Chapter 425

Principal Components Analysis

Introduction

Principal Components Analysis, or *PCA*, is a data analysis tool that is usually used to reduce the dimensionality (number of variables) of a large number of interrelated variables, while retaining as much of the information (variation) as possible. PCA calculates an uncorrelated set of variables (*factors* or *pc's*). These factors are ordered so that the first few retain most of the variation present in all of the original variables. Unlike its cousin Factor Analysis, PCA always yields the same solution from the same data (apart from arbitrary differences in the sign).

The computations of PCA reduce to an eigenvalue-eigenvector problem. **NCSS** uses a double-precision version of the modern QL algorithm as described by Press (1986) to solve the eigenvalue-eigenvector problem.

Note that PCA is a data analytical, rather than statistical, procedure. Hence, you will not find many t-tests or F-tests in PCA. Instead, you will make subjective judgments requiring you to spend a little extra time getting acquainted with the technique.

This **NCSS** program performs a PCA on either a correlation or a covariance matrix. Missing values may be dealt with using one of three methods. The analysis may be carried out using robust estimation techniques.

Chapters on PCA are contained in books dealing with multivariate statistical analysis. Books that are devoted solely to PCA include Dunteman (1989), Jolliffe (1986), Flury (1988), and Jackson (1991).

Technical Details

Mathematical Development

This section will document the basic formulas used by **NCSS** in performing a principal components analysis. We begin with an adjusted data matrix, X, which consists of n observations (rows) on p variables (columns). The adjustment is made by subtracting the variable's mean from each value. That is, the mean of each variable is subtracted from all of that variable's values. This adjustment is made since PCA deals with the covariances among the original variables, so the means are irrelevant.

New variables are constructed as weighted averages of the original variables. These new variables are called the factors, latent variables, or principal components. Their specific values on a specific row are referred to as the factor scores, the component scores, or simply the scores. The matrix of scores will be referred to as the matrix Y. The basic equation of PCA is, in matrix notation, given by:

$$Y = W'X$$

where W is a matrix of coefficients that is determined by PCA. This matrix is provided in **NCSS** in the *Score Coefficients* report. For those not familiar with matrix notation, this equation may be thought of as a set of p linear equations that form the factors out of the original variables.

These equations are also written as:

$$y_{ij} = w_{1i}x_{1j} + w_{2i}x_{2j} + \ldots + w_{pi}x_{pj}$$

As you can see, the factors are a weighted average of the original variables. The weights, *W*, are constructed so that the variance of y_1, $Var(y_1)$, is maximized. Also, so that $Var(y_2)$ is maximized and that the correlation between y_1 and y_2 is zero. The remaining y_i's are calculated so that their variances are maximized, subject to the constraint that the covariance between y_i and y_j, for all i and j (i not equal to j), is zero.

The matrix of weights, *W*, is calculated from the variance-covariance matrix, *S*. This matrix is calculated using the formula:

$$s_{ij} = \frac{\sum_{k=1}^{n}\left(x_{ik} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_j\right)}{n-1}$$

Later, we will discuss how this equation may be modified both to be robust to outliers and to deal with missing values.

The singular value decomposition of *S* provides the solution to the PCA problem. This may be defined as:

$$USU = L$$

where *L* is a diagonal matrix of the eigenvalues of *S*, and *U* is the matrix of eigenvectors of *S*. *W* is calculated from *L* and *U*, using the relationship:

$$W = UL^{-\frac{1}{2}}$$

It is interesting to note that *W* is simply the eigenvector matrix *U*, scaled so that the variance of each factor, y_i, is one.

The correlation between an i^{th} factor and the j^{th} original variable may be computed using the formula:

$$r_{ij} = \frac{u_{ji}\sqrt{l_i}}{s_{jj}}$$

Here u_{ij} is an element of *U*, l_i is a diagonal element of *L*, and s_{jj} is a diagonal element of *S*. The correlations are called the factor loadings and are provided in the *Factor Loadings* report.

When the correlation matrix, *R*, is used instead of the covariance matrix, *S*, the equation for *Y* must be modified. The new equation is:

$$Y = W'D^{-\frac{1}{2}}X$$

where *D* is a diagonal matrix made up of the diagonal elements of *S*. In this case, the correlation formula may be simplified since the s_{jj} are equal to one.

Missing Values

Missing values may be dealt with by ignoring rows with missing values, estimating the missing value with the variable's average, or estimating the missing value by regressing it on variables whose values are not missing. These will now be described in detail. Most of this information comes from Jackson (1991) and Little (1987).

When estimating statistics from data sets with missing values, you should first consider the mechanism that created the missing values. This mechanism determines whether your method of dealing with the missing values is appropriate. The worst case arises when the probability of obtaining a missing value is dependent on one or more variables in your study. For example, suppose one of your variables was a person's income level. You might suspect that the higher a person's income, the less likely he is to reveal it to you. When the probability of

obtaining a missing value is dependent on one or more variables, serious biases can occur in your results. A complete discussion of missing value mechanisms is given in Little (1987).

NCSS provides three methods of dealing with missing values. In all three cases, the overall strategy is to deal with the missing values while estimating the covariance matrix, *S*. Hence, the rest of the section will consider estimating *S*.

Complete-Case Missing-Value Analysis

One method of dealing with missing values is to remove all cases (observations or rows) that contain missing values from the analysis. The analysis is then performed only on those cases that are "complete."

The advantages of this approach are *speed* (since no iteration is required), *comparability* (since univariate statistics, such as the mean, calculated on individual variables, will be equal to the results of the multivariate calculations), and *simplicity* (since the method is easy to explain).

Disadvantages of this approach are *inefficiency* and *bias*. This method is inefficient since as the number of missing values increases, the number of discarded cases also increases. In the extreme case, suppose a data set has 100 variables and 200 cases. Suppose one value is missing at random in 80 cases, so these cases are deleted from the study. Hence, of the 20,000 values in the study, 80 values or 0.4% were missing. Yet this method has us omit 8000 values or 40%, even though 7920 of those values were actually available. This is similar to the saying that one rotten apple ruins the whole barrel.

A certain amount of bias may occur if the pattern of missing values is related to at least one of the variables in the study. This could lead to gross distortions if this variable were correlated with several other variables.

One method of determining if the complete-case methodology is causing bias is to compare the means of each variable calculated from only complete cases, with the corresponding means of each variable calculated from cases that were dropped but had this variable present. This comparison could be run using a statistic like the t-test, although we would also be interested in comparing the variances, which the t-test does not do. Significant differences would indicate the presence of a strong bias introduced by the pattern of missing values.

A modification of the complete-case method is the pairwise available-case method in which covariances are calculated one at a time from all cases that are complete for those two variables. This method is not available in this program for three reasons: the univariate statistics change from pair to pair causing serious numeric problems (such as correlations greater than one), the resulting covariance matrix may not be positive semi-definite, and the method is dominated by other methods that are available in this program.

Filling in Missing Values with Averages

A growing number of programs offer the ability to fill in (or impute) the missing values. The naive choice is to fill in with the variable average. **NCSS** offers this option, implemented iteratively. During the first iteration, no imputation occurs. On the second, third, and additional iterations, each missing value is estimated using the mean of that variable from the previous iteration. Hence, at the end of each iteration, a new set of means is available for imputation during the next iteration. The process continues until it converges.

The advantages of this method are greater efficiency (since it takes advantage of the cases in which missing values occur) and speed (since it is much faster than the EM algorithm to be presented next).

The disadvantages of this method are biases (since it consistently underestimates the variances and covariances), unreliability (since simulation studies have shown it unreliable in some cases), and domination (since it is dominated by the EM algorithm, which does much better although that method requires more computations).

Multivariate-Normal Missing-Value Imputation

Little (1987) has documented the use of the EM algorithm for estimating the covariance matrix, *S*, when the data follow the multivariate normal distribution. This might also be referred to as a regression approach or modified conditional means approach. The assumption of a multivariate normal distribution may seem limiting, but the

procedure produces estimates that are consistent under weaker assumptions. We will now define the algorithm for you.

1. Estimate the covariance matrix, *S*, with the complete-case method.
2. The E step consists of calculating the sums and sums of squares using the following formulas:

$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} x_{ij}^{(t)}}{n}$$

$$s_{jk}^{(t+1)} = \frac{\sum_{i=1}^{n}\left[\left(x_{ij}^{(t)} - \hat{\mu}_j^{(t+1)}\right)\left(x_{ik}^{(t)} - \hat{\mu}_k^{(t+1)}\right) + c_{jki}^{(t)}\right]}{n-1}$$

$$x_{ij}^{(t)} = \begin{cases} x_{ij}, & \text{if } x_{ij} \text{ is observed} \\ E\left(x_{ij} \mid x_{obs,i}, \hat{\mu}, S^{(t)}\right), & \text{if } x_{ij} \text{ is missing} \end{cases}$$

$$c_{jki}^{(t)} = \begin{cases} 0 & \text{if } x_{ij} \text{ or } x_{ik} \text{ are observed} \\ Cov(x_{ij}, x_{ik} \mid x_{obs,i}, S^{(t)}) & \text{if } x_{ij} \text{ and } x_{ik} \text{ are missing} \end{cases}$$

where $x_{obs,i}$ refers to that part of observation *i* that is not missing and $E\left(x_{ij} \mid x_{obs,i}, \hat{\mu}, S^{(t)}\right)$ refers to the regression of the variables that are missing on the variables that are observed. This regression is calculated by sweeping *S* by the variables that are observed and using the observed values as the values of the independent variables in the resulting regression equation. Essentially, we are fitting a multiple regression of each missing value on the values that are observed, using the *S(t)* matrix as our matrix of sums of squares and cross products. When both x_{ij} and x_{ik} are missing, the value of c_{jki} is the ij^{th} element of the swept *S* matrix.

Verbally, the algorithm may be stated as follows. Each missing data value is estimated by regressing it on the values that are observed. The regression coefficients are calculated from the current covariance matrix. Since this regression tends to underestimate the true covariance values, these are inflated by an appropriate amount. Once each missing value is estimated, a new covariance matrix is calculated and the process is repeated. The procedure is terminated when it converges. This convergence is measured by the trace of the covariance matrix.

NCSS first sorts the data according to the various patterns of missing values, so that the regression calculations (the sweeping of *S*) are performed a minimum number of times: once for each particular missing-value pattern.

This method has the disadvantage that it is computationally intensive and it may take twenty or more iterations to converge. However, it provides the maximum-likelihood estimate of the covariance matrix, it provides a positive semi-definite covariance matrix, and it seems to do well even when the occurrences of missing values are correlated with the values of the variables being studied. That is, it corrects for biases caused by the pattern of missing values.

Robust Estimation

Robust estimation refers to estimation techniques that decrease or completely remove the influence of observations that are outliers. These outliers can seriously distort the estimated means and covariances. The EM algorithm is employed as the robust technique used in **NCSS**. This algorithm uses weights that are inversely proportional to how "outlying" the observation is. The usual estimates of the means and covariances are modified to use these weights. The process is iterated until it converges. Note that since *S* is estimated robustly, the estimated correlation matrix is robust also.

One advantage of the EM algorithm is that it can be modified to deal with missing values and robust estimation at the same time. Hence, **NCSS** provides robust estimates that use the information in rows with missing values as well. The robust estimation formulas are:

$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_i^{(t)} x_{ij}^{(t)}}{\sum_{i=1}^{n} w_i^{(t)}}$$

$$s_{jk}^{(t+1)} = \frac{\sum_{i=1}^{n} \left[w_i^{(t)} \left(x_{ij}^{(t)} - \hat{\mu}_j^{(t+1)} \right) \left(x_{ik}^{(t)} - \hat{\mu}_k^{(t+1)} \right) + c_{jki}^{(t)} \right]}{n-1}$$

The weights, w_i, are calculated using the formula:

$$w_i = \frac{(\nu + p_i)}{(\nu + d_i^2)}$$

where ν is a parameter you supply, p_i is the number of nonmissing values in the i^{th} row, and

$$d_i^2 = \sum_{j=1}^{p} \sum_{k=1}^{p} \delta_{ijk} \left(x_{ij} - \hat{\mu}_j \right) \left(x_{ik} - \hat{\mu}_k \right) b^{jk}$$

where δ_{ijk} is equal to one if both variables x_j and x_k are observed in row i and is zero otherwise. The b^{jk} are the indicated elements of the inverse of S ($B = S^{-1}$). Note that B is found by sweeping S on all variables.

When using robust estimation, it is wise to run the analysis with the robust option turned on and then study the robust weights. When the weight is less than .4 or .3, the observation is being "removed." You should study rows that have such a weight to determine if there was an error in data entry or measurement, or if the values are valid. If the values are all valid, you have to decide whether this row should be kept or discarded. Next, make a second run without the discarded rows and without using the robust option. In this way, your results do not depend quite so much on the particular formula that was used to create the weights. Note that the weights are listed in the *Residual Report* after the values of Q_k and T^2.

How Many Factors

Several methods have been proposed for determining the number of factors that should be kept for further analysis. Several of these methods will now be discussed. However, remember that important information about possible outliers and linear dependencies may be determined from the factors associated with the relatively small eigenvalues, so these should be investigated as well.

Kaiser (1960) proposed dropping factors whose eigenvalues are less than one, since these provide less information than is provided by a single variable. Jolliffe (1972) feels that Kaiser's criterion is too large. He suggests using a cutoff on the eigenvalues of 0.7 when correlation matrices are analyzed. Other authors note that if the largest eigenvalue is close to one, then holding to a cutoff of one may cause useful factors to be dropped. However, if the largest factors are several times larger than one, then those near one may be reasonably dropped.

Cattell (1966) documented the *scree graph*, which will be described later in this chapter. Studying this chart is probably the most popular method for determining the number of factors, but it is subjective, causing different people to analyze the same data with different results.

Another criterion is to preset a certain percentage of the variation that must be accounted for and then keep enough factors so that this variation is achieved. Usually, however, this cutoff percentage is used as a lower limit. That is, if the designated number of factors do not account for at least 50% of the variance, then the whole analysis is aborted.

We cannot give a definitive answer as to which criterion is best, since most of these techniques were developed for use in factor analysis, not PCA. Perhaps the best advise we can give is to use the number of factors that agrees with the goals of your analysis. If you want to look for outliers in multivariate data, then you will want to keep

most, if not all, factors during the early stages of the analysis. If you want to reduce the dimensionality of your database, then you should keep enough factors so that you account for a reasonably large percentage of the variation.

Varimax and Quartimax Rotation

PCA finds a set of dimensions (or coordinates) in a subspace of the space defined by the set of variables. These coordinates are represented as axes. They are orthogonal (perpendicular) to one another. For example, suppose you analyze three variables that are represented in three-dimensional space. Each variable becomes one axis. Now suppose that the data lie near a two-dimensional plane within the three dimensions. A PCA of this data should uncover two factors that would account for the two dimensions. You may rotate the axes of this two-dimensional plane while keeping the 90-degree angle between them, just as the blades of a helicopter propeller rotate yet maintain the same angles among themselves. The hope is that rotating the axes will improve your ability to interpret the meaning of each component.

Many different types of rotation have been suggested. Most of them were developed for use in factor analysis. **NCSS** provides two orthogonal rotation options: varimax and quartimax.

Varimax Rotation

Varimax rotation is the most popular orthogonal rotation technique. In this technique, the axes are rotated to maximize the sum of the variances of the squared loadings within each column of the loadings matrix. Maximizing according to this criterion forces the loadings to be either large or small. The hope is that by rotating the factors, you will obtain new factors that are each highly correlated with only a few of the original variables. This simplifies the interpretation of the factor to a consideration of these two or three variables. Another way of stating the goal of varimax rotation is that it clusters the variables into groups, where each group is actually a new factor.

Since varimax seeks to maximize a specific criterion, it produces a unique solution (except for differences in sign). This has added to its popularity. Let the matrix $B = \{b_{ij}\}$ represent the rotated factors. The goal of varimax rotation is to maximize the quantity:

$$Q_1 = \sum_{j=1}^{k} \left(\frac{p \sum_{i=1}^{p} b_{ij}^4 - \sum_{i=1}^{p} b_{ij}^2}{p} \right)$$

This equation gives the raw varimax rotation. This rotation has the disadvantage of not spreading the variance very evenly among the new factors. Instead, it tends to form one large factor followed by many small ones. To correct this, **NCSS** uses the normalized-varimax rotation. The quantity maximized in this case is:

$$Q_N = \sum_{j=1}^{k} \left[\frac{p \sum_{i=1}^{p} \left(\frac{b_{ij}}{h_i} \right)^4 - \sum_{i=1}^{p} \left(\frac{b_{ij}}{h_i} \right)^2}{p^2} \right]$$

where h_i is the square root of the communality of variable i.

Quartimax Rotation

Quartimax rotation is similar to varimax rotation, except that the rows of *B* are maximized rather than the columns of *B*. This rotation is more likely to produce a general factor than will varimax. Often, the results are quite similar. The quantity maximized for the quartimax is:

$$Q_N = \sum_{j=1}^{k} \left[\frac{\sum_{i=1}^{p} \left(\frac{b_{ij}}{h_i} \right)^4}{p} \right]$$

Miscellaneous Topics

Using Correlation Matrices Directly

Occasionally, you will be provided with only the correlation (or covariance) matrix from a previous analysis. This happens frequently when you want to analyze data that is presented in a book or a report. You can perform a partial PCA on a correlation matrix using **NCSS**. We say partial because you cannot analyze the individual scores, the row-by-row values of the factors. These are often very useful to investigate, but they require the raw data.

NCSS can store the correlation (or covariance) matrix on the current database. If it takes a great deal of computer time to build the correlation matrix, you might want to save it so you can use it while you determine the number of factors**.** You could then return to the original data to analyze the factor scores.

Using PCA to Select a Subset of the Original Variables

There are at least two reasons why a researcher might want to select a subset of the original variables for further use. These will now be discussed.

1. In some data sets the number of original variables is too large, making interpretation and analysis difficult. Also, the cost of obtaining and managing so many variables is prohibitive.
2. When using PCA, it is often difficult to find a reasonable interpretation for all the factors that are kept. Instead of trying to interpret each factor, McCabe (1984) has suggested finding the principal variables. Suppose you start with *p* variables, run a PCA, and decide to retain *k* factors. McCabe suggests that it is often possible to find *k+2* or *k+3* of the original variables that will account for the same amount of variability as the *k* factors. The interpretation of the variables is much easier than the interpretation of the factors.

Jolliffe (1986) discusses several methods to reduce the number of variables in a data set while retaining most of the variability. Using **NCSS**, one of the most effective methods for selecting a subset of the original variables can easily be implemented. This method is outlined next.

1. Perform a PCA. Save the *k* most important factor scores onto your database for further analysis.
2. Use the Multivariate Variable Selection procedure to reduce the number of variables. This is done by using the saved factor scores as the dependent variables and the original variables as the independent variables. The variable selection process finds the best subset of the original variables that predicts the group of factor scores. Since the factor scores represent the original variables, you are actually finding the best subset of the original variables.

   You will usually have to select two or three more variables than you did factors, but you will end up with most of the information in your data set being represented by a fraction of the variables.

Principal Component versus Factor Analysis

Both PCA and factor analysis (FA) seek to reduce the dimensionality of a data set. The most obvious difference is that while PCA is concerned with the total variation as expressed in the correlation matrix, *R*, FA is concerned with a correlation in a partition of the total variation called the common portion. That is, FA separates *R* into two matrices R_c (common factor portion) and R_u (unique factor portion). FA models the R_c portion of the correlation matrix. Hence, FA requires the discovery of R_c as well as a model for it. The goals of FA are more concerned with finding and interpreting the underlying, common factors. The goals of PCA are concerned with a direct reduction in the dimensionality.

Put another way, PCA is directed towards reducing the diagonal elements of *R*. Factor analysis is directed more towards reducing the off-diagonal elements of *R*. Since reducing the diagonal elements reduces the off-diagonal elements and vice versa, both methods achieve much the same thing.

Further Reading

There are several excellent books that provide detailed discussions of PCA. We suggest you first read the inexpensive monograph by Dunteman (1989). More complete (and mathematical) accounts are given by Jackson (1991) and Jolliffe (1986). Several books on multivariate methods provide excellent introductory chapters on PCA.

Data Structure

The data for a PCA consist of two or more variables. We have created an artificial data set in which each of the six variables (X1 - X6) were created using weighted averages of two original variables (V1 and V2) plus a small random error. For example, X1 = 0.33 V1 + 0.65 V2 + error. Each variable had a different set of weights (0.33 and 0.65 are the weights) in the weighted average.

Rows two and three of the data set were modified to be outliers so that their influence on the analysis could be observed. Note that even though these two rows are outliers, their values on each of the individual variables are not outliers. This shows one of the challenges of multivariate analysis: multivariate outliers are not necessarily univariate outliers. In other words, a point may be an outlier in a multivariate space, and yet you cannot detect it by scanning the data one variable at a time.

This data set is contained in the database PCA2. The data given in the table below are the first few rows of this data set.

PCA2 dataset (subset)

X1	X2	X3	X4	X5	X6
50	102	103	70	75	102
4	2	5	11	11	5
81	98	94	5	85	97
31	81	86	46	50	74
65	50	51	60	57	53
22	30	39	17	15	17
36	33	39	29	27	25
31	91	96	50	56	85

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Input Variables

Variables

Designates the variables to be analyzed. If matrix input is selected, indicate the variables containing the matrix. Note that for matrix input, the number of rows used is equal to the number of variables specified. Other rows will be ignored.

Data Input Format

Indicates whether raw data is to be analyzed or if a previously summarized correlation or covariance matrix is to be used.

- **Regular Data**
  The data is to be input in its raw format.

- **Lower-Triangular**
  The data is in a correlation or covariance matrix in lower-triangular format. This matrix could have been created by a previous run, or from direct keyboard input.

- **Upper-Triangular**
  The data is in a correlation or covariance matrix in upper-triangular format. The number of rows used is equal to the number of variables specified. This matrix could have been created by a previous run, or from direct keyboard input.

Covariance Estimation Options

Robust Covariance Matrix Estimation

This option indicates whether robust estimation is to be used. A full discussion of robust estimation is provided at the beginning of this chapter. If checked, robust estimates of the means, variances, and covariances are formed.

Robust Weight

This option specifies the value of v1. This parameter controls the weighting function in robust estimation of the covariance matrix. Jackson (1991) recommends the value 4.

Missing Value Estimation

This option indicates the type of missing value imputation method that you want to use. (Note that if the number of iterations is zero, this option is ignored.)

- **None**
  No missing value imputation. Rows with missing values in any of the selected variables are ignored.

- **Average**
  The average-value imputation method is used. Each missing value is estimated by the average value of that variable. The process is iterated as many times as is indicated in the second box.

- **Multivariate Normal**

  The multivariate-normal method. Each missing value is estimated using a multiple regression of the missing variable(s) on the variables that contain data in that row. This process is iterated as many times as indicated. See the discussion of missing value imputation methods elsewhere in this chapter.

Maximum Iterations

This option specifies the number of iterations used by either Missing Value Imputation or Robust Covariance Estimation. Robust estimation usually requires only four or five iterations to converge. Missing value imputation may require as many as twenty iterations if there are a lot of missing values.

When using this option, it is better to specify too many iterations than too few. After considering the Percent Change values in the Iteration Report, you can decide upon an appropriate number of iterations and re-run the problem.

Type of Matrix Used in Analysis

Matrix Type

This option indicates whether the analysis is to be run on a correlation or covariance matrix. Normally, the analysis is run on the scale-invariant correlation matrix since the scale of the variables changes the analysis when the covariance matrix is used. (For example, a variable that was measured in yards results in a different analysis than if it were measured in feet when a covariance matrix was used.)

Factor (Component) Options

Factor Rotation

Specifies the type of rotation, if any, that should be used on the solution. If rotation is desired, either varimax or quartimax rotation is available.

Factor Selection - Method

This option specifies which of the following three methods is used to set the number of factors retained in the analysis.

- **Percent of Eigenvalues**

  Specify the total percent of variation that must be accounted for. Enough factors will be included to account for this percentage (or slightly greater) of the variation in the data.

- **Number of Factors**

  Specify the number of factors.

- **Eigenvalue Cutoff**

  Specify the minimum eigenvalue amount. All factors whose eigenvalues are greater than or equal to this value will be retained. Older statistical texts suggest that you should only keep factors whose eigenvalues are greater than one.

Factor Selection - Value

This option sets a quantity corresponding to the method selected by the last option. For example, if you specified a Percent of Eigenvalues in the last option, you would enter the percentage (perhaps 80) here.

Reports Tab

The following options control the format of the reports.

Select Reports

Descriptive Statistics - Residuals (Q and T2)

These options let you specify which reports are displayed.

Alpha

The alpha value that is used in the residual reports to test if the observation is an outlier.

Report Options

Minimum Loading

Specifies the minimum absolute value that a loading can have and still remain in the Variable List report.

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Plots Tab

These sections specify the pair-wise plots of the scores and loadings.

Select Plots

Scores Plot - Loadings Plot

These options let you specify which reports and plots are displayed. Click the plot format button to change the plot settings.

Plot Options

Number of Factors Plotted

You can limit the number of plots generated using this parameter. Usually, you will only have interest in the first three or four factors.

Storage Tab

The factor scores and/or the correlation matrix may be stored on the current dataset for further analysis. This group of options let you designate which statistics (if any) should be stored and which columns should receive these statistics. The selected statistics are automatically stored to the current dataset. Note that existing data are replaced.

Data Storage Columns

Factor Scores

You can automatically store the factor scores for each row into the columns specified here. These scores are generated for each row of data in which all independent variable values are nonmissing.

Correlation or Covariance Matrix

Specifies columns to receive the correlation or covariance matrix. The type of matrix saved (i.e., correlation matrix or covariance matrix) depends on the Matrix Type option specified on the Variables tab.

Example 1 – Principal Components Analysis

This section presents an example of how to run a principal components analysis. The data used are found in the PCA2 dataset.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Principal Components Analysis window.

1 Open the PCA2 dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **PCA2.NCSS**.
- Click **Open**.

2 Open the Principal Components Analysis window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Principal Components Analysis** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Principal Components Analysis window, select the **Variables tab**.
- Double-click in the **Variables** box. This will bring up the variable selection window.
- Select **X1** through **X6** from the list of variables and then click **Ok**. "X1-X6" will appear in the Variables box.
- Check the **Robust Covariance Matrix Estimation** box.

4 Specify which reports.

- Select the **Reports tab**.
- Check all reports and plots. Normally you would only view a few of these reports, but we are selecting them all so that we can document them.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Descriptive Statistics Section

Descriptive Statistics Section

Variables	Count	Mean	Standard Deviation	Communality
X1	30	44.2	24.66241	1.000000
X2	30	51.53333	30.57803	1.000000
X3	30	54.93333	29.05753	1.000000
X4	30	41.7	25.3175	1.000000
X5	30	43.66667	26.65143	1.000000
X6	30	47.63334	34.18962	1.000000

This report lets us compare the relative sizes of the standard deviations. In this data set, they are all about the same size, so we could analyze either the correlation or the covariance matrix. We will analyze the correlation matrix.

Count, Mean, and Standard Deviation

These are the familiar summary statistics of each variable. They are displayed to allow you to make sure that you have specified the correct variables. Note that using missing value imputation or robust estimation will change these values.

Communality

The communality shows how well this variable is predicted by the retained factors. It is the R^2 that would be obtained if this variable were regressed on the factors that were kept. In this example, all factors were kept, so the R^2 is one.

Correlation Section

Correlation Section

	Variables				
Variables	X1	X2	X3	X4	X5
X1	1.000000	0.347229	0.224730	0.734112	0.819983
X2	0.347229	1.000000	0.990372	0.557526	0.799049
X3	0.224730	0.990372	1.000000	0.475404	0.710086
X4	0.734112	0.557526	0.475404	1.000000	0.830195
X5	0.819983	0.799049	0.710086	0.830195	1.000000
X6	0.514102	0.974167	0.935223	0.693869	0.907416

Phi=0.735970 Log(Det|R|)=-22.779188 Bartlett Test=596.06 DF=15 Prob=0.000000

	Variables
Variables	X6
X1	0.514102
X2	0.974167
X3	0.935223
X4	0.693869
X5	0.907416
X6	1.000000

Phi=0.735970 Log(Det|R|)=-22.779188 Bartlett Test=596.06 DF=15 Prob=0.000000

Bar Chart of Absolute Correlation Section

	Variables				
Variables	X1	X2	X3	X4	X5
X1		|||||||	|||||	||||||||||||||	||||||||||||||||
X2	|||||||		||||||||||||||||||	|||||||||||	||||||||||||||
X3	|||||	|||||||||||||||||		|||||||||	||||||||||||||
X4	|||||||||||||||	||||||||||||	|||||||||		||||||||||||||||
X5	||||||||||||||||	|||||||||||||||	|||||||||||||||	||||||||||||||||	
X6	||||||||||	|||||||||||||||||	|||||||||||||||||	||||||||||||	||||||||||||||||||

Phi=0.735970 Log(Det|R|)=-22.779188 Bartlett Test=596.06 DF=15 Prob=0.000000

Bar Chart of Absolute Correlation Section

	Variables
Variables	**X6**
X1	||||||||||
X2	||||||||||||||||||
X3	||||||||||||||||
X4	||||||||||||
X5	||||||||||||||||
X6	

Phi=0.735970 Log(Det|R|)=-22.779188 Bartlett Test=596.06 DF=15 Prob=0.000000

The report gives the correlations for a test of the overall correlation structure in the data. In this example, we notice several high correlation values. The Gleason-Staelin redundancy measure, phi, is 0.736, which is quite large. There is apparently some correlation structure in this data set that can be modeled. If all the correlations were small, there would be no need for a PCA.

Correlations

The simple correlations between each pair of variables. Note that using the missing value imputation or robust estimation options will affect the correlations in this report. When the above options are not used, the correlations are constructed from those observations having no missing values in any of the specified variables.

Phi

This is the Gleason-Staelin redundancy measure of how interrelated the variables are. A zero value of φ means that there is no correlation among the variables, while a value of one indicates perfect correlation among the variables. This coefficient may have a value less than 0.5 even when there is obvious structure in the data, so care should to be taken when using it. This statistic is especially useful for comparing two or more sets of data. The formula for computing φ is:

$$\varphi = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 - p}{p(p-1)}}$$

Log(Det|R|)

This is the log (base e) of the determinant of the correlation matrix. If you used the covariance matrix, this is the log (base e) of the determinant of the covariance matrix.

Bartlett Test, DF, Prob

This is Bartlett's sphericity test (Bartlett, 1950) for testing the null hypothesis that the correlation matrix is an identity matrix (all correlations are zero). If you get a probability (Prob) value greater than 0.05, you should not perform a PCA on the data. The test is valid for large samples (*N>150*). It uses a Chi-square distribution with *p(p-1)/2* degrees of freedom. Note that this test is only available when you analyze a correlation matrix. The formula for computing this test is:

$$\chi^2 = \frac{(11 + 2p - 6N)}{6} \text{Log}_e |R|$$

Bar Chart of Absolute Correlation Section

This chart graphically displays the absolute values of the correlations. It lets you quickly find high and low correlations.

Eigenvalues Section

Eigenvalues

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	4.562633	76.04	76.04	|||||||||||||||
2	1.171509	19.53	95.57	||||
3	0.242834	4.05	99.62	|
4	0.022878	0.38	100.00	|
5	0.000105	0.00	100.00	|
6	0.000041	0.00	100.00	|

Eigenvalue

The eigenvalues. Often, these are used to determine how many factors to retain. (In this example, we would retain the first two eigenvalues.)

When the PCA is run on the correlations, one rule-of-thumb is to retain those factors whose eigenvalues are greater than one. The sum of the eigenvalues is equal to the number of variables. Hence, in this example, the first factor retains the information contained in 4.563 of the original variables.

When the PCA is run on the covariances, the sum of the eigenvalues is equal to the sum of the variances of the variables.

Individual and Cumulative Percents

The first column gives the percentage of the total variation in the variables accounted for by this factor. The second column is the cumulative total of the percentage. Some authors suggest that the user pick a cumulative percentage, such as 80% or 90%, and keep enough factors to attain this percentage.

Scree Plot

This is a rough bar plot of the eigenvalues. It enables you to quickly note the relative size of each eigenvalue. Many authors recommend it as a method of determining how many factors to retain.

The word *scree*, first used by Cattell (1966), is usually defined as the rubble at the bottom of a cliff. When using the scree plot, you must determine which eigenvalues form the “cliff” and which form the “rubble.” You keep the factors that make up the cliff. Cattell and Jaspers (1967) suggest keeping those that make up the cliff plus the first factor of the rubble.

Interpretation of the Example

This table presents the eigenvalues of the correlation (covariance) matrix. The first question that we would ask is how many factors should be kept. The scree plot shows that the first two factors are indeed the largest. The cumulative percentages show that the first two factors account for over 95% of the variation. Only the first two eigenvalues are greater than one. We begin to get the impression that the correct answer is that two factors will adequately approximate these data.

We note in passing that the third and fourth eigenvalues are several orders of magnitude larger than the fifth and sixth. We will keep our eyes on these factors as well. Although they are not significant, they certainly represent some artifact in the data.

Eigenvectors Section

Eigenvectors

	Factors				
Variables	Factor1	Factor2	Factor3	Factor4	Factor5
X1	-0.315300	-0.639819	0.507144	0.437189	0.012097
X2	-0.427719	0.372949	0.077702	0.188932	0.786262
X3	-0.399630	0.476348	0.016354	0.485911	-0.590471
X4	-0.379732	-0.385432	-0.830977	0.126310	0.027636
X5	-0.452963	-0.197773	0.209428	-0.567541	-0.140774
X6	-0.456690	0.192251	0.046011	-0.446098	-0.111393

	Factors
Variables	Factor6
X1	0.206739
X2	-0.134256
X3	-0.168389
X4	-0.002258
X5	-0.608219
X6	0.735489

Bar Chart of Absolute Eigenvectors

	Factors				
Variables	Factor1	Factor2	Factor3	Factor4	Factor5
X1	|||||||	||||||||||||	||||||||||	||||||||		
X2	|||||||||	||||||||	||	||||	||||||||||||||||
X3	|||||||||	|||||||||			|||||||||	||||||||||||
X4	|||||||||	||||||||	||||||||||||||||	|||		
X5	||||||||||	||||	|||||	||||||||||||	|||
X6	||||||||||	||||			|||||||||	|||

Bar Chart of Absolute Eigenvectors

	Factors
Variables	Factor6
X1	|||||
X2	|||
X3	||||
X4		
X5	||||||||||||
X6	||||||||||||||

Eigenvector

The eigenvectors are the weights that relate the scaled original variables, $x_i = (X_i - Mean_i)/Sigma_i$, to the factors. For example, the first factor, $Factor_1$, is the weighted average of the scaled variables, the weight of each variable given by the corresponding element of the first eigenvector. Mathematically, the relationship is given by:

$$Factor_1 = v_{11}x_{11} + v_{12}x_{12} + ... + v_{1p}x_{1p}$$

These coefficients may be used to determine the relative importance of each variable in forming the factor. Often, the eigenvectors are scaled so that the variances of the factor scores are equal to one. These scaled eigenvectors are given in the *Score Coefficients* section described later.

Bar Chart of Absolute Eigenvectors

This chart graphically displays the absolute values of the eigenvectors. It lets you quickly interpret the eigenvector structure. By looking at which variables correlate highly with a factor, you can determine what underlying structure it might represent.

Factor Loadings Section

Factor Loadings

Variables	Factors Factor1	Factor2	Factor3	Factor4	Factor5
X1	-0.673491	-0.692517	0.249911	0.066127	0.000124
X2	-0.913622	0.403667	0.038290	0.028577	0.008051
X3	-0.853623	0.515581	0.008059	0.073497	-0.006046
X4	-0.811120	-0.417177	-0.409490	0.019105	0.000283
X5	-0.967543	-0.214062	0.103202	-0.085844	-0.001441
X6	-0.975505	0.208086	0.022673	-0.067475	-0.001141

Variables	Factors Factor6
X1	0.001326
X2	-0.000861
X3	-0.001080
X4	-0.000014
X5	-0.003900
X6	0.004716

Bar Chart of Absolute Factor Loadings

Variables	Factors Factor1	Factor2	Factor3	Factor4	Factor5
X1	|||||||||||||	|||||||||||||	|||||	||	|
X2	||||||||||||||||||	||||||||	|	|	|
X3	|||||||||||||||||	|||||||||	|	||	|
X4	||||||||||||||||	||||||||	||||||||	|	|
X5	|||||||||||||||||||	|||||	|||	||	|
X6	|||||||||||||||||||	|||||	|	||	|

Bar Chart of Absolute Factor Loadings

Variables	Factors Factor6
X1	|
X2	|
X3	|
X4	|
X5	|
X6	|

Factor Loadings

These are the correlations between the variables and factors.

Bar Chart of Absolute Factor Loadings

This chart graphically displays the absolute values of the factor loadings. It lets you quickly interpret the correlation structure. By looking at which variables correlate highly with a factor, you can determine what underlying structure it might represent.

Interpretation of the Example

We now go through the interpretation of each factor. Factor one appears to be an average of all six variables. Although the weights of all variables are large, the weights on X5 and X6 are the largest. Factor two appears to be a contrast (difference) of X2+X3 and X1+X4. Factor three is most highly correlated to X4. Factor four appears to be associated with several variables, but most highly with X5. Factor five is a contrast of X2 and X3. Factor six is a contrast of X5 and X6. If these data were real, we could try to attach meaning to these patterns.

Communality Section

Communalities

Variables	Factors Factor1	Factor2	Factor3	Factor4	Factor5
X1	0.453590	0.479580	0.062456	0.004373	0.000000
X2	0.834705	0.162947	0.001466	0.000817	0.000065
X3	0.728671	0.265824	0.000065	0.005402	0.000037
X4	0.657916	0.174037	0.167682	0.000365	0.000000
X5	0.936140	0.045823	0.010651	0.007369	0.000002
X6	0.951610	0.043300	0.000514	0.004553	0.000001

Variables	Factors Factor6	Communality
X1	0.000002	1.000000
X2	0.000001	1.000000
X3	0.000001	1.000000
X4	0.000000	1.000000
X5	0.000015	1.000000
X6	0.000022	1.000000

Bar Chart of Communalities

Variables	Factors Factor1	Factor2	Factor3	Factor4	Factor5
X1	|||||||||	|||||||||	||	|	|
X2	||||||||||||||||	||||	|	|	|
X3	||||||||||||||	||||||	|	|	|
X4	|||||||||||||	||||	||||	|	|
X5	|||||||||||||||||||	|	|	|	|
X6	|||||||||||||||||||	|	|	|	|

Variables	Factors Factor6	Communality
X1	|	||||||||||||||||||
X2	|	||||||||||||||||||
X3	|	||||||||||||||||||
X4	|	||||||||||||||||||
X5	|	||||||||||||||||||
X6	|	||||||||||||||||||

Communality

The communality is the proportion of the variation of a variable that is accounted for by the factors that are retained. It is the R^2 value that would be achieved if this variable were regressed on the retained factors. This table value gives the amount added to the communality by each factor.

Bar Chart of Communalities

This chart graphically displays the values of the communalities.

Factor Structure Summary Section

Factor Structure Summary Section

Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
X6	X1	X4			
X5	X3				
X2	X4				
X3	X2				
X4					
X1					

Interpretation

This report is provided to summarize the factor structure. Variables with an absolute loading greater than the amount set in the *Minimum Loading* option are listed under each factor. Using this report, you can quickly see which variables are related to each factor. Notice that it is possible for a variable to have high loadings on several factors.

Score Coefficients Section

Score Coefficients

	Factors				
Variables	**Factor1**	**Factor2**	**Factor3**	**Factor4**	**Factor5**
X1	-0.1476102	-0.5911322	1.029146	2.890409	1.181451
X2	-0.20024	0.3445697	0.1576812	1.249092	76.79078
X3	-0.1870899	0.4401002	3.318755E-02	3.212524	-57.6688
X4	-0.1777746	-0.356102	-1.686299	0.8350784	2.699127
X5	-0.2120581	-0.1827235	0.4249913	-3.752211	-13.7488
X6	-0.2138031	0.1776219	9.337004E-02	-2.949311	-10.87929

Score Coefficients

	Factors
Variables	**Factor6**
X1	32.24512
X2	-20.93996
X3	-26.26369
X4	-0.3522398
X5	-94.86387
X6	114.7142

Score Coefficients

These are the coefficients that are used to form the factor scores. The factor scores are the values of the factors for a particular row of data. These score coefficients are similar to the eigenvectors. They have been scaled so that the scores produced have a variance of one rather than a variance equal to the eigenvalue. This causes each of the factors to have the same variance.

You would use these scores if you wanted to calculate the factor scores for new rows not included in your original analysis.

Residual Section

Residual Section

Row	T2	T2 Prob	Q0	Q1	Q2	Q3	Q5
1	4.68	0.6932	10.68	0.91	0.11	0.00	0.00
2	27.94*	0.0078	12.76	0.66	0.65	0.57*	0.00
3	28.02*	0.0077	12.93	6.84*	6.23*	0.01	0.00
4	2.25	0.9250	3.04	1.61	0.06	0.00	0.00
5	8.20	0.3742	1.53	0.95	0.01	0.00	0.00*
6	4.20	0.7427	4.52	0.23	0.01	0.01	0.00
7	1.33	0.9785	1.86	0.01	0.00	0.00	0.00
8	3.06	0.8573	5.47	2.29	0.07	0.00	0.00

(report continues through all thirty rows)

This report is useful for detecting outliers--observations that are very different from the bulk of the data. To do this, two quantities are displayed: T^2 and Q_k. We will now define these two quantities.

T^2 measures the combined variability of all the variables in a single observation. Mathematically, T^2 is defined as:

$$T^2 = [\underline{x} - \underline{\bar{x}}]' S^{-1} [\underline{x} - \underline{\bar{x}}]$$

where $\underline{x}$ represents a *p*-variable observation, $\overline{\underline{x}}$ represents the *p*-variable mean vector and S^{-1} represents the inverse of the covariance matrix.

T is not affected by a change in scale. It is the same whether the analysis is performed on the covariance or the correlation matrix. T^2 gives a scaled distance measure of an individual observation from the overall mean. The closer an observation is to its mean, the smaller will be the value of T^2.

If the variables follow a multivariate normal distribution, then the probability distribution of T^2 may be related to the common F distribution using the formula:

$$T^2_{p,n,\alpha} = \frac{p(n-1)}{n-p} F_{p,n-p,\alpha}$$

Using this relationship, we can perform a statistical test at a given level of significance to determine if the observation is significantly different from the vector of means. You set the α value using the *Alpha* option. Since this test is being performed *N* times, you would anticipate about *N(1-α)* observations to be significant by chance variation. In our current example, rows two and three are starred (which means they were significant at the .05 significance level). You would probably want to check for data entry or transcription errors. (Of course, in this data set, these rows were made to be outliers.)

T^2 is really not part of a normal PCA since it may be calculated independently. It is presented to help detect observations that may have an undue influence on the analysis. You can read more about its use and interpretation in Jackson (1991).

The other quantity shown on this report is Q_k. Q_k represents the sum of squared residuals when an observation is predicted using the first *k* factors. Mathematically, the formula for Q_k is:

$$\begin{aligned} Q_k &= (\underline{x} - \underline{\hat{x}})'(\underline{x} - \underline{\hat{x}}) \\ &= \sum_{i=1}^{p} \left(x_i - {}_k\hat{x}_i\right)^2 \\ &= \sum_{i=k+1}^{p} \lambda_i \left(pc_i\right)^2 \end{aligned}$$

Here ${}_k x_i$ refers to the value of variable *i* predicted from the first *k* factors, λ_i refers to the i^{th} eigenvalue, and pc_i is the score of the i^{th} factor for this particular observation. Further details are given in Jackson (1991) on pages 36 and 37.

An upper limit for Q_k is given by the formula:

$$Q_\alpha = a\left[\frac{z_\alpha\sqrt{2bh^2}}{a} + \frac{bh(h-1)}{a^2} + 1\right]^{1/h}$$

where

$$a = \sum_{i=k+1}^{p} \lambda_i$$

$$b = \sum_{i=k+1}^{p} \lambda_i^2$$

$$c = \sum_{i=k+1}^{p} \lambda_i^3$$

$$h = 1 - \frac{2ac}{3b^2}$$

and z_α is the upper normal deviate of area α if h is positive or the lower normal deviate of area α if h is negative.

This limit is valid for any value of k, whether too many or too few factors are kept. Note that these formulas are for the case when the correlation matrix is being used. When the analysis is being run on the covariance matrix, the pc_i's must be adjusted. Further details are given in Jackson (1991).

Notice that significant (starred) values of Q_k indicate observations that are not duplicated well by the first k factors. These should be checked to see if they are valid. Q_k and T^2 provide an initial data screening tool.

Interpretation of the Example

We are interested in two columns in this report: Q2 and T2. Notice that rows two and three are significantly large (shown by the asterisk) for both measurements. If these were real data, we would investigate these two rows very carefully. We would first check for data entry errors and next for errors that might have occurred when the measurements were actually taken. In our case, we know that these two rows are outliers (since they were artificially made to be outliers).

Factor Scores Section

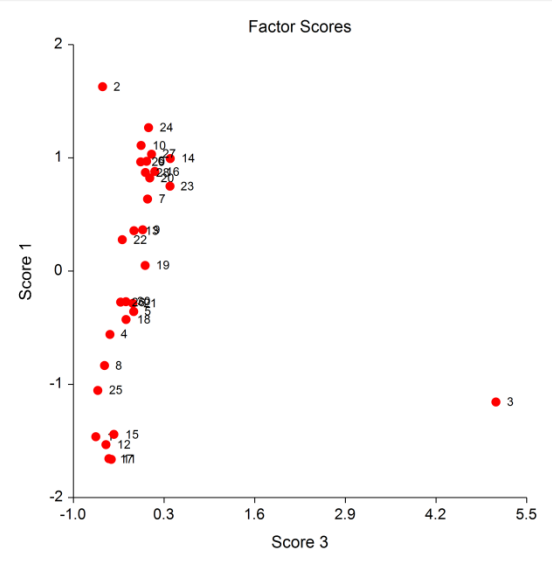
Factor Score Factors

Row	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
1	-1.4627	0.8272	-0.6797	-0.1124	1.1732	0.0689
2	1.6286	0.0834	-0.5825	-4.9911	-0.0743	0.1499
3	-1.1560	0.7225	5.0582	-0.7581	-0.0226	0.0079
4	-0.5595	1.1520	-0.4768	0.0670	0.5125	0.3465
5	-0.3573	-0.8963	-0.1360	0.2037	-1.6830	2.0931
6	0.9696	0.4329	0.0488	0.6208	-1.6155	-0.2796
7	0.6364	-0.0783	0.0624	0.4003	-0.8677	-0.0678
8	-0.8339	1.3759	-0.5545	-0.0806	-0.3899	-0.0447

(report continues through all thirty rows)

This report presents the individual factor scores scaled so each column has a mean of zero and a standard deviation of one. These are the values that are plotted in the plots to follow. Remember, there is one row of score values for each observation and one column for each factor that was kept.

Factor Score Plots

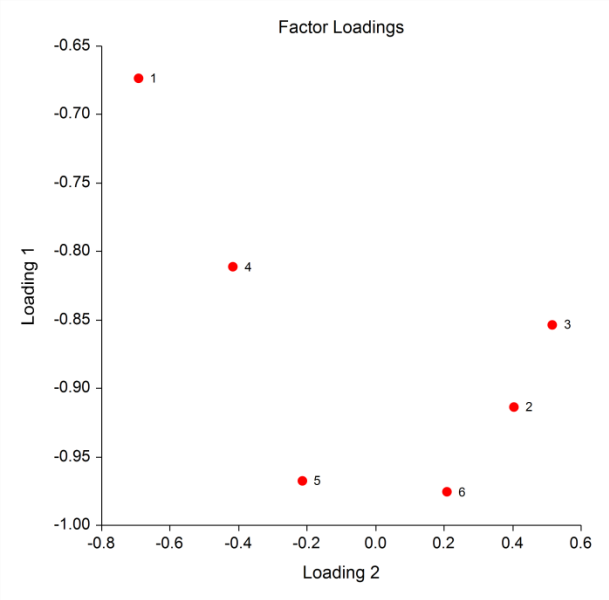


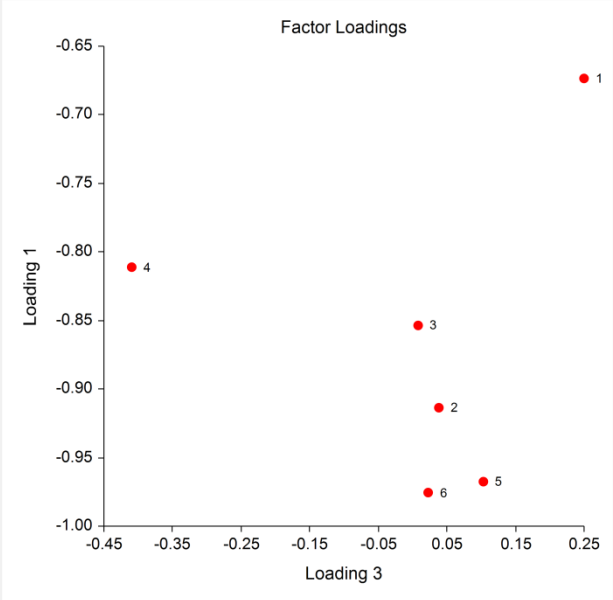
This set of plots shows each factor plotted against every other factor. The first k factors (where k is the number of large eigenvalues) usually show the major structure that will be found in the data. The rest of the factors show outliers and linear dependencies. Note that in our present example, outliers are displayed in both plots that include factor three. We would now investigate these rows much more closely.

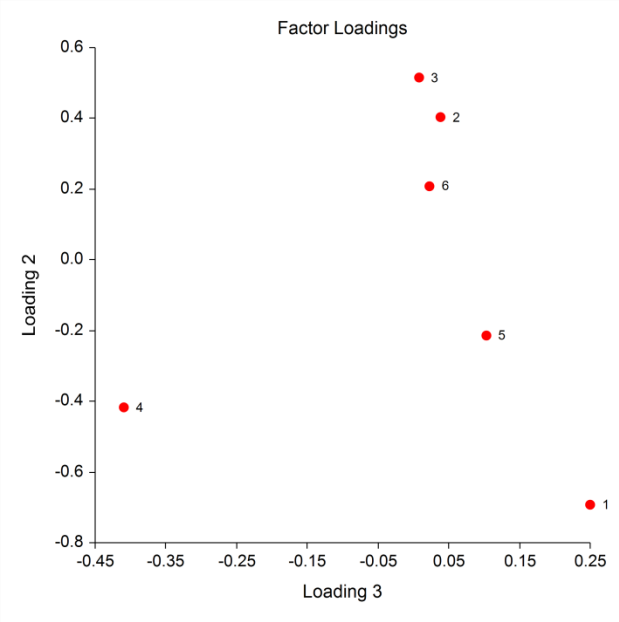
Interpretation of the Example

We first notice the presence of an outlier in plots containing factor three. If we were not convinced before that this row was an outlier, we would be now. Factor three fits row three. If we had called for the plot of factor four, we would see that it fits row four. Hence, these plots of nonsignificant factors show outliers.

Factor Loading Plots







Discussion of Factor Loading Plots

This set of plots shows each of the factor loading columns plotted against each other. The data points represent variables. The plot allows you to find variables that are highly correlated with both factors. It is anticipated that this will aid in the interpretation of the factors.

Robust and Missing-Value Iteration Section

The following report is not part of the preceding tutorial. We have re-run the problem calling for robust estimation so that we could show you this iteration report. We have set the number of robust iterations at six.

Robust and Missing-Value Estimation Iteration Section

No.	Count	Trace of Covar Matrix	Percent Change
0	30	4907.795	0.00
1	30	4907.795	0.00
2	30	4423.718	-9.86
3	30	4423.718	0.00
4	30	4353.748	-1.58
5	30	4353.748	0.00
6	30	4335.77	-0.41

This report presents the progress of the iterations. The trace of the covariance matrix gives a measure of what is happening at each iteration. When this value stabilizes, the program has converged. The percent change is reported to let you determine how much the trace has changed. In this particular example, we see very little change between iterations five and six. We would feel comfortable in stopping at this point. A look at the Descriptive Statistics section will let you see how much the means and standard deviations have changed.

A look at the Residual Section will let you see the robust weights that are assigned to each row. Those weights that are near zero indicate observations whose influence has been removed by the robust procedure.