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 (*components* or *pc's*). These components 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.

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 principal components, latent variables, or factors. 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 components 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 components 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 component, y_i, is one.

The correlation between an i^{th} component 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 component loadings and are provided in the *Component 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 other 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 $E\left(x_{ij} | x_{obs,i}, \hat{\mu}, S^{(t)}\right)$ 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 Components

Several methods have been proposed for determining the number of components 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 components associated with the relatively small eigenvalues, so these should be investigated as well.

Kaiser (1960) proposed dropping components 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 components to be

dropped. However, if the largest components 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 components, 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 components so that this variation is achieved. Usually, however, this cutoff percentage is used as a lower limit. That is, if the designated number of components 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 components 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, components during the early stages of the analysis. If you want to reduce the dimensionality of your database, then you should keep enough components 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 components 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 components, you will obtain new components that are each highly correlated with only a few of the original variables. This simplifies the interpretation of the component 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 component.

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 components. 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 components. Instead, it tends to form one large component 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 component 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 components. 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 components**.** You could then return to the original data to analyze the component 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 components that are kept. Instead of trying to interpret each component, McCabe (1984) has suggested finding the principal variables. Suppose you start with p variables, run a PCA, and decide to retain k components. 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 components. The interpretation of the variables is much easier than the interpretation of the components.

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 component 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 component 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 component scores. Since the component 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 components, 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

Example 1 – Principal Components Analysis

Even though we go directly into running PCA here, it is important to realize that the first step in any real PCA is to investigate all appropriate graphics. In this case, we recommend that you run NCSS's *Scatter Plot Matrix* procedure which will allow you to look at all individual, pairwise scatter plots quickly and easily. Only then should you begin an analysis.

This example shows the basics of how to conduct a principal components analysis. The data used are found in the *Death Rates – States – 2016* dataset. This dataset presents state-by-state mortality rates for various causes of death in 2016. The dataset was obtained from the National Center for Health Statistics.

The example will allow us to document the various tables that are available. This example will perform a PCA of variables Alzheimer - Accidents. It will do a robust adjustment and correct for any missing values.

Setup

To run this example, complete the following steps:

1 Open the Death Rates – States – 2016 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Death Rates – States – 2016** and click **OK**.

2 Specify the Principal Components Analysis procedure options

- Find and open the **Principal Components Analysis** procedure using the menus or the Procedure Navigator.
- Set **Variable Labels** to **Column Names** using the **Report Options** dropdown in the toolbar.
- The settings for this example are listed below and are stored in the **Example 1** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Variables	**Alzheimers, LowResDis, Cancer, Diabetes, HeartDis, FluPneum, Kidney, Stroke, Suicide, Accidents**
Data Label Variable	**State**
Robust Covariance Matrix Estimation	**Checked**
Reports Tab	
All Reports and Plots	**Checked** (Normally you would only view a few of these reports, but we are selecting them all so that we can document them.)
Report Options (*in the Toolbar*)	
Variable Labels	**Column Names**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Robust and Missing-Value Estimation Iteration

Robust and Missing-Value Estimation Iteration
Robust Estimation: Iterations = 6 and Weight = 4
Missing-Value Estimation: Average

Iteration Number	Count	Trace of Covar Matrix	Percent Change
0	52	1421.12	0.00
1	52	1421.12	0.00
2	52	1152.677	-18.89
3	52	1152.677	0.00
4	52	1108.958	-3.79
5	52	1108.958	0.00
6	52	1098.059	-0.98

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 algorithm 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.

Iteration Number

This is the iteration number. The number of iterations was set by the *Maximum Iterations* option.

Count

The number of observations that were used.

Trace of Covar Matrix

This is the sum of the variances of the covariance matrix. We want these values as small as possible.

Percent Change

This provides a percentage of how much each iteration improves the model. The largest change occurred at iteration 2. After iteration 4, little changes.

Descriptive Statistics

Descriptive Statistics
Robust Estimation: Iterations = 6 and Weight = 4
Missing-Value Estimation: Average

Variables	Count	Mean	Standard Deviation	Communality
Alzheimers	52	31.35265	7.513908	1.000000
LowResDis	52	42.99033	8.751616	1.000000
Cancer	52	158.9506	12.3722	1.000000
Diabetes	52	21.25077	3.470444	1.000000
HeartDis	52	165.4697	25.29784	1.000000
FluPneum	52	13.26525	2.908935	1.000000
Kidney	52	13.29424	4.095566	1.000000
Stroke	52	37.07346	5.490243	1.000000
Suicide	52	15.14693	3.963418	1.000000
Accidents	52	51.6677	9.424813	1.000000

This report lets us compare the relative sizes of the standard deviations. Since the robust estimation and missing-value imputation options were selected, the descriptive statistics shown here are adjusted estimates.

Count, Mean, and Standard Deviation

These are the familiar (robust) summary statistics of each variable. They allow you to check that you have specified the correct variables. Note that using missing value imputation or robust estimation will change these values from their standard values.

Communality

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

Correlations

Correlations

Variables	Variables Alzheimers	LowResDis	Cancer	Diabetes	HeartDis	FluPneum
Alzheimers	1.000000	0.481162	0.328182	0.508583	0.348971	0.125531
LowResDis	0.481162	1.000000	0.658578	0.675307	0.618861	0.337299
Cancer	0.328182	0.658578	1.000000	0.554234	0.733415	0.376618
Diabetes	0.508583	0.675307	0.554234	1.000000	0.499199	0.338081
HeartDis	0.348971	0.618861	0.733415	0.499199	1.000000	0.556298
FluPneum	0.125531	0.337299	0.376618	0.338081	0.556298	1.000000
Kidney	0.245905	0.426098	0.615882	0.384131	0.650337	0.539397
Stroke	0.650977	0.574768	0.562345	0.557224	0.646170	0.459609
Suicide	0.195359	0.521825	-0.034368	0.296968	0.032518	-0.001083
Accidents	0.078286	0.483689	0.540200	0.427036	0.334065	0.139587

Variables	Variables Kidney	Stroke	Suicide	Accidents
Alzheimers	0.245905	0.650977	0.195359	0.078286
LowResDis	0.426098	0.574768	0.521825	0.483689
Cancer	0.615882	0.562345	-0.034368	0.540200
Diabetes	0.384131	0.557224	0.296968	0.427036
HeartDis	0.650337	0.646170	0.032518	0.334065
FluPneum	0.539397	0.459609	-0.001083	0.139587
Kidney	1.000000	0.683541	-0.151774	0.227860
Stroke	0.683541	1.000000	0.102694	0.132398
Suicide	-0.151774	0.102694	1.000000	0.398038
Accidents	0.227860	0.132398	0.398038	1.000000

Phi=0.453362 Log(Det|R|)=-6.638676 Bartlett Test=310.91 DF=45 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 0.4533, which is moderately 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 Correlations

Bar Chart of Absolute Correlations

Variables	Variables Alzheimers	LowResDis	Cancer	Diabetes	HeartDis	FluPneum
Alzheimers		|||||||	|||||||	||||||||	|||
LowResDis	||||||||||		|||||||||||||	||||||||||||	||||||||	|||||||
Cancer	|||||||	||||||||||||||	||||||||||||	||||||||||||||	||||||||
Diabetes	||||||||||	||||||||||||||	||||||||||||	|||||||||	|||||||
HeartDis	|||||||	||||||||||||	|||||||||||||||	|||||||||		||||||||||||
FluPneum	|||	|||||||	|||||||	|||||||	||||||||||||	
Kidney	|||||	||||||||	||||||||||||	|||||||	||||||||||||||	||||||||||
Stroke	||||||||||||||	||||||||||||	||||||||||||	||||||||||||	||||||||||||	|||||||||
Suicide	||||	||||||||||			||||||			
Accidents	||	|||||||||	||||||||||	|||||||||	|||||||	|||

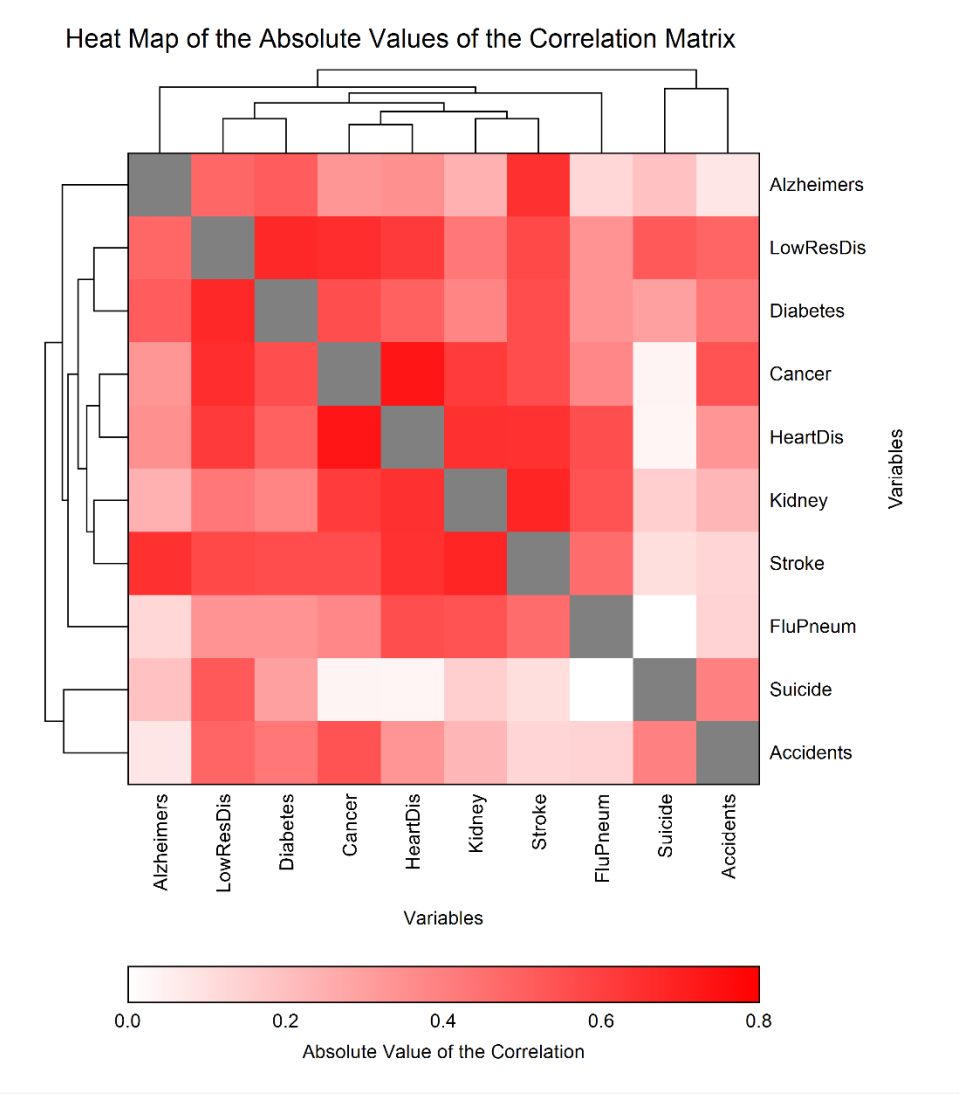
Variables	Variables Kidney	Stroke	Suicide	Accidents
Alzheimers	|||||	||||||||||||||	||||	||
LowResDis	|||||||||	||||||||||||	||||||||||	|||||||||
Cancer	||||||||||||||	||||||||||||		|||||||||||
Diabetes	|||||||	||||||||||||	||||||	|||||||||
HeartDis	||||||||||||||	||||||||||||		|||||||
FluPneum	||||||||||	|||||||||		|||
Kidney		||||||||||||||	||||	|||||
Stroke	||||||||||||||		|||	|||
Suicide	||||	|||		||||||||
Accidents	|||||	|||	||||||||	

Phi=0.453362 Log(Det|R|)=-6.638676 Bartlett Test=310.91 DF=45 Prob=0.000000

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

Heat Map of the Absolute Values of the Correlation Matrix

Heat Map of the Absolute Values of the Correlation Matrix



This report displays a heat map of the adjusted correlation matrix. Note that the rows and columns are sorted in the order suggested by a hierarchical clustering of the correlation matrix.

This plot allows you to discover various subsets of the variables that seem to be highly correlated. You can see that the diseases LowResDis, Diabetes, Cancer, Heart Disease, Kidney disease, Stroke, and FluPneum are highly correlated. Similarly, Suicides and Accidents seem to be correlated.

This plot was suggested by Friendly (2002) and Friendly and Kwan (2003). It actually does not involve the PCA results. It is presented to get you acquainted with the data.

Cluster Detail Report

Cluster Detail Report for the Absolute Values of the Correlation Matrix
Clustering Method Group Average

Cluster	Variables in this Cluster
1	Suicide, Accidents
2	LowResDis, Cancer, Diabetes, HeartDis, FluPneum, Kidney, Stroke
None	Alzheimers

This report displays the results of a hierarchical cluster analysis of the variables. It lists the variables contained in the clusters. Those variables that cannot be classified are listed in the “None” cluster.

Linkage Report

Linkage Report for the Absolute Values of the Correlation Matrix
Clustering Method Group Average

Link	Number Clusters	Distance Value	Distance Bar
9	1	0.768768	||||||||||||||||||||||||||||
8	2	0.615813	||||||||||||||||||||||
7	3	0.601962	|||||||||||||||||||||
6	4	0.565450	||||||||||||||||||||
5	5	0.465863	||||||||||||||||
4	6	0.381316	||||||||||||||
3	7	0.324693	||||||||||||
2	8	0.316459	|||||||||||
1	9	0.266585	|||||||||

Cophenetic Correlation	0.756968
Delta(0.5)	0.172485
Delta(1.0)	0.216258

This report displays the number of clusters that exist at each link. The links are displayed in reverse order so that you can quickly determine an appropriate number of clusters to use. It displays the distance level at which the fusion took place. It will let you precisely determine the best value of the number of clusters.

The cophenetic correlation and two delta goodness of fit statistics are reported at the bottom of this report. These values compare the fit of various cluster configurations.

Link

This is the sequence number of the fusion.

Number Clusters

This is the number of clusters that would result if the cluster cutoff value were set to the corresponding Distance Value or higher. Note that this number includes outliers.

Distance Value

This is distance value between the two joining clusters that is used by the algorithm. Normally, this value is monotonically increasing. When backward linking occurs, this value will no longer exhibit a strictly increasing behavior.

Distance Bar

This is a bar graph of the Distance Values. Choose the number of clusters by finding a jump in the decreasing pattern shown in this bar chart.

Cophenetic Correlation

This is the Pearson correlation between the actual distances and the predicted distances based on this particular hierarchical configuration. A value of 0.75 or above needs to be achieved in order for the clustering to be considered useful.

Delta (0.5, 1)

These are the values of the goodness of fit deltas. When comparing to clustering configurations, the configuration with the smallest delta value fits the data better.

Eigenvalues

Eigenvalues

Robust Estimation: Iterations = 6 and Weight = 4
Missing-Value Estimation: Average

Number	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	4.843781	48.44	48.44	|||||||||
2	1.631417	16.31	64.75	||||
3	1.122669	11.23	75.98	|||
4	0.750564	7.51	83.48	||
5	0.451893	4.52	88.00	|
6	0.383123	3.83	91.83	|
7	0.341897	3.42	95.25	|
8	0.218917	2.19	97.44	|
9	0.162260	1.62	99.07	|
10	0.093478	0.93	100.00	|

Eigenvalue

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

When the PCA is run on the correlations, one rule-of-thumb is to retain those components whose eigenvalues are greater than one. The sum of the eigenvalues is equal to the number of variables. Hence, in this example, the first component retains the information contained in 4.844 of the original six variables. This represents 48.44% of the variation in the data.

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 component. 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 components 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 components 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 components that make up the cliff. Cattell and Jaspers (1967) suggest keeping those that make up the cliff plus the first component 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 components should be kept. The scree plot shows that the first four, or even five, components are needed. The cumulative percentages show that the first three components account for over 76% of the variation. Only the first three eigenvalues are greater than one. We decide to look at the first three components.

Eigenvectors

Eigenvectors

	Components					
Variables	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Alzheimers	-0.262945	0.107263	-0.657130	-0.224078	-0.080610	0.085875
LowResDis	-0.378847	0.287112	-0.007869	0.058632	0.183915	-0.360938
Cancer	-0.374378	-0.090726	0.269319	-0.406199	0.018514	-0.252261
Diabetes	-0.348820	0.198532	-0.107873	-0.075675	-0.716916	0.070522
HeartDis	-0.378947	-0.187392	0.127786	0.024211	0.187116	-0.514084
FluPneum	-0.263096	-0.294449	0.158806	0.713201	-0.364717	0.010449
Kidney	-0.329981	-0.374817	0.107777	0.019920	0.333267	0.519371
Stroke	-0.371317	-0.160716	-0.360854	0.039330	0.236756	0.264152
Suicide	-0.112328	0.657616	-0.047248	0.457507	0.325770	0.062863
Accidents	-0.230868	0.371244	0.546202	-0.234795	-0.063950	0.431131

	Components			
Variables	**PC7**	**PC8**	**PC9**	**PC10**
Alzheimers	0.525764	-0.105735	-0.350204	0.114894
LowResDis	-0.249717	-0.470917	-0.189500	-0.535173
Cancer	0.055684	-0.385315	0.320172	0.544079
Diabetes	-0.488838	0.238152	-0.023773	0.079755
HeartDis	0.147047	0.670488	-0.183568	-0.014087
FluPneum	0.341566	-0.243342	0.030894	0.015502
Kidney	-0.335321	-0.056892	-0.462606	0.167684
Stroke	-0.042816	0.143828	0.697453	-0.264086
Suicide	-0.056417	0.107822	0.057063	0.463935
Accidents	0.404231	0.132492	0.004340	-0.290642

Bar Chart of Absolute Eigenvectors

	Components						
Variables	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**
Alzheimers	||||||	|||	||||||||||||	|||||	||	||	||||||||||
LowResDis	||||||||	||||||	|	||	||||	||||||||	|||||
Cancer	|||||||	||	||||||	||||||||	|	||||||	||
Diabetes	|||||||	||||	|||	||	|||||||||||||	||	|||||||||
HeartDis	||||||||	||||	|||	|	||||	||||||||||	|||
FluPneum	||||||	||||||	||||	||||||||||||||	||||||||	|	|||||||
Kidney	|||||||	||||||||	|||	|	|||||||	||||||||||	|||||||
Stroke	||||||||	||||	||||||||	|	|||||	||||||	|
Suicide	|||	||||||||||||	|	|||||||||	|||||||	||	||
Accidents	|||||	||||||||	||||||||||	|||||	||	||||||||	||||||||

	Components		
Variables	**PC8**	**PC9**	**PC10**
Alzheimers	|||	||||||||	|||
LowResDis	|||||||||	||||	||||||||||
Cancer	||||||||	|||||||	||||||||||
Diabetes	|||||	|	||
HeartDis	||||||||||||	||||	|
FluPneum	|||||	|	|
Kidney	||	|||||||||	||||
Stroke	|||	||||||||||||	||||||
Suicide	|||	||	|||||||||
Accidents	|||	|	||||||

Eigenvector

The eigenvectors are the weights that relate the scaled original variables, $x_i = (X_i - Mean_i)/Sigma_i$, to the components. For example, the first component, PC_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:

$$PC_1 = v_{11}x_{11} + v_{12}x_{12} + \cdots + v_{1p}x_{1p}$$

These coefficients may be used to determine the relative importance of each variable in forming the component. Often, the eigenvectors are scaled so that the variances of the component 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 component, you can determine what underlying structure it might represent.

Component Loadings

Component Loadings

	Components					
Variables	PC1	PC2	PC3	PC4	PC5	PC6
Alzheimers	-0.578705	0.137004	-0.696269	-0.194130	-0.054188	0.053154
LowResDis	-0.833788	0.366720	-0.008338	0.050796	0.123633	-0.223409
Cancer	-0.823952	-0.115881	0.285360	-0.351911	0.012445	-0.156142
Diabetes	-0.767703	0.253579	-0.114298	-0.065561	-0.481932	0.043651
HeartDis	-0.834008	-0.239350	0.135397	0.020975	0.125785	-0.318203
FluPneum	-0.579037	-0.376091	0.168264	0.617883	-0.245174	0.006468
Kidney	-0.726241	-0.478742	0.114196	0.017258	0.224032	0.321475
Stroke	-0.817217	-0.205278	-0.382346	0.034073	0.159154	0.163502
Suicide	-0.247218	0.839953	-0.050062	0.396362	0.218993	0.038911
Accidents	-0.508108	0.474179	0.578734	-0.203415	-0.042989	0.266857

	Components			
Variables	PC7	PC8	PC9	PC10
Alzheimers	0.307425	-0.049472	-0.141067	0.035128
LowResDis	-0.146014	-0.220336	-0.076333	-0.163625
Cancer	0.032560	-0.180283	0.128970	0.166348
Diabetes	-0.285833	0.111428	-0.009576	0.024385
HeartDis	0.085981	0.313712	-0.073944	-0.004307
FluPneum	0.199721	-0.113856	0.012445	0.004740
Kidney	-0.196069	-0.026619	-0.186345	0.051268
Stroke	-0.025035	0.067295	0.280945	-0.080742
Suicide	-0.032988	0.050449	0.022986	0.141844
Accidents	0.236362	0.061991	0.001748	-0.088862

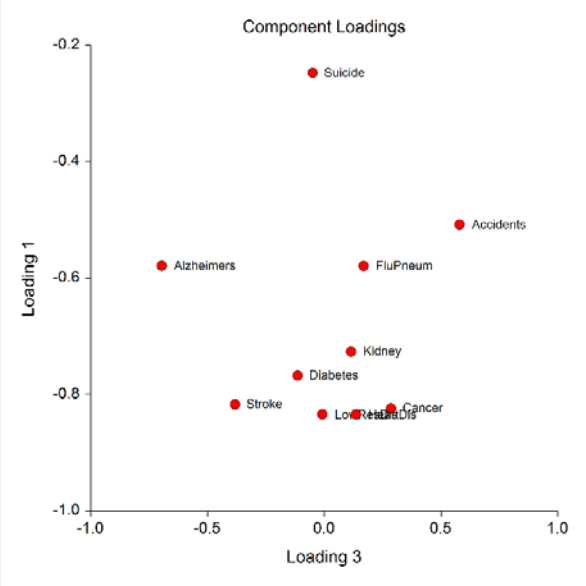
Bar Chart of Absolute Component Loadings

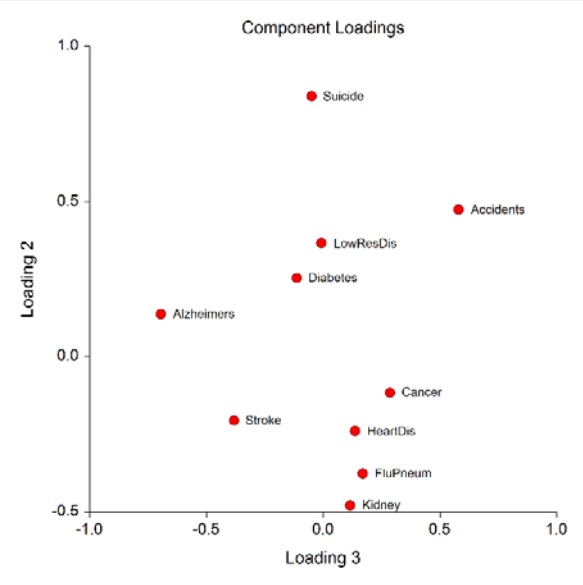
	Components							
Variables	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Alzheimers	|||||||||||	|||	||||||||||||||	||	||	|||||||	
LowResDis	||||||||||||||||	||||||||	||	|||	|||||	|||	|||||
Cancer	||||||||||||||||	|||	||||||	||||||||			||||			||||
Diabetes	||||||||||||||	||||||	|||	||	||||||||||			||||||	|||
HeartDis	||||||||||||||||	|||||	|||			|||	|||||||	||	|||||||
FluPneum	||||||||||	||||||||	||||	||||||||||||	|||||			||||	|||
Kidney	||||||||||||||	|||||||||	|||			|||||	|||||||	||||		
Stroke	||||||||||||||||	|||||	||||||||			||||	||||			||
Suicide	|||||	||||||||||||||||	||	||||||||	|||||				||
Accidents	||||||||||	||||||||||	||||||||||	|||||			|||||||	|||||	||

Principal Components Analysis

Variables	Components PC9	PC10
Alzheimers	|||	
LowResDis	||	||||
Cancer	|||	||||
Diabetes				
HeartDis	||		
FluPneum				
Kidney	||||	||
Stroke	||||||	||
Suicide			|||
Accidents			||

Component Loadings Plot(s)





Component Loadings

The loadings are the correlations between the original variables and constructed principal components.

Bar Chart of Absolute Component Loadings

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

Component Loadings Scatter Plots

These plots display scatter plots of the loading values. The points on the plots represent the variables. These plots allow you to see which variables are similar and which are different.

Interpretation

The numeric reports allows us a quick general impression of the magnitudes of the loadings. However, the plots allow many conclusions to be drawn.

1. Heart Disease, Stroke, and Cancer are highly correlated. With Diabetes and LowResDis, they appear to be related to PC1.
2. Suicide are highly correlated to PC2.
3. Accidents are highly correlated to PC3.

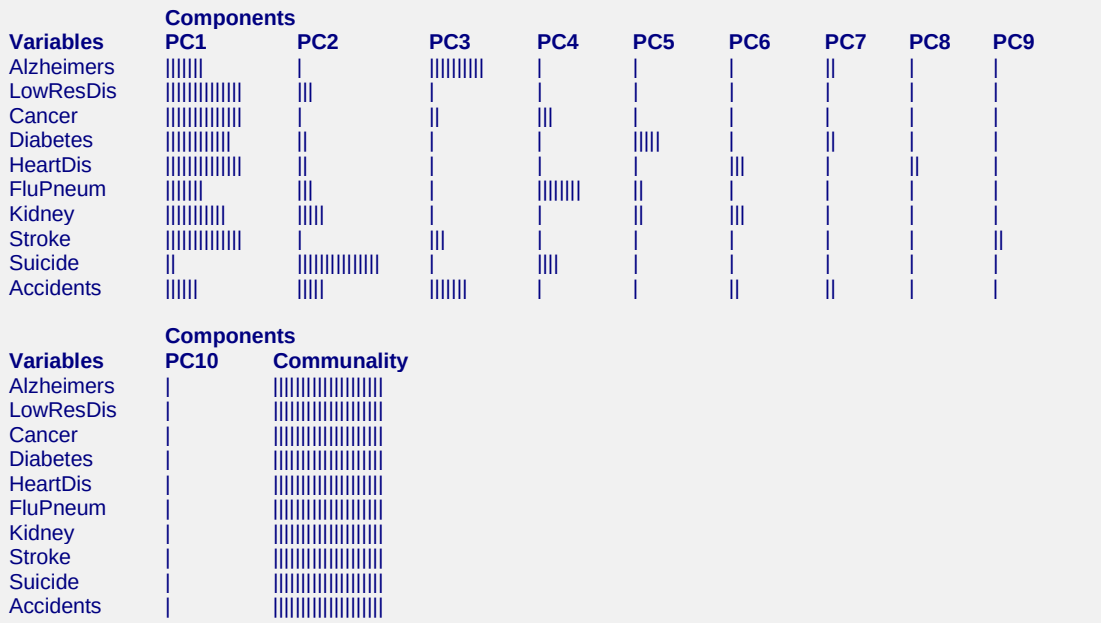
Communalities

Communalities

	Components					
Variables	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Alzheimers	0.334899	0.018770	0.484791	0.037686	0.002936	0.002825
LowResDis	0.695202	0.134483	0.000070	0.002580	0.015285	0.049912
Cancer	0.678897	0.013428	0.081430	0.123841	0.000155	0.024380
Diabetes	0.589368	0.064302	0.013064	0.004298	0.232259	0.001905
HeartDis	0.695570	0.057289	0.018332	0.000440	0.015822	0.101253
FluPneum	0.335283	0.141444	0.028313	0.381779	0.060110	0.000042
Kidney	0.527427	0.229194	0.013041	0.000298	0.050190	0.103346
Stroke	0.667844	0.042139	0.146189	0.001161	0.025330	0.026733
Suicide	0.061117	0.705521	0.002506	0.157103	0.047958	0.001514
Accidents	0.258173	0.224846	0.334933	0.041377	0.001848	0.071213

	Components				
Variables	**PC7**	**PC8**	**PC9**	**PC10**	**Communality**
Alzheimers	0.094510	0.002447	0.019900	0.001234	1.000000
LowResDis	0.021320	0.048548	0.005827	0.026773	1.000000
Cancer	0.001060	0.032502	0.016633	0.027672	1.000000
Diabetes	0.081701	0.012416	0.000092	0.000595	1.000000
HeartDis	0.007393	0.098415	0.005468	0.000019	1.000000
FluPneum	0.039888	0.012963	0.000155	0.000022	1.000000
Kidney	0.038443	0.000709	0.034724	0.002628	1.000000
Stroke	0.000627	0.004529	0.078930	0.006519	1.000000
Suicide	0.001088	0.002545	0.000528	0.020120	1.000000
Accidents	0.055867	0.003843	0.000003	0.007896	1.000000

Bar Chart of Communalities



Communality

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

Bar Chart of Communalities

This chart graphically displays the values of the communalities.

Interpretation

These results substantiate those found in looking at the *Loadings* reports and plots.

1. Heart Disease, Stroke, and Cancer are highly correlated. With Diabetes and LowResDis, they appear to be related to PC1.
2. Suicide is highly correlated to PC2.
3. PC3 and PC4 are hard to interpret since all of the values are less than 0.5.

Component Structure Summary

Component Structure Summary

	Components							
PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**
HeartDis	Suicide	Alzheimers	FluPneum	Diabetes				
LowResDis	Kidney	Accidents						
Cancer	Accidents							
Stroke								
Diabetes								
Kidney								
FluPneum								
Alzheimers								

	Components
PC1	**PC10**
HeartDis	
LowResDis	
Cancer	
Stroke	
Diabetes	
Kidney	
FluPneum	
Alzheimers	
Accidents	

Interpretation

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

Score Coefficients

Score Coefficients

	Components				
Variables	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
Alzheimers	-0.1194738	0.08397864	-0.6201912	-0.2586453	-0.1199137
LowResDis	-0.1721357	0.224786	-0.007427108	0.06767751	0.2735889
Cancer	-0.1701052	-0.07103108	0.2541803	-0.4688619	0.02754074
Diabetes	-0.1584926	0.1554346	-0.1018092	-0.08734928	-1.066473
HeartDis	-0.1721813	-0.1467131	0.120603	0.02794582	0.2783509
FluPneum	-0.1195423	-0.2305299	0.1498786	0.8232246	-0.5425473
Kidney	-0.1499327	-0.2934517	0.1017182	0.02299283	0.4957627
Stroke	-0.1687147	-0.1258281	-0.340569	0.04539716	0.3521945
Suicide	-0.05103818	0.514861	-0.04459185	0.5280852	0.4846113
Accidents	-0.104899	0.2906547	0.5154984	-0.2710156	-0.09513136

	Components				
Variables	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
Alzheimers	0.1387386	0.8991737	-0.2259844	-0.8693901	0.3757879
LowResDis	-0.5831264	-0.4270717	-1.006479	-0.4704384	-1.750407
Cancer	-0.4075499	0.09523229	-0.823524	0.7948371	1.779539
Diabetes	0.1139342	-0.8360211	0.5089967	-0.05901764	0.2608584
HeartDis	-0.8305485	0.2514835	1.433017	-0.4557119	-0.04607541
FluPneum	0.01688136	0.5841543	-0.5200886	0.07669481	0.05070405
Kidney	0.8390904	-0.5734735	-0.1215941	-1.148431	0.5484502
Stroke	0.4267608	-0.07322455	0.3073988	1.731447	-0.8637542
Suicide	0.1015615	-0.09648632	0.2304459	0.1416607	1.517407
Accidents	0.6965303	0.6913254	0.283171	0.01077313	-0.9506139

Score Coefficients

These are the coefficients that are used to form the component scores. The component scores are the values of the components 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 components to have the same variance.

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

Residuals

Residuals

State	T2	T2 Prob	Weight	Q0	Q1	Q2	Q9
AL	19.02	0.1507	0.60	22.18	6.07	5.23	0.31
AK	23.21	0.0706	0.51	11.36	11.16	4.99	0.39*
AZ	15.23	0.2865	0.72	11.90	8.46	3.12	0.38*
AR	8.83	0.6948	1.07	21.87	1.72	1.55	0.12
CA	10.23	0.5918	0.97	12.12	7.61	6.31	0.06
CO	14.78	0.3083	0.74	12.26	7.80	3.83	0.47*
CT	8.82	0.6951	1.06	15.41	3.50	2.54	0.06
DE	11.77	0.4838	0.88	5.88	5.88	4.57	0.04
DC	50.89*	0.0005	0.26	19.57	18.05*	13.64*	0.44*
FL	14.42	0.3263	0.76	6.46	2.71	2.52	0.64*
GA	7.09	0.8179	1.24	7.76	4.90	3.69	0.03
HI	42.78*	0.0019	0.30	40.80	29.00*	20.94*	0.03
ID	4.47	0.9536	1.63	4.39	4.09	1.64	0.02
IL	5.27	0.9212	1.48	5.17	4.96	0.67	0.08
IN	7.57	0.7850	1.20	7.22	2.10	2.10	0.03
IA	6.86	0.8324	1.27	3.41	2.70	2.32	0.03
KS	7.10	0.8172	1.23	2.92	2.91	2.91	0.00
KY	14.52	0.3214	0.75	31.32	4.79	4.46	0.00
LA	16.91	0.2170	0.66	17.80	5.05	3.70	0.27
ME	5.04	0.9316	1.52	3.74	3.60	2.27	0.00
MD	17.63	0.1918	0.64	11.44	8.90	3.08	0.00
MA	11.00	0.5361	0.91	15.41	6.32	4.21	0.11
MI	3.90	0.9709	1.75	2.85	1.67	1.08	0.00
MN	6.71	0.8420	1.27	12.09	3.17	2.85	0.00
MS	19.04	0.1500	0.60	54.99	6.07	3.27	0.04
MO	6.57	0.8508	1.29	6.89	2.90	2.88	0.00
MT	15.26	0.2851	0.72	13.07	12.77*	4.39	0.01
NE	15.17	0.2894	0.72	6.31	4.65	4.61	0.16
NV	25.81*	0.0436	0.47	13.59	13.23*	12.81*	0.09
NH	9.26	0.6626	1.02	9.03	7.01	5.16	0.11
NJ	9.44	0.6496	1.02	14.00	6.25	1.83	0.00
NM	25.21*	0.0487	0.49	14.52	14.51*	8.21*	0.38*
NY	22.83	0.0757	0.52	26.11	16.46*	9.15*	0.02
NC	6.65	0.8458	1.30	4.82	2.87	2.13	0.05
ND	10.60	0.5647	0.96	6.57	3.82	3.51	0.36*
OH	4.35	0.9576	1.66	6.84	1.82	1.82	0.04
OK	20.11	0.1240	0.58	25.63	8.71	4.46	0.00
OR	10.47	0.5740	0.94	7.38	6.12	3.81	0.13
PA	6.55	0.8521	1.31	4.18	4.17	3.94	0.04
RI	8.45	0.7224	1.09	10.07	3.85	3.80	0.06
SC	7.29	0.8045	1.22	7.80	3.85	3.72	0.02
SD	16.46	0.2343	0.68	6.67	6.67	5.00	0.52*
TN	13.78	0.3603	0.78	19.35	3.50	3.45	0.00
TX	7.09	0.8180	1.23	5.76	5.63	4.40	0.04
UT	31.05*	0.0164	0.40	16.25	15.93*	15.30*	0.00
VT	13.46	0.3788	0.78	13.09	9.69	4.67	0.06
VA	7.46	0.7925	1.19	3.64	3.03	1.71	0.22
WA	7.39	0.7976	1.20	8.99	5.47	4.97	0.04
WV	30.83*	0.0171	0.40	46.73	16.50*	11.91*	0.24
WI	1.76	0.9988	2.36	1.56	0.74	0.70	0.03
WY	18.13	0.1761	0.63	15.87	14.52*	9.10*	0.18
US	0.51	1.0000	3.04	0.55	0.40	0.11	0.01

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. These quantities are defined as follows.

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

$$T^2 = (x - \bar{x})' S^{-1} (x - \bar{x})$$

where x represents a p-variable observation, $\bar{x}$ 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 vector. The closer an observation is to the means, the smaller will be the value of T^2.

If the variables follow a multivariate normal distribution, then the probability distribution of T^2 is 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 $N(1 - \alpha)$ observations to be significant by chance variation.

In our current example, six rows are starred (which means they were significant at the 0.05 significance level). You would probably want to check for data entry or transcription errors.

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 components. 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 components, λ_i refers to the i^{th} eigenvalue, and pc_i is the score of the i^{th} component 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 components 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 components. 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: T2 and Q2. Notice that six rows are significantly large (shown by the asterisk) for both measurements. These are the DC, Hawaii, Nevada, New Mexico, Utah, and Wyoming. Apparently, mortality rates for these states have a different pattern from the rest of the country.

Component Scores

Component Scores

	Components						
State	PC1	PC2	PC3	PC4	PC5	PC6	PC7
AL	-1.8234	-0.7178	-0.9210	0.3459	1.9524	-0.7441	2.2176
AK	0.2051	1.9446	0.6597	1.0827	1.4264	1.3326	0.4941
AZ	0.8419	1.8104	-0.8022	-0.0520	-1.7757	-0.1482	0.1431
AR	-2.0396	-0.3200	-0.4629	0.5862	0.7701	-1.1722	-0.2109
CA	0.9647	-0.8950	-1.9810	0.5838	-1.8215	-0.3901	0.1463
CO	0.9589	1.5603	-1.0175	0.4762	1.8549	0.5951	0.6677
CT	1.5680	-0.7666	1.2596	-0.2191	0.3263	0.2559	0.5173
DE	-0.0386	-0.8955	0.4456	-1.3772	2.1691	0.5017	-0.3809
DC	0.5601	-1.6438	1.5748	-1.6823	-0.8170	-1.1306	1.1404
FL	0.8797	0.3457	0.4239	-0.6034	0.5071	0.6929	-0.8991
GA	-0.7679	-0.8613	-1.5902	-0.1568	0.8094	0.6069	0.5663
HI	1.5607	-2.2234	0.1586	4.6316	-1.9270	1.6671	2.1938

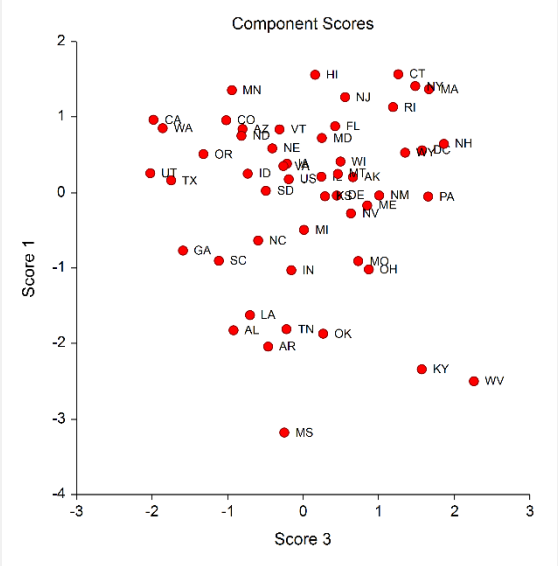
(report continues through all 52 rows)

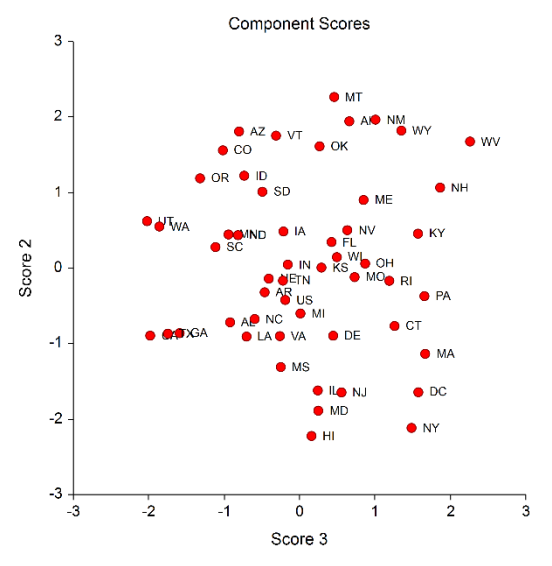
This report presents the individual component scores scaled so each column has a mean of zero and a standard deviation of one. These are the values that are plotted in the plots to follow.

These values are more easily interpreted by looking at them in scatter plots.

Component Scores Plots

Component Scores Plot(s)





This set of plots shows the data for each component plotted against every other component. The first *k* components (where *k* is the number of large eigenvalues) usually show the major structure that will be found in the data. The rest of the components show outliers and linear dependencies.

Interpretation of the Example

We start by considering the plot of Score 1 versus Score 2. Together, these represent 65% of the variation in the data. In this plot, Mississippi appears to be different from the rest of the states. Also, several of the states are close to other states that are their neighbors geographically. For example, New York and New Jersey are close to each other as are Idaho, Utah, Nevada, and Wyoming. These plots let you discover such clusters of states that behave similarly.

Example 2a – Storing the PC Scores

This example continues the analysis of the *Death Rates – States – 2016* dataset which was begun in Example 1. It will show how to store the PC scores on the dataset for further analysis.

This example will demonstrate a technique for directly and easily finding a reduced set of variables for further analysis. This reduced set of variables will contain most of the variation (information) that was in the full set.

It is not uncommon for a dataset to have 10, 20, or even 100 variables. One of the main reasons for using PCA is to reduce the number of variables down to a more manageable number. The technique follows the following steps.

1. Save the scores to the dataset. Using the Eigenvalue report, it was decided to keep the first four components. This set accounts for 83% of the variation in the data.
2. Use the *Subset Selection in Multivariate Y Regression* procedure to reduce the number of variables by finding the best subset of the original variables that predicts the group of component scores. Since the component scores represent the original variables, you are actually finding the best subset of the original variables.

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

Setup

To run this example, complete the following steps:

1 Open the Death Rates – States – 2016 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Death Rates – States – 2016** and click **OK**.

2 Specify the Principal Components Analysis procedure options

- Find and open the **Principal Components Analysis** procedure using the menus or the Procedure Navigator.
- Set **Variable Labels** to **Column Names** using the **Report Options** dropdown in the toolbar.
- The settings for this example are listed below and are stored in the **Example 2a** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Variables	**Alzheimers, LowResDis, Cancer, Diabetes, HeartDis, FluPneum, Kidney, Stroke, Suicide, Accidents**
Robust Covariance Matrix Estimation	**Checked**
Reports Tab	
Eigenvalue Summary	**Checked**
Score Report	**Checked**
Storage Tab	
Component Scores	**PC1, PC2, PC3, PC4** (We set these column names previously.)

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Eigenvalues

Eigenvalues

Robust Estimation: Iterations = 6 and Weight = 4
Missing-Value Estimation: Average

Number	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	4.843781	48.44	48.44	|||||||||
2	1.631417	16.31	64.75	||||
3	1.122669	11.23	75.98	|||
4	0.750564	7.51	83.48	||
5	0.451893	4.52	88.00	|
6	0.383123	3.83	91.83	|
7	0.341897	3.42	95.25	|
8	0.218917	2.19	97.44	|
9	0.162260	1.62	99.07	|
10	0.093478	0.93	100.00	|

This report shows again the percentage of variation accounted for by the first four components.

Component Scores

Component Scores

	Components						
State	PC1	PC2	PC3	PC4	PC5	PC6	PC7
AL	-1.8234	-0.7178	-0.9210	0.3459	1.9524	-0.7441	2.2176
AK	0.2051	1.9446	0.6597	1.0827	1.4264	1.3326	0.4941
AZ	0.8419	1.8104	-0.8022	-0.0520	-1.7757	-0.1482	0.1431
AR	-2.0396	-0.3200	-0.4629	0.5862	0.7701	-1.1722	-0.2109
CA	0.9647	-0.8950	-1.9810	0.5838	-1.8215	-0.3901	0.1463
CO	0.9589	1.5603	-1.0175	0.4762	1.8549	0.5951	0.6677
CT	1.5680	-0.7666	1.2596	-0.2191	0.3263	0.2559	0.5173
DE	-0.0386	-0.8955	0.4456	-1.3772	2.1691	0.5017	-0.3809
DC	0.5601	-1.6438	1.5748	-1.6823	-0.8170	-1.1306	1.1404
FL	0.8797	0.3457	0.4239	-0.6034	0.5071	0.6929	-0.8991
GA	-0.7679	-0.8613	-1.5902	-0.1568	0.8094	0.6069	0.5663
HI	1.5607	-2.2234	0.1586	4.6316	-1.9270	1.6671	2.1938

(report continues through all 52 rows)

This report presents the individual component scores. You can compare this report to the dataset to note that the first four columns have been stored to the dataset.

Example 2b – Variable Selection Analysis

This section presents an example of how to run a variable selection analysis of the data stored in Example 2a using the *Subset Selection in Multivariate Y Multiple Regression* procedure.

Setup

To run this example, complete the following steps:

1 Open the Death Rates – States – 2016 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Death Rates – States – 2016** and click **OK**.

2 Specify the Subset Selection in Multivariate Y Multiple Regression procedure options

- Find and open the **Subset Selection in Multivariate Y Multiple Regression** procedure using the menus or the Procedure Navigator.
- Set **Variable Labels** to **Column Names** using the **Report Options** dropdown in the toolbar.
- The settings for this example are listed below and are stored in the **Example 2b** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Y's Dependent Variables	**PC1, PC2, PC3, PC4**
X's Independent Variables	**Alzheimers, LowResDis, Cancer, Diabetes, HeartDis, FluPneum, Kidney, Stroke, Suicide, Accidents**
Maximum Variables	**7**
Report Options (*in the Toolbar*)	
Variable Labels	**Column Names**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Selection Results Section

Selection Results Section with Variable Names

Model Size	Wilks Lambda	Wilks Lambda Change	Variable Names
1	0.066454	-0.933546	Suicide
2	0.003471	-0.062983	FluPneum, Suicide
3	0.000181	-0.003289	Alzheimers, FluPneum, Suicide
4	**0.000012**	**-0.000169**	**Alzheimers, Cancer, FluPneum, Suicide**
5	0.000001	-0.000012	Alzheimers, Cancer, FluPneum, Suicide, Accidents
6	0.000000	-0.000001	Alzheimers, Cancer, Diabetes, FluPneum, Suicide, Accidents
7	0.000000	0.000000	LowResDis, Cancer, Diabetes, HeartDis, Kidney, Stroke, Accidents

This report presents the results of the search procedure. We see that the Suicide variable is the most highly correlated variable. Looking at the Wilks Lambda Change values, we select four variables to represent all eleven variables in the dataset. These variables are **Alzheimers, Cancer, FluPneum, and Suicide**.

Example 2c – Creating a 3D Scatter Plot of the Component Scores

This section presents an example of how to generate a 3D scatter plot. It assumes that you have run Example 2a and saved the scores (PC1, PC2, PC3, PC4) to the dataset.

Setup

To run this example, complete the following steps:

1 Open the Death Rates – States – 2016 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Death Rates – States – 2016** and click **OK**.

2 Specify the 3D Scatter Plot procedure options

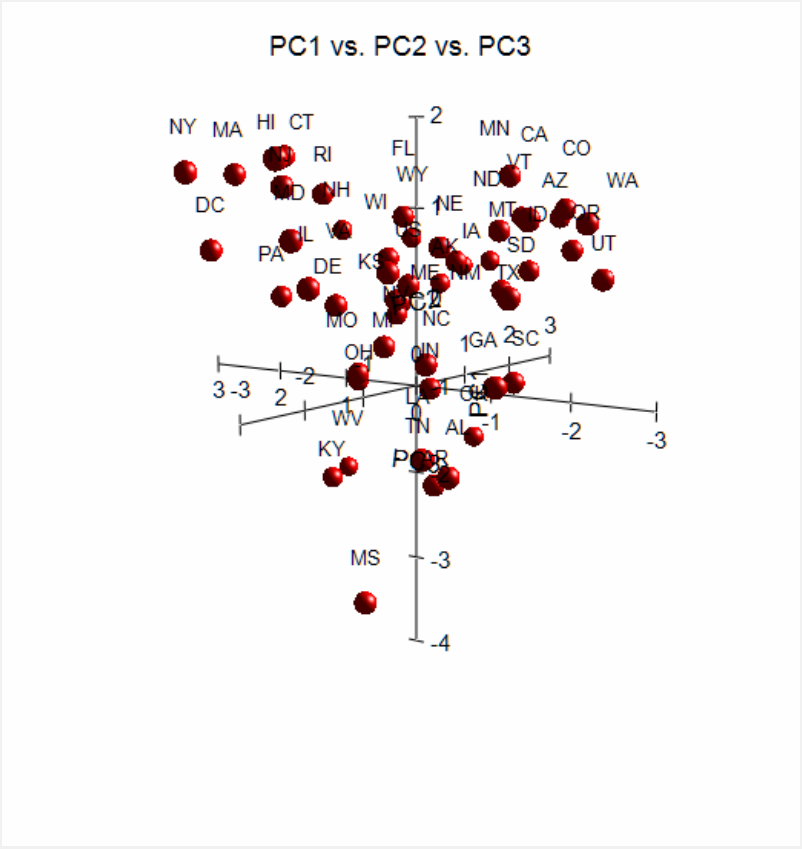
- Find and open the **3D Scatter Plot** procedure using the menus or the Procedure Navigator.
- Set **Variable Labels** to **Column Names** using the **Report Options** dropdown in the toolbar.
- The settings for this example are listed below and are stored in the **Example 2c** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
X (Horizontal) Variable	**PC2**
Y (Vertical) Variable	**PC1**
Z (Depth) Variable	**PC3**
Data Label Variable	**State**
Edit During Run (in Plot Format Button)	**Checked** (allows you to edit the 3D plot in real time)

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

3D Scatter Plot Output



The plot is a little crowded. Perhaps the default symbol size is too large. We suggest that you rotate the plot in real time so that you can obtain a better interpretation of the data. To do this, in the 3D Scatter Plot Format window,

1. Uncheck the *Actual Size* box.
2. Click the *Show in New Window* button.
3. Experiment with all the options to find the most informative view.

Example 3a – Storing the PC Loadings

This example continues the analysis of the *Death Rates – States – 2016* dataset which was begun in Example 1. It will show how to store the PC Loadings on the dataset for further analysis. It will then show how to create a three-dimensional plot of the loadings.

Setup

To run this example, complete the following steps:

1 Open the Death Rates – States – 2016 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Death Rates – States – 2016** and click **OK**.

2 Specify the Principal Components Analysis procedure options

- Find and open the **Principal Components Analysis** procedure using the menus or the Procedure Navigator.
- Set **Variable Labels** to **Column Names** using the **Report Options** dropdown in the toolbar.
- The settings for this example are listed below and are stored in the **Example 3a** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Variables	**Alzheimers, LowResDis, Cancer, Diabetes, HeartDis, FluPneum, Kidney, Stroke, Suicide, Accidents**
Robust Covariance Matrix Estimation	**Checked**
Reports Tab	
Eigenvalue Summary	**Checked**
Loadings Report	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Eigenvalues

Eigenvalues

Robust Estimation: Iterations = 6 and Weight = 4
Missing-Value Estimation: Average

Number	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	4.843781	48.44	48.44	|||||||||
2	1.631417	16.31	64.75	||||
3	1.122669	11.23	75.98	|||
4	0.750564	7.51	83.48	||
5	0.451893	4.52	88.00	|
6	0.383123	3.83	91.83	|
7	0.341897	3.42	95.25	|
8	0.218917	2.19	97.44	|
9	0.162260	1.62	99.07	|
10	0.093478	0.93	100.00	|

This report shows again the percentage of variation accounted for by the components.

Component Loadings

Component Loadings

	Components					
Variables	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Alzheimers	-0.578705	0.137004	-0.696269	-0.194130	-0.054188	0.053154
LowResDis	-0.833788	0.366720	-0.008338	0.050796	0.123633	-0.223409
Cancer	-0.823952	-0.115881	0.285360	-0.351911	0.012445	-0.156142
Diabetes	-0.767703	0.253579	-0.114298	-0.065561	-0.481932	0.043651
HeartDis	-0.834008	-0.239350	0.135397	0.020975	0.125785	-0.318203
FluPneum	-0.579037	-0.376091	0.168264	0.617883	-0.245174	0.006468
Kidney	-0.726241	-0.478742	0.114196	0.017258	0.224032	0.321475
Stroke	-0.817217	-0.205278	-0.382346	0.034073	0.159154	0.163502
Suicide	-0.247218	0.839953	-0.050062	0.396362	0.218993	0.038911
Accidents	-0.508108	0.474179	0.578734	-0.203415	-0.042989	0.266857

	Components			
Variables	**PC7**	**PC8**	**PC9**	**PC10**
Alzheimers	0.307425	-0.049472	-0.141067	0.035128
LowResDis	-0.146014	-0.220336	-0.076333	-0.163625
Cancer	0.032560	-0.180283	0.128970	0.166348
Diabetes	-0.285833	0.111428	-0.009576	0.024385
HeartDis	0.085981	0.313712	-0.073944	-0.004307
FluPneum	0.199721	-0.113856	0.012445	0.004740
Kidney	-0.196069	-0.026619	-0.186345	0.051268
Stroke	-0.025035	0.067295	0.280945	-0.080742
Suicide	-0.032988	0.050449	0.022986	0.141844
Accidents	0.236362	0.061991	0.001748	-0.088862

This report shows again the component loadings which are the correlations between each component (PC1, PC2, etc.) and each variable (Alzheimers, LowResDis, etc.).

Example 3b – Creating a 3D Scatter Plot of the Component Loadings

This section presents an example of how to generate a 3D scatter plot of the loadings. It assumes that you have just run Example 3a and you are looking at the Components Loadings report.

Setup

To run this example, complete the following steps:

1 Open the Death Rates – States – 2016 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Death Rates – States – 2016** and click **OK**.

2 Specify the 3D Scatter Plot procedure options

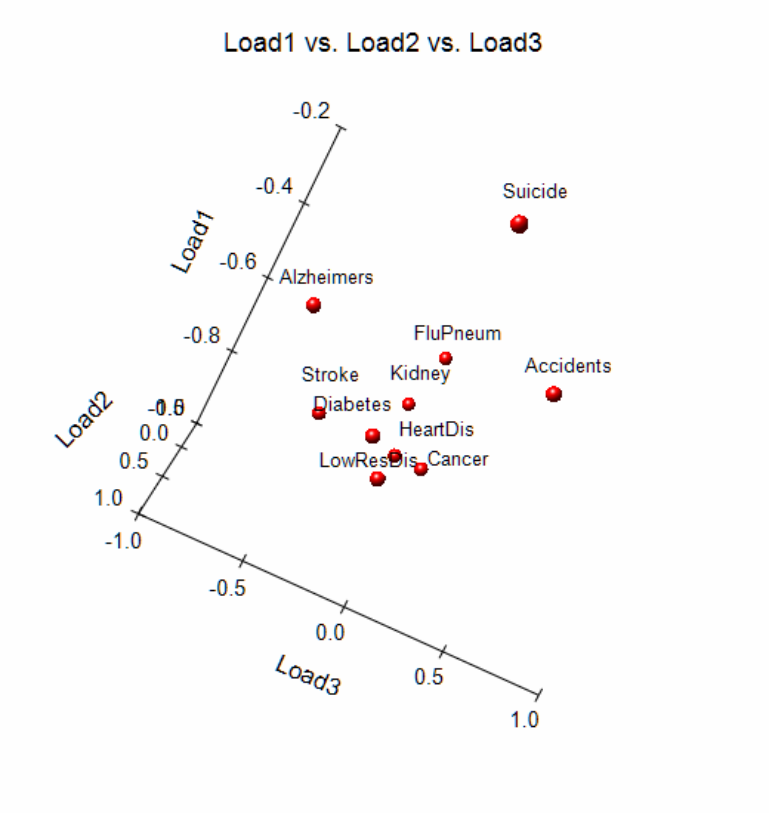
- Find and open the **3D Scatter Plot** procedure using the menus or the Procedure Navigator.
- Set **Variable Labels** to **Column Names** using the **Report Options** dropdown in the toolbar.
- The settings for this example are listed below and are stored in the **Example 3b** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
X (Horizontal) Variable	**Load2**
Y (Vertical) Variable	**Load1**
Z (Depth) Variable	**Load3**
Data Label Variable	**Variables**
Edit During Run (in Plot Format Button)	**Checked** (allows you to edit the 3D plot in real time)

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

3D Scatter Plot Output



Your plot may not look like this at first. While looking at the plot,

1. Check the *Labels* under *Data Point Labels*.
2. Click the *Walls* tab. Uncheck all of the walls.
3. We have found that motion is needed to allow the eye to see the spatial relationship among the points. To accomplish this, turn off (uncheck) all lines, tick marks, and labels.
4. Click the *Show in New Window* button at the bottom right of the plot.
5. Click the *3D Orientation* tab at the bottom left of the plot.
6. Check the three *Auto Spin* boxes to cause the plot to rotate. This will allow you to see which points a near each other and which are far from each other. This will allow you to determine which mortality patterns occur in the same way across all states.
7. You can experiment with all the various settings. Once you find a plot you like, click the *Close* button and the current plot will be displayed in the report.

From our brief inspection of the 3D plot, we concluded the following.

1. Suicide and Accidents loadings are each different from other causes of death.
2. Alzheimers loadings are different from other causes of death.
3. Diabetes and lower respiratory loadings across the states appear to be similar.
4. Cancer, Heart Disease, Stroke, Kidney Disease, and Flu/Pneumonia have similar loadings.

More plots will be needed to fully understand what these data can show us.

Example 4a – Using Robust Estimation to Find and Remove Outliers

This section presents an example of how to find and remove outliers from a principal components analysis. The data used are found in the PCA2 dataset. This example will look for outliers in variables X1-X6.

Setup

To run this example, complete the following steps:

1 Open the PCA2 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **PCA2** and click **OK**.

2 Specify the Principal Components Analysis procedure options

- Find and open the **Principal Components Analysis** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 4a** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Variables	**X1, X2, X3, X4, X5, X6**
Robust Covariance Matrix Estimation	**Checked**
Reports Tab	
Eigenvalue Summary	**Checked**
Iteration Report	**Checked**
Residuals (Q and T2)	**Checked**
Plots Tab	
Scores Plot	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Robust and Missing-Value Estimation Iteration

Robust and Missing-Value Estimation Iteration

Robust Estimation: Iterations = 6 and Weight = 4
Missing-Value Estimation: None

Iteration Number	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

In this particular example, we see very little change between iterations five and six. We would feel comfortable in stopping at this point.

Eigenvalues

Eigenvalues

Robust Estimation: Iterations = 6 and Weight = 4
Missing-Value Estimation: None

Number	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	4.702330	78.37	78.37	|||||||||||||||
2	1.290318	21.51	99.88	|||||
3	0.006414	0.11	99.98	|
4	0.000780	0.01	100.00	|
5	0.000115	0.00	100.00	|
6	0.000042	0.00	100.00	|

The first two components are all that are needed.

Residuals

Residuals

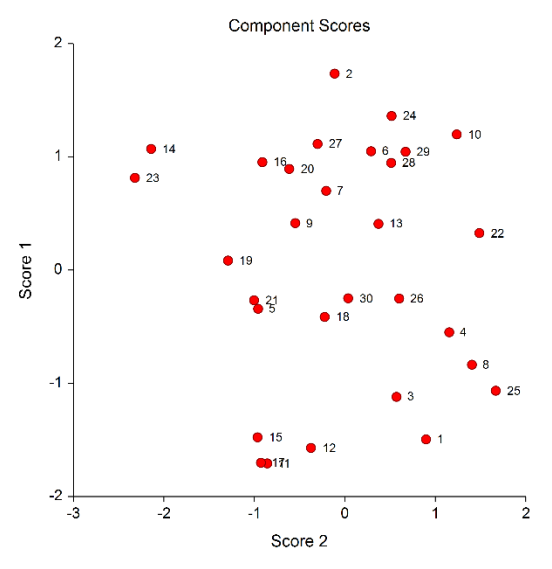
Row	T2	T2 Prob	Weight	Q0	Q1	Q2	Q5
1	4.57	0.7047	1.01	11.55	1.04	0.00	0.00
2	1104.76*	0.0000	0.02	15.20	1.04	1.03*	0.00*
3	1312.74*	0.0000	0.02	14.55	8.65*	8.23*	0.00
4	1.95	0.9458	1.42	3.14	1.72	0.00	0.00
5	9.27	0.3040	0.64	1.73	1.19	0.00	0.00*
6	4.36	0.7261	1.04	5.29	0.11	0.00	0.00
7	1.35	0.9778	1.60	2.36	0.06	0.00	0.00
8	3.25	0.8394	1.20	5.83	2.55	0.00	0.00
9	3.22	0.8417	1.20	1.20	0.39	0.00	0.00
10	3.76	0.7884	1.09	8.73	1.98	0.00	0.00
11	4.02	0.7614	1.06	14.65	0.95	0.00	0.00
12	4.72	0.6887	0.98	11.78	0.18	0.00	0.00
13	6.48	0.5152	0.81	0.96	0.18	0.00	0.00
14	9.21	0.3079	0.65	11.30	5.91*	0.00	0.00
15	3.64	0.8001	1.11	11.45	1.20	0.00	0.00
16	4.67	0.6937	0.98	5.34	1.07	0.00	0.00
17	4.36	0.7262	1.06	14.72	1.11	0.00	0.00
18	1.98	0.9438	1.42	0.87	0.06	0.00	0.00
19	5.15	0.6448	0.93	2.18	2.15	0.00	0.00
20	6.61	0.5037	0.81	4.24	0.49	0.00	0.00
21	4.11	0.7520	1.05	1.63	1.30	0.00	0.00
22	4.04	0.7592	1.09	3.36	2.85	0.00	0.00
23	6.69	0.4962	0.81	10.07	6.95*	0.00	0.00

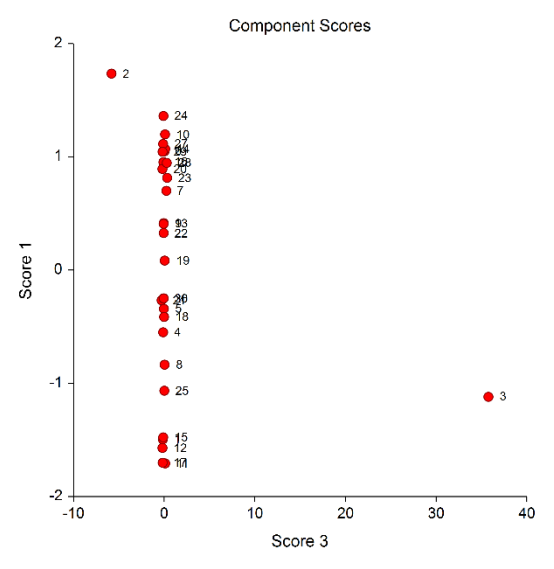
24	2.45	0.9103	1.32	9.05	0.34	0.00	0.00
25	4.23	0.7403	1.03	8.93	3.59	0.00	0.00
26	3.03	0.8597	1.24	0.76	0.47	0.00	0.00
27	5.55	0.6046	0.89	5.95	0.12	0.00	0.00
28	6.72	0.4936	0.81	4.56	0.34	0.00	0.00
29	1.81	0.9544	1.46	5.72	0.58	0.00	0.00
30	3.03	0.8600	1.21	0.29	0.00	0.00	0.00

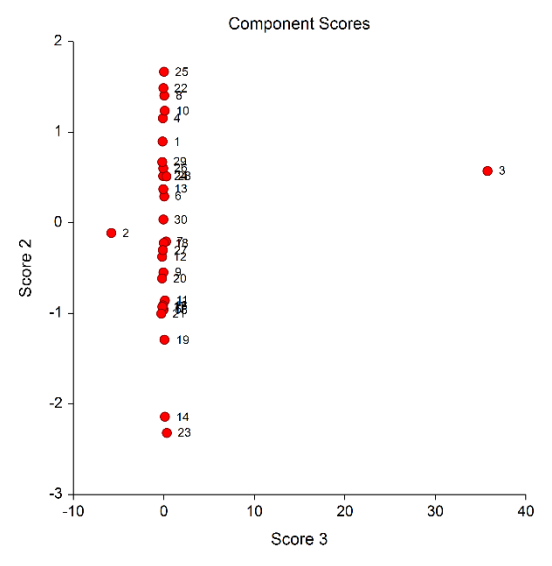
It is obvious that there are two outliers: rows 2 and 3.

Component Scores Plots

Component Scores Plot(s)







These reports also lead to the conclusion that rows 2 and 3 are outliers.

Example 4b – Removing Outliers

This section presents an example of how to remove outliers from a principal components analysis. The data used are found in the PCA2 dataset. This example will look for outliers in variables X1-X6. In Example 2a, we found that rows 2 and 3 are outliers, so we want to remove them from the analysis. We could simply select and delete those two rows, but that does not leave a good trail as to what was done.

So, in this example, we will remove those two rows using a filter.

Setup

To run this example, complete the following steps:

1 Open the PCA2 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **PCA2** and click **OK**.

2 Create a data filter

- Enter **1** in the first 30 rows of column **C18**.
- Enter **0** in rows 2 and 3 of column **C18**.
- Press the blue **Column Info** box if the light-yellow variable info spreadsheet is not showing.
- Enter **1** as the Filter in column **C18**.

3 Specify the Principal Components Analysis procedure options

- Find and open the **Principal Components Analysis** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 4b** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Variables	**X1, X2, X3, X4, X5, X6**
Robust Covariance Matrix Estimation	**Checked**
Reports Tab	
Eigenvalue Summary	**Checked**
Iteration Report	**Checked**
Residuals (Q and T2)	**Checked**
Plots Tab	
Scores Plot	**Checked**

4 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Robust and Missing-Value Estimation Iteration

Robust and Missing-Value Estimation Iteration

Robust Estimation: Iterations = 6 and Weight = 4
Missing-Value Estimation: None

Iteration Number	Count	Trace of Covar Matrix	Percent Change
0	28	4490.09	0.00
1	28	4490.09	0.00
2	28	4316.382	-3.87
3	28	4316.382	0.00
4	28	4280	-0.84
5	28	4280	0.00
6	28	4270.101	-0.23

Again, we see very little change between iterations five and six. We would feel comfortable in stopping at this point.

Eigenvalues

Eigenvalues

Robust Estimation: Iterations = 6 and Weight = 4
Missing-Value Estimation: None

Number	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	4.693290	78.22	78.22	|||||||||||||||
2	1.306325	21.77	99.99	|||||
3	0.000149	0.00	100.00	|
4	0.000108	0.00	100.00	|
5	0.000091	0.00	100.00	|
6	0.000037	0.00	100.00	|

The first two components are all that are needed.

Residuals

Residuals

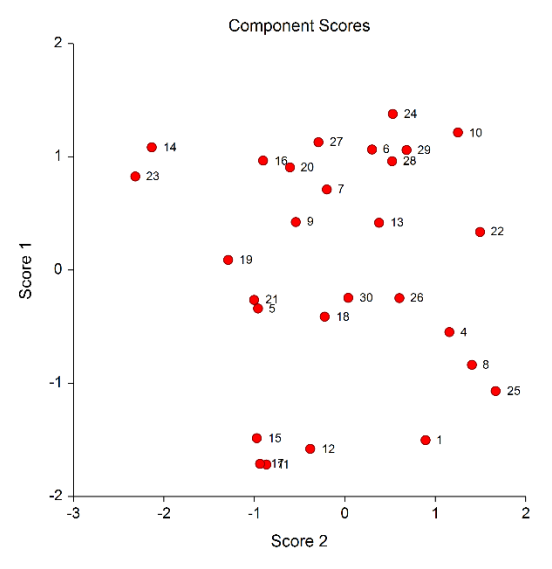
Row	T2	T2 Prob	Weight	Q0	Q1	Q2	Q5
1	7.59	0.4320	0.85	11.63	1.04	0.00	0.00
4	2.15	0.9346	1.58	3.15	1.74	0.00	0.00
5	9.92	0.2788	0.71	1.74	1.20	0.00*	0.00
6	6.70	0.5058	0.91	5.44	0.12	0.00	0.00
7	5.45	0.6228	1.04	2.43	0.05	0.00	0.00
8	6.03	0.5675	0.97	5.86	2.58	0.00	0.00
9	4.82	0.6865	1.11	1.23	0.38	0.00	0.00*
10	4.79	0.6892	1.11	8.98	2.04	0.00	0.00
11	6.30	0.5422	0.95	14.82	0.98	0.00	0.00
12	6.52	0.5220	0.93	11.89	0.19	0.00	0.00
13	6.07	0.5635	0.97	1.01	0.19	0.00	0.00
14	9.15	0.3230	0.75	11.47	5.95*	0.00	0.00
15	4.06	0.7639	1.21	11.58	1.24	0.00	0.00
16	5.24	0.6441	1.06	5.46	1.07	0.00	0.00
17	8.90	0.3391	0.76	14.89	1.15	0.00	0.00
18	2.40	0.9163	1.53	0.86	0.06	0.00	0.00
19	5.36	0.6319	1.05	2.21	2.17	0.00	0.00
20	10.55	0.2468	0.68	4.34	0.48	0.00*	0.00
21	7.39	0.4481	0.86	1.64	1.31	0.00	0.00
22	7.28	0.4566	0.87	3.45	2.92	0.00	0.00
23	14.14	0.1224	0.55	10.21	7.00*	0.00	0.00
24	3.24	0.8438	1.34	9.29	0.37	0.00	0.00
25	5.00	0.6679	1.08	8.98	3.63	0.00	0.00

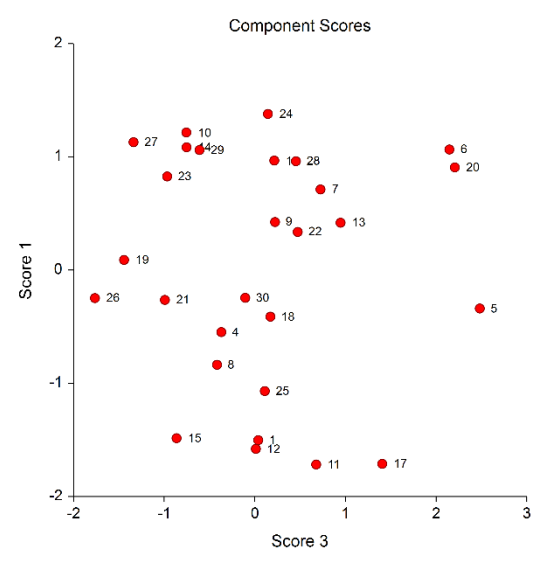
26	5.23	0.6446	1.05	0.76	0.48	0.00	0.00
27	6.49	0.5250	0.93	6.10	0.11	0.00	0.00
28	14.23	0.1203	0.54	4.70	0.36	0.00*	0.00
29	3.72	0.7973	1.26	5.89	0.61	0.00	0.00
30	3.17	0.8507	1.36	0.28	0.00	0.00	0.00

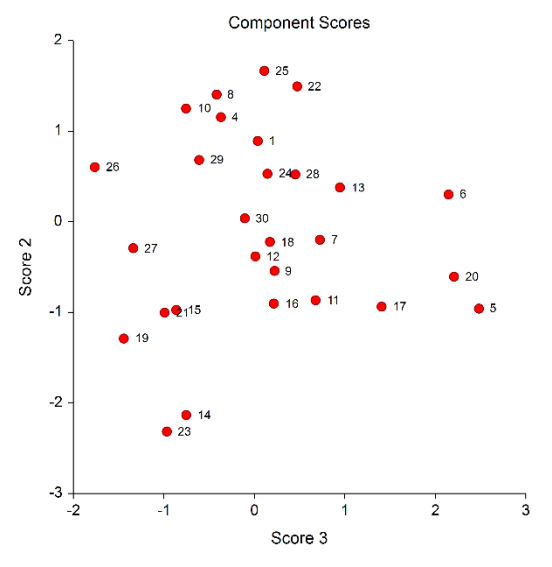
Note that rows 2 and 3 have been omitted. There does not appear to be any further outliers.

Component Scores Plots

Component Scores Plot(s)







These reports also lead to the conclusion all outliers have been found and removed. You can now rerun the analysis with more reports so that you can carry-out the PCA.