judgment Under Pressure	0.3733	0.4004	0.6226
Observational Skills	1.0000	0.7300	0.2616
Willingness to Confront Problems	0.7300	1.0000	0.2233
Interest in People	0.2616	0.2233	1.0000
Interpersonal Sensitivity	0.1655	0.1291	0.8051
Desire for Self-Improvement	0.5985	0.5307	0.4857
Appearance	0.4177	0.4825	0.2679
Dependability	0.5626	0.4870	0.6074
Physical Ability	0.4274	0.4872	0.3768
Integrity	0.3906	0.3260	0.7452

Correlation Matrix

	Interpersonal Sensitivity	Desire for Self-Improvement	Appearance
Communication Skills	0.5030	0.5642	0.4913
Problem Solving	0.4398	0.4090	0.3873
Learning Ability	0.1855	0.5737	0.3988
Judgment Under Pressure	0.6134	0.4826	0.2266

Output 70.3.1 *continued*

Correlation Matrix

	Interpersonal Sensitivity	Desire for Self-Improvement	Appearance
Observational Skills	0.1655	0.5985	0.4177
Willingness to Confront Problems	0.1291	0.5307	0.4825
Interest in People	0.8051	0.4857	0.2679
Interpersonal Sensitivity	1.0000	0.3713	0.2600
Desire for Self-Improvement	0.3713	1.0000	0.4474
Appearance	0.2600	0.4474	1.0000
Dependability	0.5408	0.5981	0.5089
Physical Ability	0.2182	0.3752	0.3820
Integrity	0.6920	0.5664	0.4135

Correlation Matrix

	Dependability	Physical Ability	Integrity
Communication Skills	0.5471	0.2192	0.5081
Problem Solving	0.4546	0.3201	0.3846
Learning Ability	0.5110	0.2269	0.3142
Judgment Under Pressure	0.5471	0.3476	0.5883
Observational Skills	0.5626	0.4274	0.3906
Willingness to Confront Problems	0.4870	0.4872	0.3260
Interest in People	0.6074	0.3768	0.7452
Interpersonal Sensitivity	0.5408	0.2182	0.6920
Desire for Self-Improvement	0.5981	0.3752	0.5664
Appearance	0.5089	0.3820	0.4135
Dependability	1.0000	0.4461	0.6536
Physical Ability	0.4461	1.0000	0.3810
Integrity	0.6536	0.3810	1.0000

Figure 70.3.2 displays the eigenvalues. The first principal component explains about 50% of the total variance, the second principal component explains about 13.6%, and the third principal component explains about 7.7%. Note that the eigenvalues sum to the total variance. The eigenvalues indicate that three to five components provide a good summary of the data, with three components accounting for about 71.7% of the total variance and five components explaining about 82.7%. Subsequent components contribute less than 5% each.

Output 70.3.2 Eigenvalues and Eigenvectors from the PRINCOMP Procedure

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	6.54740242	4.77468744	0.5036	0.5036
2	1.77271499	0.76747933	0.1364	0.6400
3	1.00523565	0.26209665	0.0773	0.7173
4	0.74313901	0.06479499	0.0572	0.7745
5	0.67834402	0.22696368	0.0522	0.8267
6	0.45138034	0.06922167	0.0347	0.8614
7	0.38215866	0.08432613	0.0294	0.8908
8	0.29783254	0.02340663	0.0229	0.9137
9	0.27442591	0.01208809	0.0211	0.9348
10	0.26233782	0.01778332	0.0202	0.9550
11	0.24455450	0.04677622	0.0188	0.9738
12	0.19777828	0.05508241	0.0152	0.9890
13	0.14269586		0.0110	1.0000

Output 70.3.2 *continued*

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
Communication Skills	0.303548	0.052039	-.329181	-.227039
Problem Solving	0.278034	0.057046	-.400112	0.300476
Learning Ability	0.266521	0.288152	-.354591	-.020735
Judgment Under Pressure	0.294376	-.199458	-.255164	0.397306
Observational Skills	0.276641	0.366979	0.065959	0.035711
Willingness to Confront Problems	0.267580	0.392989	0.098723	0.184409
Interest in People	0.278060	-.432916	0.118113	0.046047
Interpersonal Sensitivity	0.253814	-.495662	-.064547	-.060000
Desire for Self-Improvement	0.299833	0.099077	0.061097	-.211279
Appearance	0.237358	0.190065	0.248353	-.544587
Dependability	0.319480	-.049742	0.169476	-.156070
Physical Ability	0.213868	0.097499	0.614959	0.514519
Integrity	0.298246	-.301812	0.190222	-.169062

Eigenvectors

	Prin5	Prin6	Prin7	Prin8
Communication Skills	0.181087	-.416563	0.143543	0.333846
Problem Solving	0.453604	0.096750	0.048904	0.199259
Learning Ability	-.219329	0.578388	-.114808	0.064088
Judgment Under Pressure	-.030188	0.102087	0.068204	-.591822
Observational Skills	-.325257	-.301254	-.297894	0.163484
Willingness to Confront Problems	0.038278	-.458585	-.044796	-.365684
Interest in People	-.111279	0.030870	-.011105	0.154829
Interpersonal Sensitivity	0.107807	-.170305	-.088194	0.192725
Desire for Self-Improvement	-.427477	0.105369	0.689011	0.087453
Appearance	0.568044	0.221643	0.049267	-.257497
Dependability	-.130575	0.202301	-.594850	0.081242
Physical Ability	0.203995	0.173168	0.169247	0.302536
Integrity	-.130757	-.100039	0.029456	-.317545

Eigenvectors

	Prin9	Prin10	Prin11	Prin12
Communication Skills	-.430955	0.375983	0.028370	-.252778
Problem Solving	0.256098	-.372914	-.434417	0.069863
Learning Ability	0.224706	0.287031	0.210540	-.284355
Judgment Under Pressure	-.358618	0.178270	0.118318	0.306490
Observational Skills	0.258377	0.223793	-.079692	0.565290
Willingness to Confront Problems	0.129976	-.330710	0.275249	-.386151
Interest in People	0.321200	-.081470	0.393841	-.210915
Interpersonal Sensitivity	0.137468	-.074821	0.285447	0.276824
Desire for Self-Improvement	-.121474	-.363854	-.052085	0.151436
Appearance	0.087395	0.061890	0.168369	0.236655

Output 70.3.2 *continued*

Eigenvectors

	Prin9	Prin10	Prin11	Prin12
Dependability	-.495598	-.377561	-.164909	-.090904
Physical Ability	-.149625	0.258321	-.006202	-.055828
Integrity	0.271060	0.297010	-.612497	-.276273

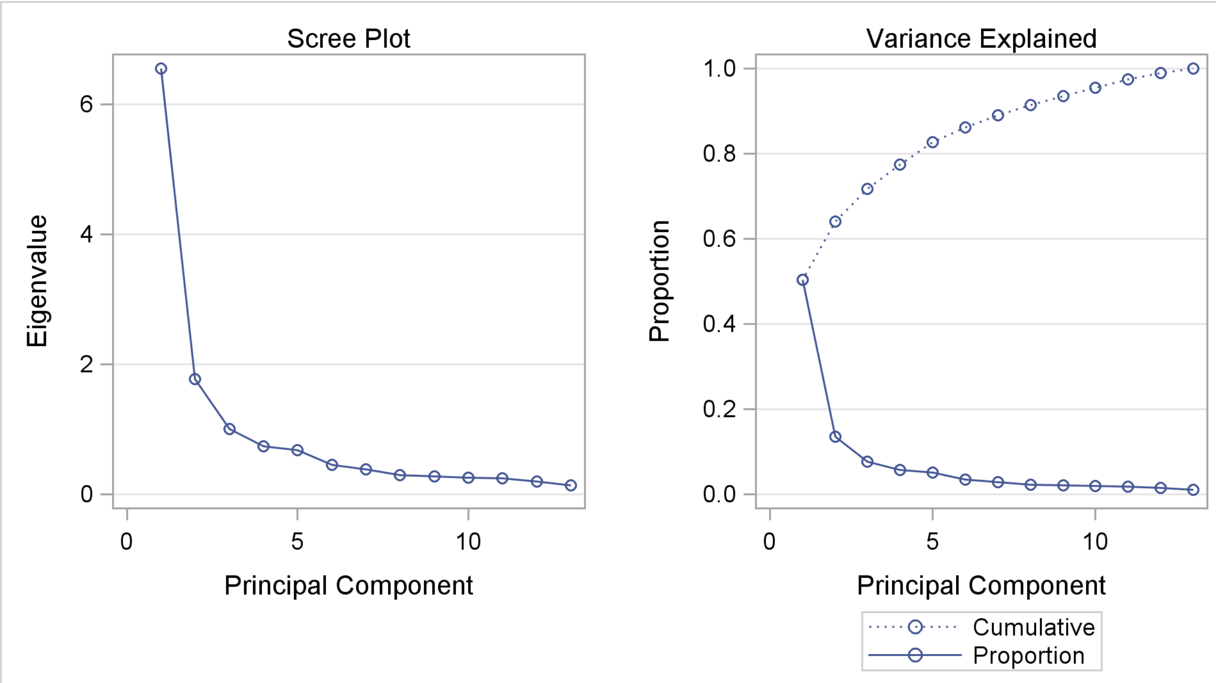
Eigenvectors

	Prin13
Communication Skills	-.122809
Problem Solving	-.116642
Learning Ability	0.248555
Judgment Under Pressure	-.126636
Observational Skills	-.168555
Willingness to Confront Problems	0.177688
Interest in People	-.610215
Interpersonal Sensitivity	0.643410
Desire for Self-Improvement	0.053834
Appearance	-.113705
Dependability	-.018094
Physical Ability	0.133430
Integrity	0.114965

When the **ods graphics on** statement is specified, PROC PRINCOMP produces the scree plot as shown in Figure 70.3.3 by default, which helps to visualize and choose the number of components. You can obtain more plots by specifying the PLOTS= option in the PROC PRINCOMP statement.

The “Scree Plot” on the left shows that the eigenvalue of the first component is approximately 6.5 and the eigenvalue of the second component is largely decreased to under 2.0. The “Variance Explained” plot on the right shows that you can explain a near 80% of total variance with the first four principal components.

Output 70.3.3 Scree Plot from the PRINCOMP Procedure



The first component reflects overall performance since the first eigenvector shows approximately equal loadings on all variables. The second eigenvector has high positive loadings on the variables Observational Skills and Willingness to Confront Problems but even higher negative loadings on the variables Interest in People and Interpersonal Sensitivity. This component seems to reflect the ability to take action, but it also reflects a lack of interpersonal skills. The third eigenvector has a very high positive loading on the variable Physical Ability and high negative loadings on the variables Problem Solving and Learning Ability. This component seems to reflect physical strength, but also shows poor learning and problem-solving skills.

In short, the three components represent the following:

First Component:	overall performance
Second Component:	smart, tough, and introverted
Third Component:	superior strength and average intellect

PROC PRINCOMP also produces other plots besides the scree plot, which are helpful while interpreting the results. The following statements request plots from the PRINCOMP procedure:

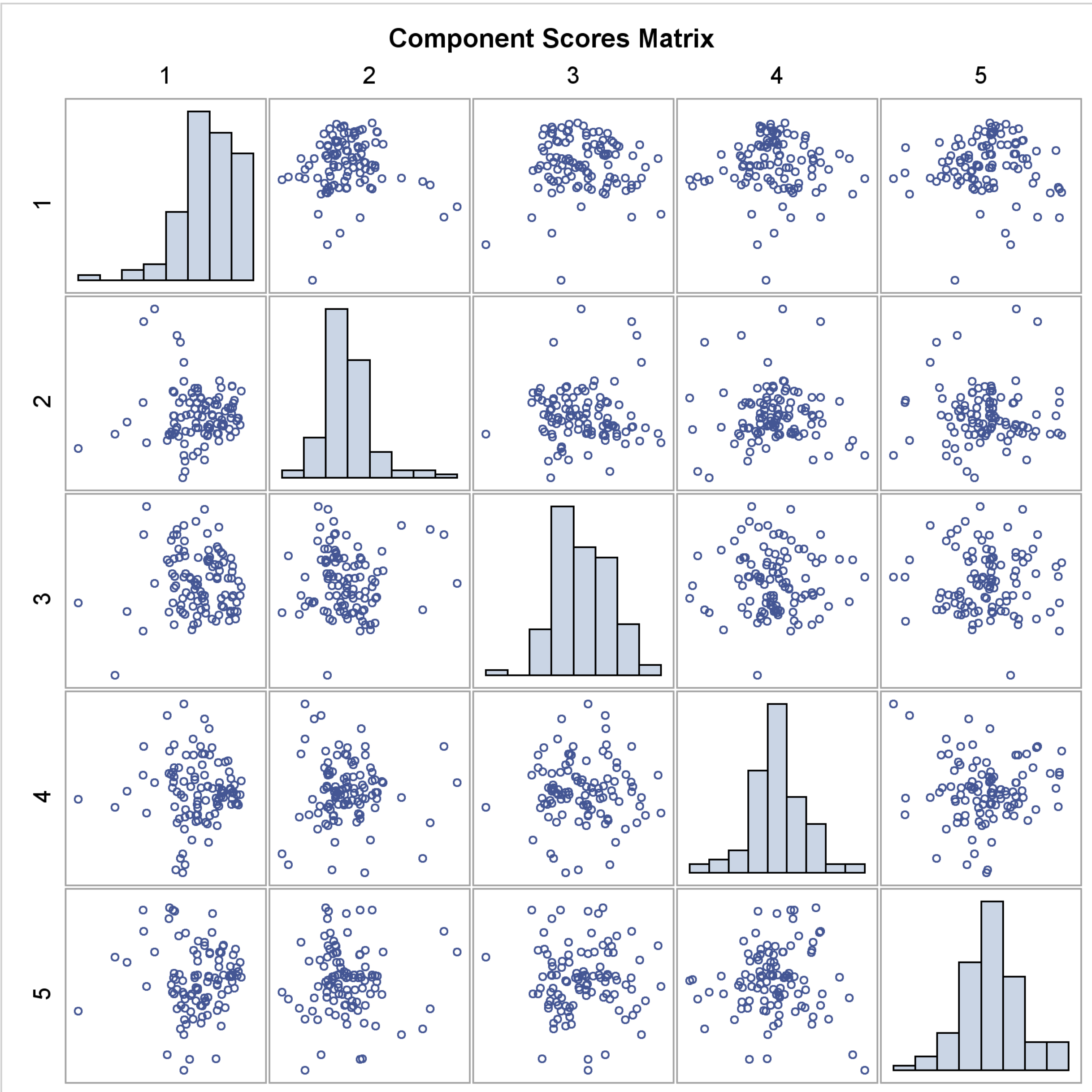
```
proc princomp data=Jobratings(drop='Overall Rating'n)
              plots(ncomp=3)=all n=5;
run;
ods graphics off;
```

PLOTS=ALL(NCOMP=3) in the PROC PRINCOMP statement requests all plots to be produced but limits the number of components to be plotted in the component pattern plots and the component score plots to three. The N=5 option sets the number of principal components to be computed to five.

Besides a scree plot similar to the one shown before, the rest of plots are displayed in the following context.

Output 70.3.4 shows a matrix plot of component scores between the first five principal components. The histogram of each component is displayed in the diagonal element of the matrix. The histograms indicate that the first principal component is skewed to the left and the second principal component is slightly skewed to the right.

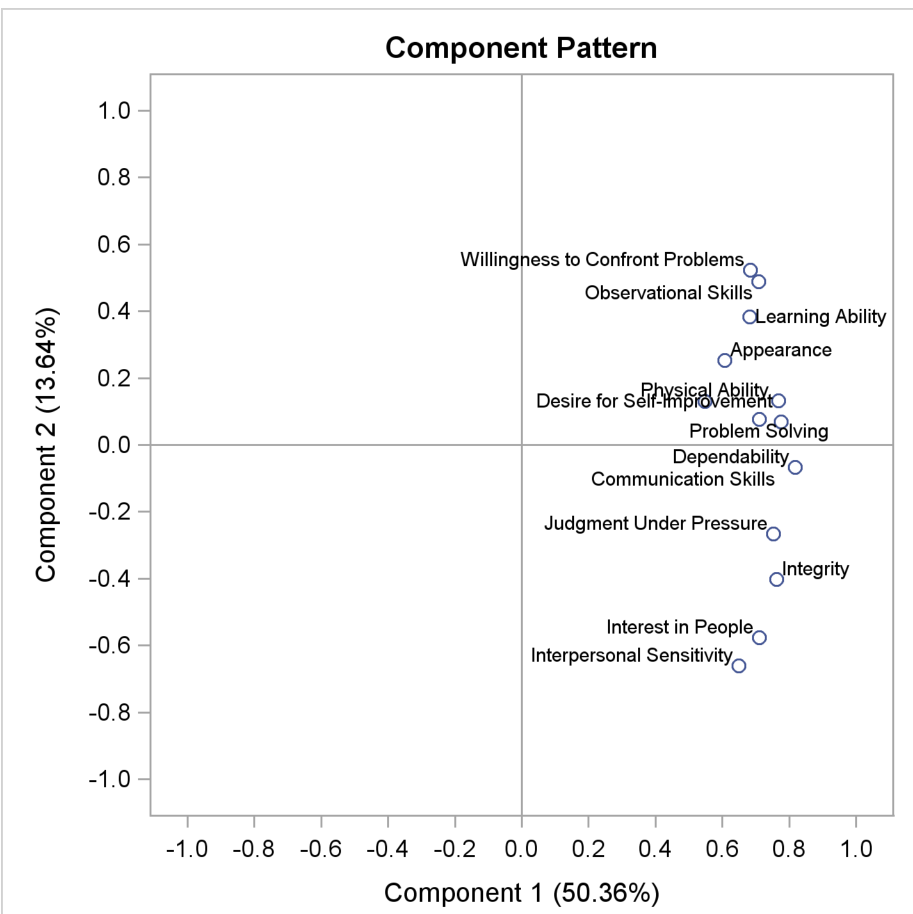
Output 70.3.4 Matrix Plot of Component Scores



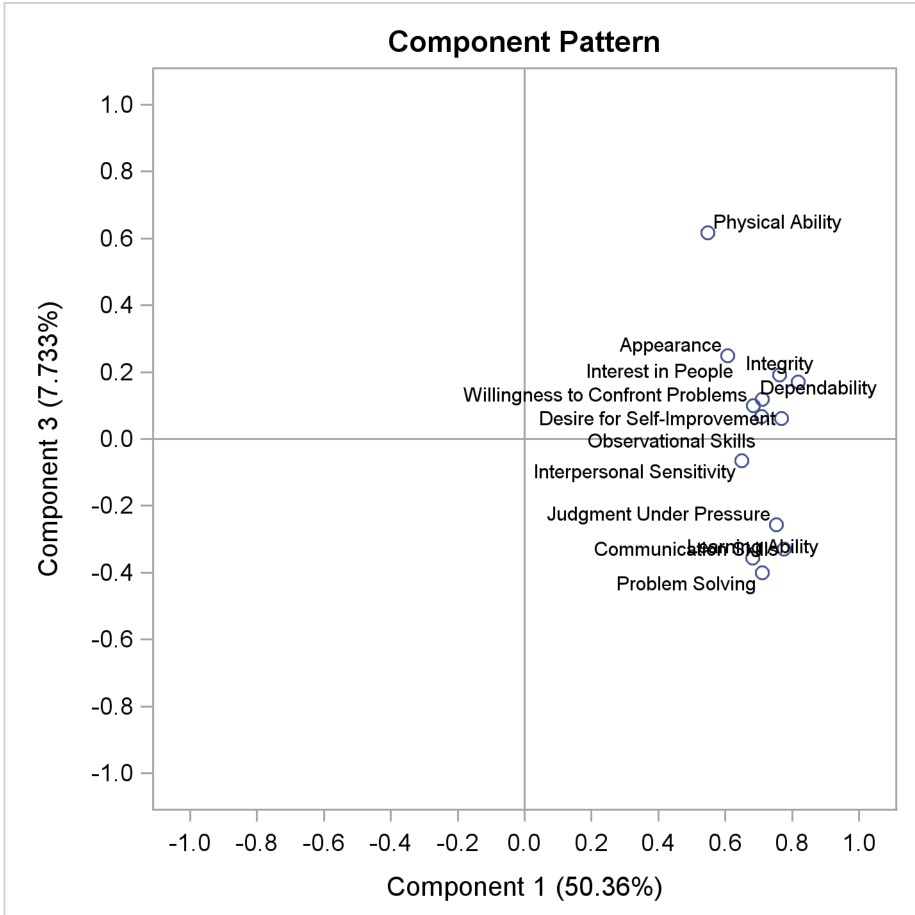
The pairwise component pattern plots are shown in Output 70.3.5 to Output 70.3.7. The pattern plots show the following:

- All variables positively and evenly correlate with the first principal component (Output 70.3.5 and Output 70.3.6).
- The variables Observational Skills and Willingness to Confront Problems correlate highly with the second component, and the variables Interest in People and Interpersonal Sensitivity correlate highly but negatively with the second component (Output 70.3.5).
- The variable Physical Ability correlates highly with the third component, and the variables Problem Solving and Learning Ability correlate highly but negatively with the third component (Output 70.3.6).
- The variable Observational Skills, Willingness to Confront Problems, Interest in People, and Interpersonal Sensitivity correlate highly (either positively or negatively) with the second component, but all have very low correlations with the third component; the variables Physical Ability and Problem Solving correlate highly (either positively or negatively) with the third component, but both have very low correlations with the second component (Output 70.3.7).

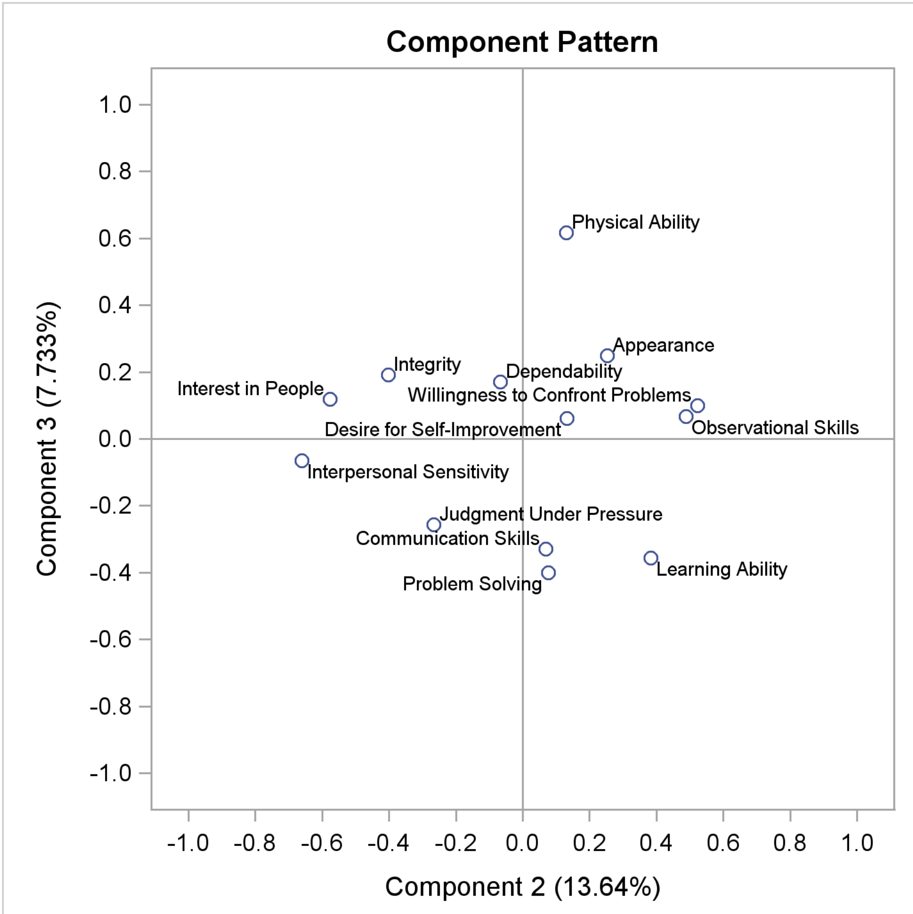
Output 70.3.5 Pattern Plot of Component 2 by Component 1



Output 70.3.6 Pattern Plot of Component 3 by Component 1

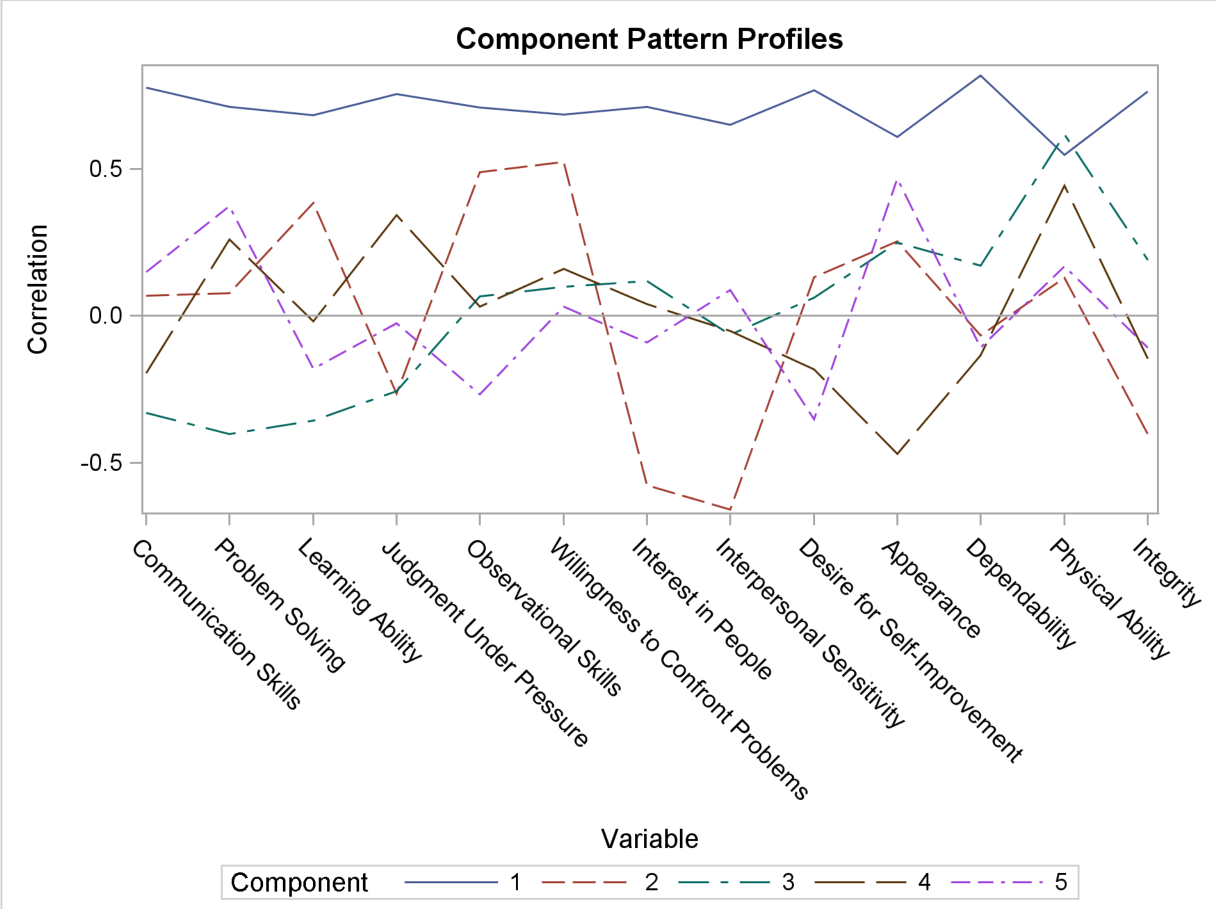


Output 70.3.7 Pattern Plot of Component 3 by Component 2



Output 70.3.8 shows a component pattern profile. As it was shown in the pattern plots, the nearly horizontal profile from the first component indicates that the first component is mostly correlated evenly across all variables.

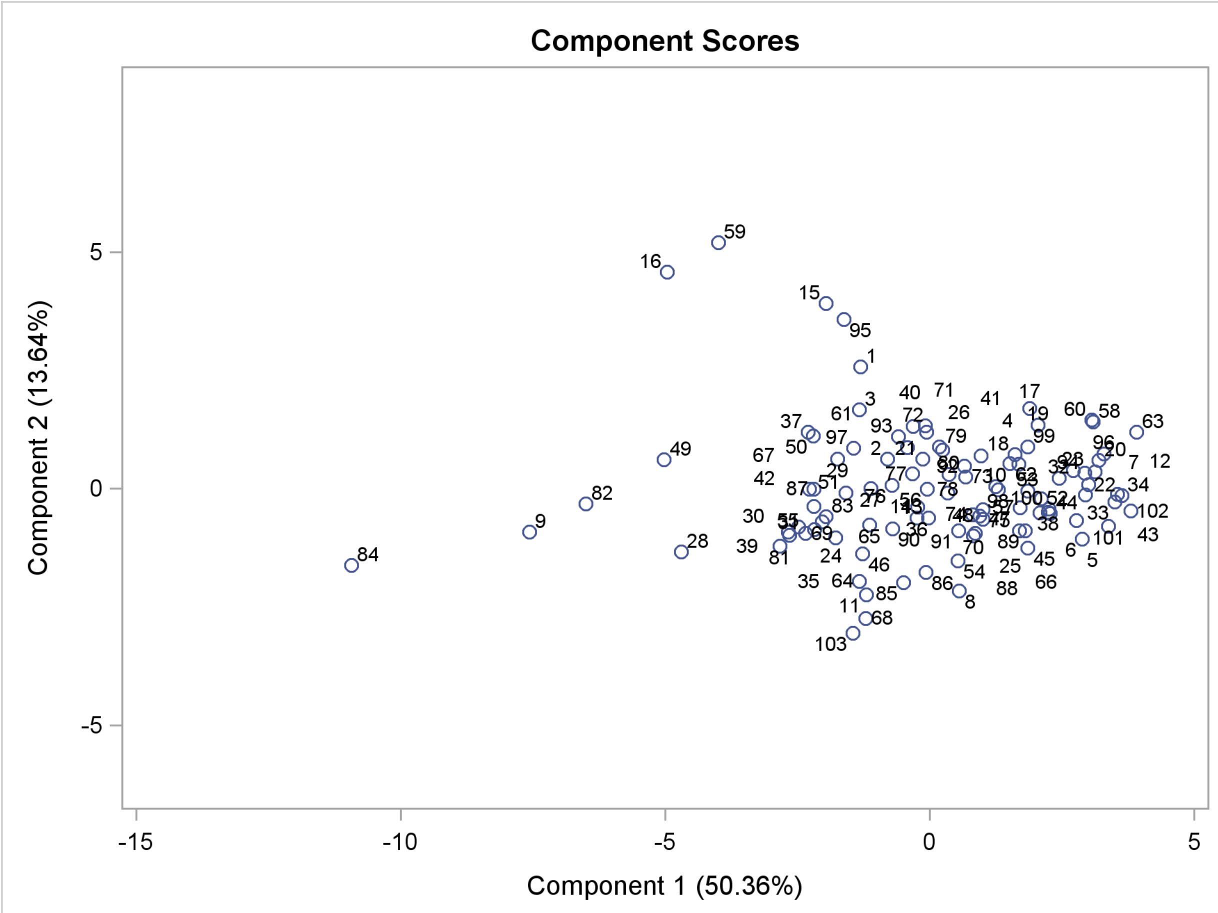
Output 70.3.8 Component Pattern Profile Plot from the PRINCOMP Procedure



Output 70.3.9 through Output 70.3.11 display the pairwise component score plots. Observation numbers are used as the plotting symbol.

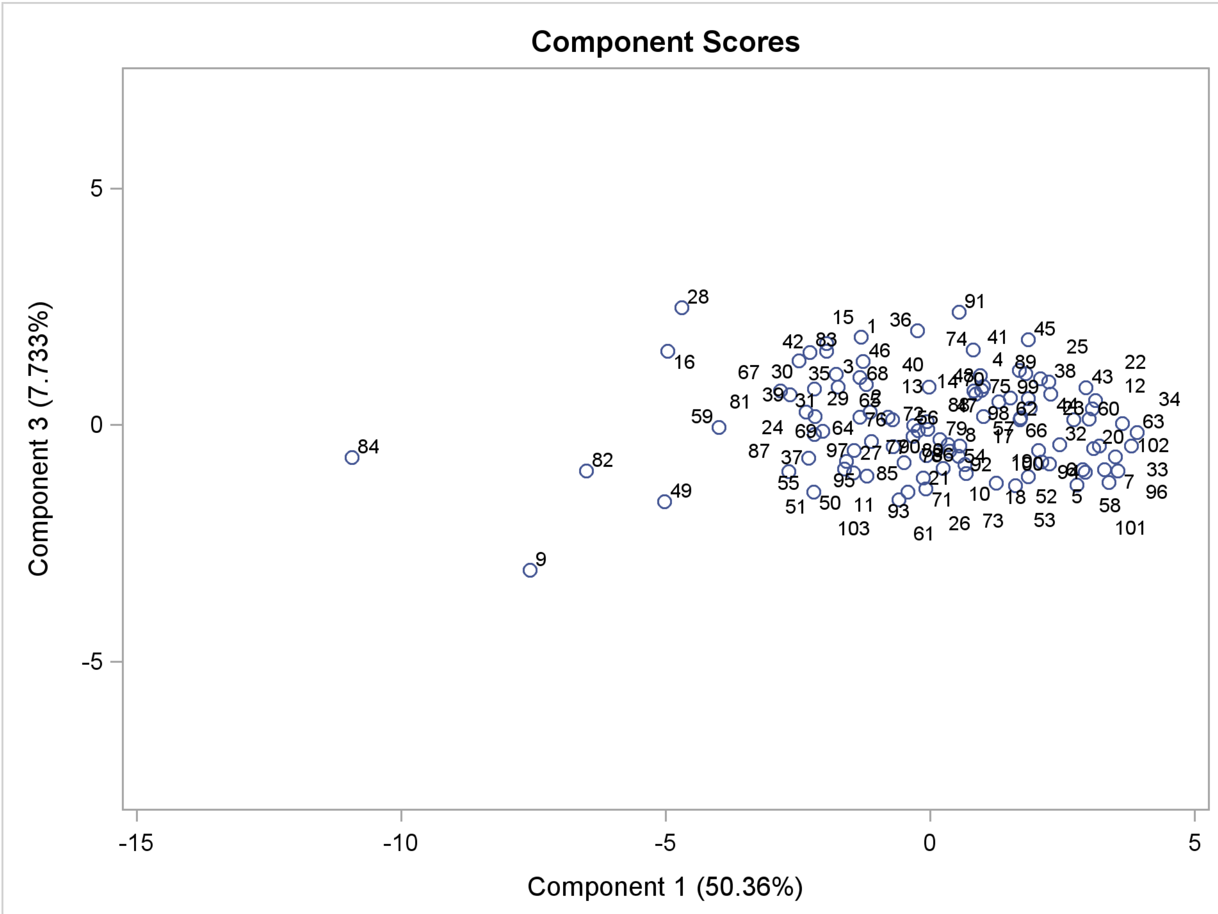
Output 70.3.9 shows a scatter plot of the first and third components. Observations 82, 9, and 84 seem like outliers on the first component; Observations 16 and 59 can be potential outliers on the second component.

Output 70.3.9 Component 2 versus Component 1



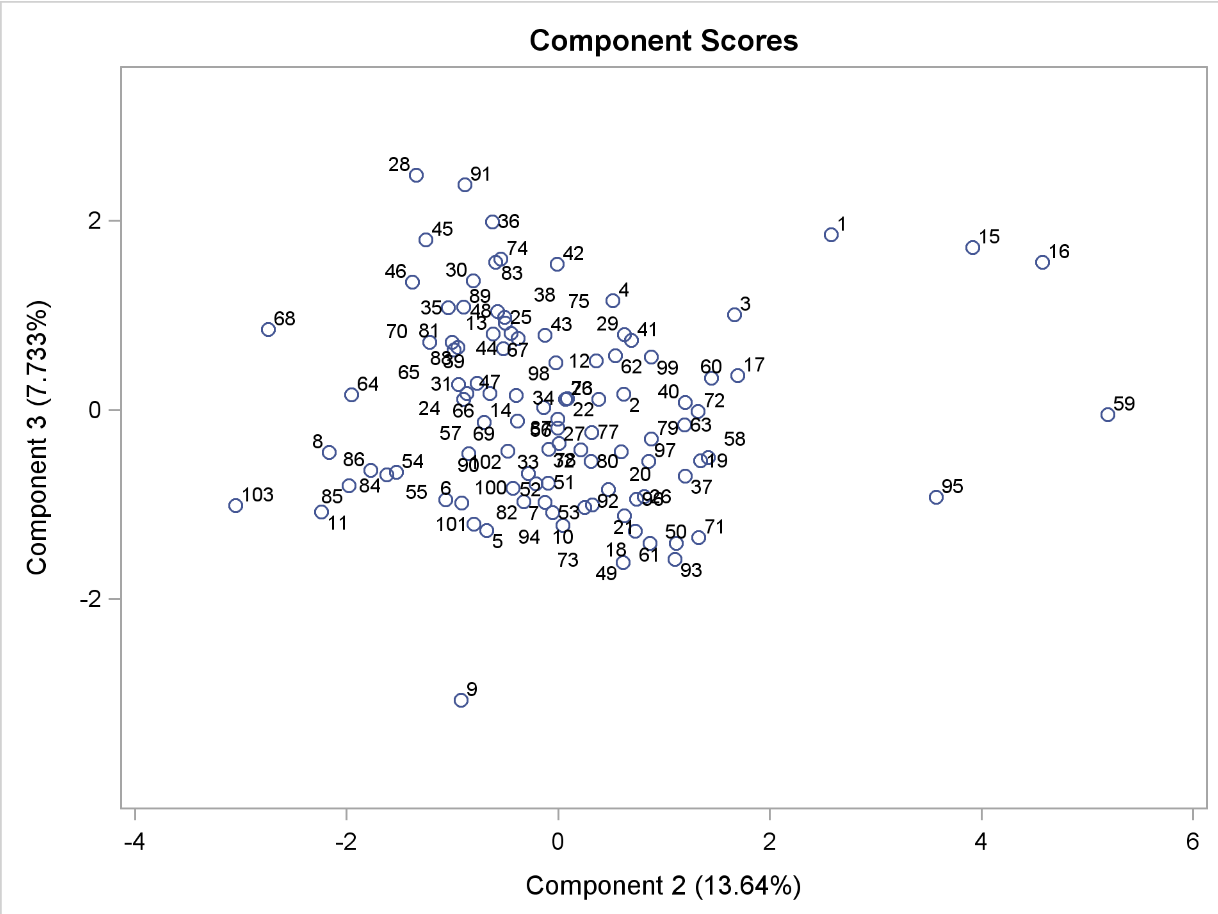
Output 70.3.10 shows a scatter plot of the first and third components. Observations 82, 9, and 84 seem like outliers on the first component.

Output 70.3.10 Component 3 versus Component 1



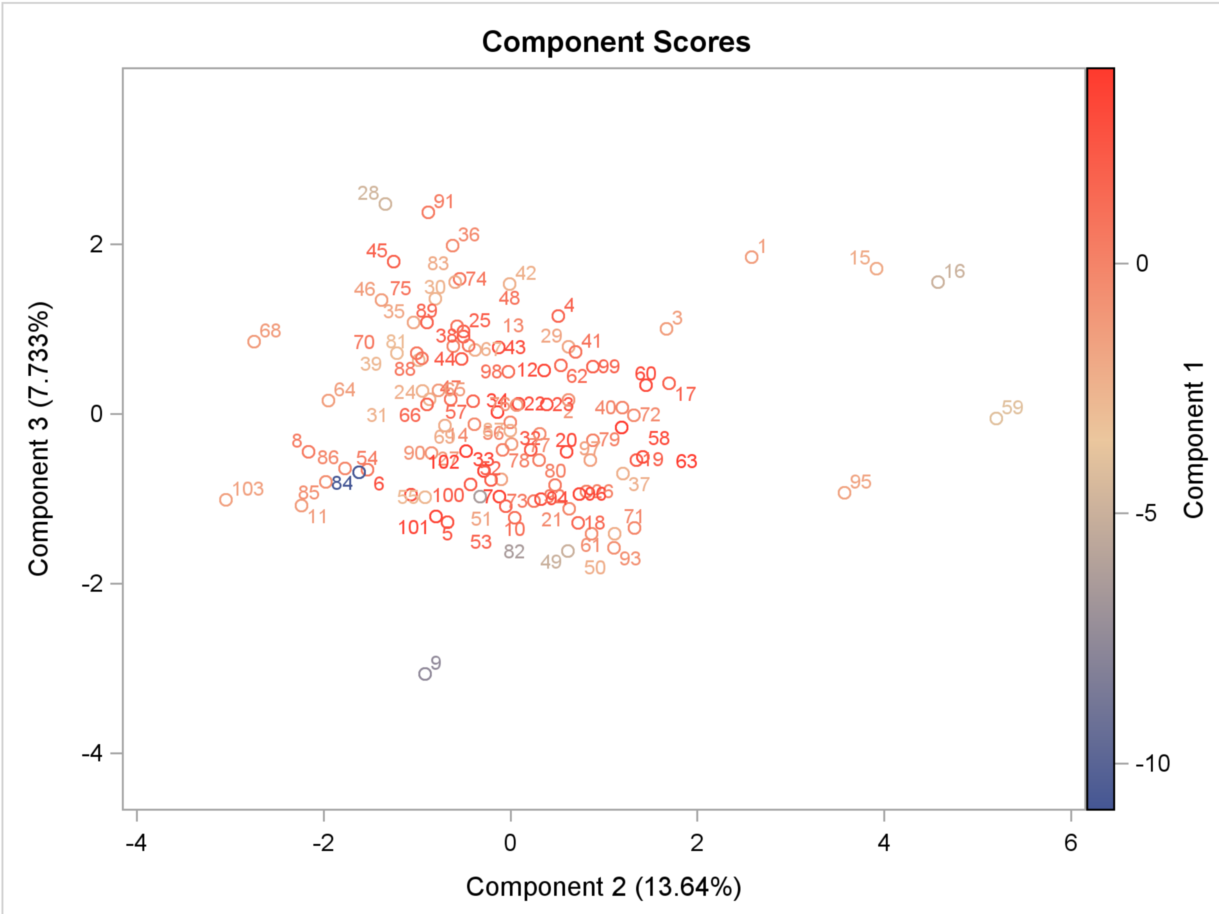
Output 70.3.11 shows a scatter plot of the second and third components. Observations 95, 15, 16, and 59 can be potential outliers on the second component.

Output 70.3.11 Component 3 versus Component 2



Output 70.3.12 shows a scatter plot of the second and third components, displaying density with color. Color interpolation is based on the first component, such as in the statistical style, going from blue (minimum density) to tan (median density) and to red (maximum density).

Output 70.3.12 Component 3 versus Component 2, Painted by Component 1



References

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

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

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.

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.

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

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

Subject Index

Syntax Index

Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **yourturn@sas.com**. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to **suggest@sas.com**.