Chapter 420

Factor Analysis

Introduction

Factor Analysis (FA) is an exploratory technique applied to a set of observed variables that seeks to find underlying factors (subsets of variables) from which the observed variables were generated. For example, an individual’s response to the questions on a college entrance test is influenced by underlying variables such as intelligence, years in school, age, emotional state on the day of the test, amount of practice taking tests, and so on. The answers to the questions are the observed variables. The underlying, influential variables are the factors.

Factor analysis is carried out on the correlation matrix of the observed variables. A factor is a weighted average of the original variables. The factor analyst hopes to find a few factors from which the original correlation matrix may be generated.

Usually, the goal of factor analysis is to aid data interpretation. The factor analyst hopes to identify each factor as representing a specific theoretical factor. Therefore, many of the reports from factor analysis are designed to aid in the interpretation of the factors.

Another goal of factor analysis is to reduce the number of variables. The analyst hopes to reduce the interpretation of a 200-question test to the study of 4 or 5 factors. One of the most subtle tasks in factor analysis is determining the appropriate number of factors.

Factor analysis has an infinite number of solutions. If a solution contains two factors, these may be rotated to form a new solution that does just as good a job at reproducing the correlation matrix. Hence, one of the biggest complaints of factor analysis is that the solution is not unique. Two researchers can find two different sets of factors that are interpreted quite differently yet fit the original data equally well.

NCSS provides the *principal axis method* of factor analysis. The results may be rotated using varimax or quartimax rotation. The factor scores may be stored for further analysis.

Many books are devoted to factor analysis. We suggest you obtain a book on the subject from an author in your own field. An excellent introduction to the subject is provided by Tabachnick (1989).

Technical Details

Mathematical Development

This section will document the basic formulas used by **NCSS** in performing a factor analysis. The following table lists many of the matrices that are used in the discussion to follow.

Label	Matrix Name	Size	Description
R	Correlation	pxp	Matrix of correlations between each pair of variables.
X	Data	Nxp	Observed data matrix with N rows (observations) and p columns (variables).
Z	Standardized data	Nxp	Matrix of standardized data. The standardization of each variable is made by subtracting its mean and dividing by its standard deviation.
A	Factor loading	pxm	Matrix of correlations between the original variables and the factors. Also represents the contribution of each factor in estimating the original variables.
L	Eigenvalue	mxm	Diagonal matrix of eigenvalues. Only the first m eigenvalues are considered.
V	Eigenvector	pxm	Matrix of eigenvectors. Only the first m columns of this matrix are used.
B	Factor-score coefficients	pxm	Matrix of regression weights used to construct the factor scores from the original variables.
U	Uniqueness	pxp	Matrix of uniqueness values.
F	Factor score	Nxm	Matrix of factor scores. For each observation in the original data, the values of each of the retained factors are estimated. These are the factor scores.

The principal-axis method is used by **NCSS** to solve the factor analysis problem. Factor analysis assumes the following partition of the correlation matrix, *R*:

$$R = AA' + U$$

The principal-axis method proceeds according to the following steps:

1. Estimate *U* from the communalities as discussed below.
2. Find *L* and *V*, the eigenvalues and eigenvectors of *R*-*U* using standard eigenvalue analysis.
3. Calculate the loading matrix as follows:

$$A = VL^{\frac{1}{2}}$$

4. Calculate the score matrix as follows:

$$B = VL^{-\frac{1}{2}}$$

5. Calculate the factor scores as follows:

$$F = ZB$$

Steps 1-3 may be iterated since a new *U* matrix may be estimated from the current loading matrix.

Initial Communality Estimation

We close this section with a discussion of obtaining an initial value of U. **NCSS** uses the initial estimation of Cureton (1983), which will be outlined here. The initial communality estimates, c_{ii}, are calculated from the correlation and inverse correlation matrices as follows:

$$c_{ii} = \left(1 - \frac{1}{R^{ii}}\right) \frac{\sum_{k=1}^{p} \max_{over\ j \neq k} \left(\left|r_{jk}\right|\right)}{\sum_{k=1}^{p} \left(1 - \frac{1}{R^{kk}}\right)}$$

where R_{ii} is the i^{th} diagonal element of R-1 and r_{jk} is an element of R. The value of U is then estimated by 1-c_{ii}.

Missing Values and Robust Estimation

Missing values and robust estimation are done the same way as in principal components analysis. Refer to that chapter for details.

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.

Varimax and Quartimax Rotation

Factor analysis 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 factor analysis 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 factor.

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; 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 $G = \{g_{ij}\}$ represent the rotated factors. The goal of varimax rotation is to maximize the quantity:

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

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:

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

where c_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 G are maximized rather than the columns of G. 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:

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

Miscellaneous Topics

Using Correlation Matrices Directly

Occasionally, you will be provided with only the correlation 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 factor analysis on a correlation matrix using **NCSS**.

NCSS can store the correlation 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.

Principal Component Analysis versus Factor Analysis

Both principal component analysis (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.

Data Structure

The data for a factor analysis consists of two or more columns. 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 = .33 V1 + .65 V2 + error. Each variable had a different set of weights (.33 and .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 is contained in the dataset PCA2. The data given below are the first few rows of this dataset.

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 – Factor Analysis

Even though we here go directly into running factor analysis here, it is important to realize that the first step in any real factor analysis 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 section presents an example of how to run a factor analysis. The data used are shown in the table above and stored in the PCA2 dataset.

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 Factor Analysis procedure options

- Find and open the **Factor Analysis** procedure using the menus or the Procedure Navigator.
- 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	**X1-X6**
Robust Covariance Matrix Estimation	**Checked**
Maximum Iterations	**6**
Factor Rotation	**Varimax**
Number of Factors	**2**
Communality Iterations	**6**
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.)

3 Run the procedure

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

Robust and Missing-Value Iteration Section

This report is only produced when robust or missing value estimation is used.

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

Iteration                   Trace of    Percent
Number        Count      Covar Matrix    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 robust 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 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 have been removed by the robust procedure.

Descriptive Statistics Section

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

                                        Standard
Variables     Count        Mean        Deviation     Communality
X1               30    42.83667         23.18579        0.997983
X2               30    53.25062         27.93123        0.999791
X3               30    57.13034          26.3737        0.999585
X4               30     43.5617         24.56474        0.992023
X5               30    43.20835         25.75021        1.000072
X6               30    48.61827         32.49559        0.999944
```

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 similar to the R-Squared that would be obtained if this variable were regressed on the factors that were kept. However, remember that this is not based directly on the correlation matrix. Instead, calculations are based on an adjusted correlation matrix.

Correlation Section

```
Correlations ──────────────────────────────────────────────────────────────────

            Variables
Variables          X1          X2          X3          X4          X5          X6
X1           1.000000    0.271780    0.127016    0.881604    0.814686    0.484907
X2           0.271780    1.000000    0.988909    0.683206    0.778649    0.973093
X3           0.127016    0.988909    1.000000    0.568933    0.677480    0.928454
X4           0.881604    0.683206    0.568933    1.000000    0.986945    0.831949
X5           0.814686    0.778649    0.677480    0.986945    1.000000    0.901975
X6           0.484907    0.973093    0.928454    0.831949    0.901975    1.000000

Phi=0.769781  Log(Det|R|)=-29.547320  Bartlett Test=773.15  DF=15  Prob=0.000000


Bar Chart of Absolute Correlations ─────────────────────────────────────────────

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

Phi=0.769781  Log(Det|R|)=-29.547320  Bartlett Test=773.15  DF=15  Prob=0.000000
```

This report gives the correlations alone 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 are small (say less then .3), there would be no need for a factor analysis.

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 φ1 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 φ2 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 factor analysis 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:

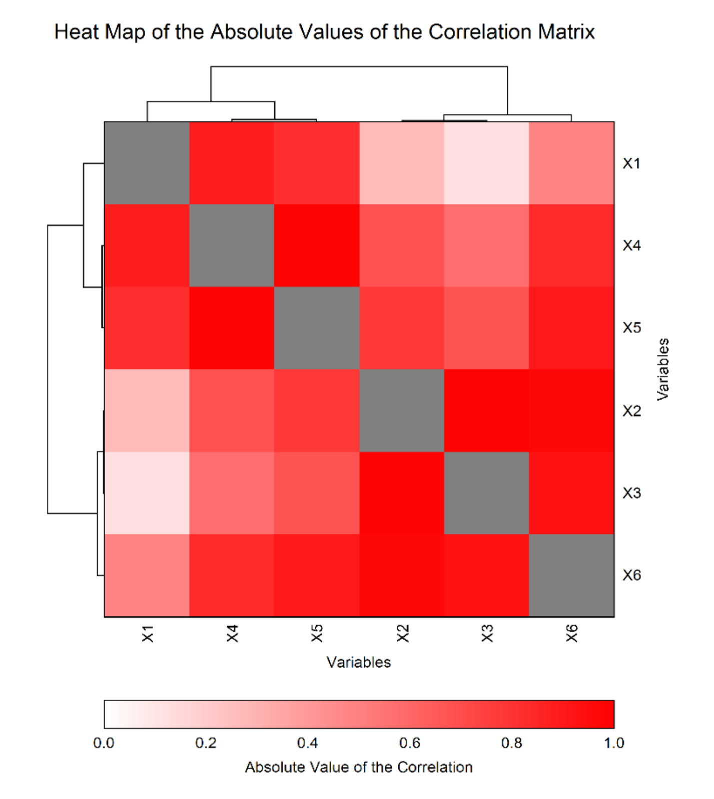
$$\chi^2 = \frac{(11 + 2p - 6N)}{6}\,\mathrm{Log}_e|R|$$

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.

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.

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            X4, X5
2            X2, X3, X6
None         X1
```

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

        Number     Distance      Distance
Link    Clusters      Value      Bar
5       1          0.408234      |IIIIIIIIIIIIIIIIIIIIIIIIIIIIII
4       2          0.151855      |IIIIIIIIIII
3       3          0.049227      |IIII
2       4          0.013055      |I
1       5          0.011091      |I

Cophenetic Correlation       0.666568
Delta(0.5)                   0.473224
Delta(1.0)                   0.588991
```

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 after Varimax Rotation ──────────────────────────────────────────────
Robust Estimation:          Iterations = 6 and Weight = 4
Missing-Value Estimation:   Average

                                 Individual    Cumulative
Number        Eigenvalue            Percent       Percent     Scree Plot
1               3.288191              54.89         54.89     |||||||||||
2               2.701207              45.09         99.99     ||||||||||
3               0.001207               0.02        100.01     |
4              -0.000099               0.00        100.01     |
5              -0.000121               0.00        100.00     |
6              -0.000295               0.00        100.00     |
```

Eigenvalues

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

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 3.3 of the original variables.

Note that, unlike in PCA where all eigenvalues are positive, the eigenvalues may be negative in factor analysis. Usually, these factors would be discarded and the analysis would be re-run.

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

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 99.99% of the variation.

Eigenvectors

```
Eigenvectors after Varimax Rotation ─────────────────────────────────────────

                   Factors
Variables          Factor1       Factor2
X1                -0.303444     -0.662220
X2                -0.416551      0.378018
X3                -0.382768      0.491154
X4                -0.428167     -0.317929
X5                -0.448606     -0.204840
X6                -0.450912      0.185189

Bar Chart of Absolute Eigenvectors after Varimax Rotation ───────────────────

                 Factors
Variables        Factor1         Factor2
X1               |||||||         ||||||||||||||
X2               |||||||||       ||||||||
X3               ||||||||        ||||||||||
X4               |||||||||       |||||||
X5               |||||||||       |||||
X6               ||||||||||      ||||
```

Eigenvector

The eigenvectors of the R-U matrix.

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

```
Factor Loadings after Varimax Rotation ──────────────────────────────────────

                     Factors
Variables            Factor1        Factor2
X1                  -0.019936      -0.998792
X2                  -0.967470      -0.252572
X3                  -0.994037      -0.107126
X4                  -0.478418      -0.873578
X5                  -0.594943      -0.803812
X6                  -0.883654      -0.468080

Bar Chart of Absolute Factor Loadings after Varimax Rotation ────────────────

                 Factors
Variables        Factor1                 Factor2
X1               |                       ||||||||||||||||||||
X2               ||||||||||||||||||||    ||||||
X3               ||||||||||||||||||||    |||
X4               ||||||||||              ||||||||||||||||||
X5               ||||||||||||            |||||||||||||||||
X6               ||||||||||||||||||      ||||||||||
```

Factor Loadings

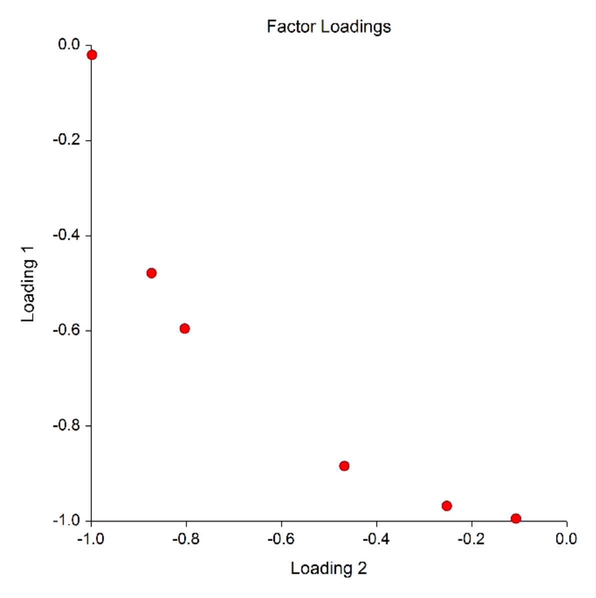
These are the correlations between the variables and factors.

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.

Factor Loading Plots

Factor Loadings Plot(s)



This set of plots shows each of the factor loading columns plotted against each other.

Communality

Communalities after Varimax Rotation

	Factors		
Variables	Factor1	Factor2	Communality
X1	0.000397	0.997586	0.997983
X2	0.935999	0.063793	0.999791
X3	0.988109	0.011476	0.999585
X4	0.228883	0.763139	0.992023
X5	0.353958	0.646114	1.000072
X6	0.780845	0.219099	0.999944

Bar Chart of Communalities after Varimax Rotation

```
            Factors
Variables   Factor1               Factor2               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 similar to the R-Squared 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

```
Factor Structure Summary after Varimax Rotation ─────────────────────────────────

          Factors
Factor1        Factor2
X3             X1
X2             X4
X6             X5
X5             X6
X4
```

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. Note that it is possible for a variable to have high loadings on several factors, although varimax rotation makes this very unlikely.

Score Coefficients

```
Score Coefficients after Varimax Rotation ─────────────────────────────────────

                Factors
Variables       Factor1         Factor2
X1              0.268188        -0.5366089
X2              -0.3613275      0.1312937
X3              -0.4135219      0.2176106
X4              0.02901571      -0.3414549
X5              -0.0422971      -0.2712604
X6              -0.2641693      -0.008934696
```

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.

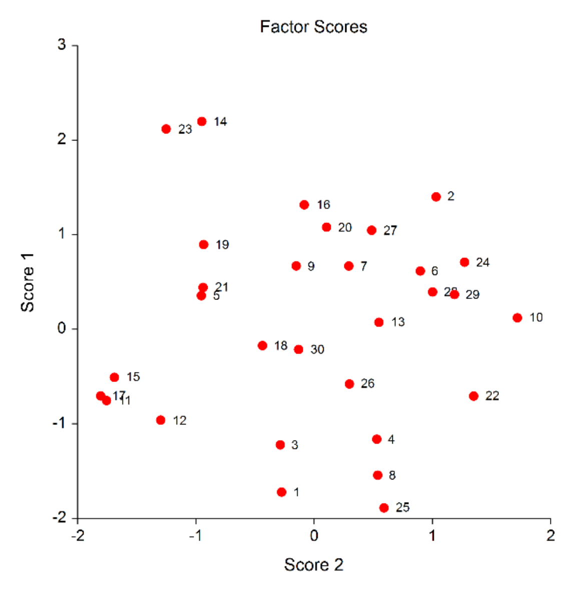
Factor Scores

```
Factor Scores after Varimax Rotation ─────────────────────────────

         Factors
Row      Factor1      Factor2
1        -1.7219      -0.2752
2         1.4002       1.0317
3        -1.2231      -0.2862
4        -1.1632       0.5302

(report continues through all thirty rows)
```

The factor scores are the values of the factors for a particular row of data. They have been scaled so they have a variance of one.

Factor Score Plots

Factor Scores Plot(s)



This set of plots shows each factor plotted against every other factor.

Example 2a – Storing the Factor Scores

This example analyzes the data 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. It will show how to store the factor scores on the dataset for further analysis.

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 Factor Analysis procedure options**

- Find and open the **Factor 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**
Data Label Variable	**State**
Factor Rotation	**Varimax**
Number of Factors	**3**
Reports Tab	
Eigenvalue Summary	**Checked**
Score Report	**Checked**
Storage Tab	
Factor Scores	**F1, F2, F3** (We set these column names previously.)

3 **Run the procedure**

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

Eigenvalues

```
Eigenvalues after Varimax Rotation ─────────────────────────────────────────
Robust Estimation:           None
Missing-Value Estimation:    None

                            Individual   Cumulative
Number       Eigenvalue        Percent      Percent     Scree Plot
1              2.803478          43.28        43.28     |||||||||
2              2.102399          32.46        75.74     |||||||
3              1.549342          23.92        99.66     |||||
4              0.308242           4.76       104.42     |
5              0.137108           2.12       106.54     |
6              0.039781           0.61       107.15     |
7              0.006154           0.10       107.24     |
8             -0.113017          -1.74       105.50     |
9             -0.140473          -2.17       103.33     |
10            -0.215772          -3.33       100.00     |
```

This report shows the percentage of variation accounted for by the eigenvalues. Notice that only the first three are greater than 1.0.

Factor Scores

```
Factor Scores after Varimax Rotation ─────────────────────────────────────────

          Factors
State     Factor1      Factor2       Factor3
AL        -1.3536      -0.3282       -1.8275
AK         1.2778       1.3426       -0.0098
AZ         1.8702       0.6774       -0.3623
AR        -1.4028       0.3351       -1.3665
CA         0.7738      -1.9210       -1.3009
CO         2.0111       0.2099       -0.9566
CT         0.1614      -0.6961        2.0344
DE        -0.8710      -0.5216        0.5454
DC        -1.4835      -0.8404        2.3234
FL         0.5822      -0.0421        0.7152
GA        -0.5416      -1.0412       -1.8152
HI         0.3398      -2.5941       -0.0033

(report continues through all 52 rows)
```

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

Example 2b – Creating a 3D Scatter Plot of the Factor 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 (F1, F2, F3) 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

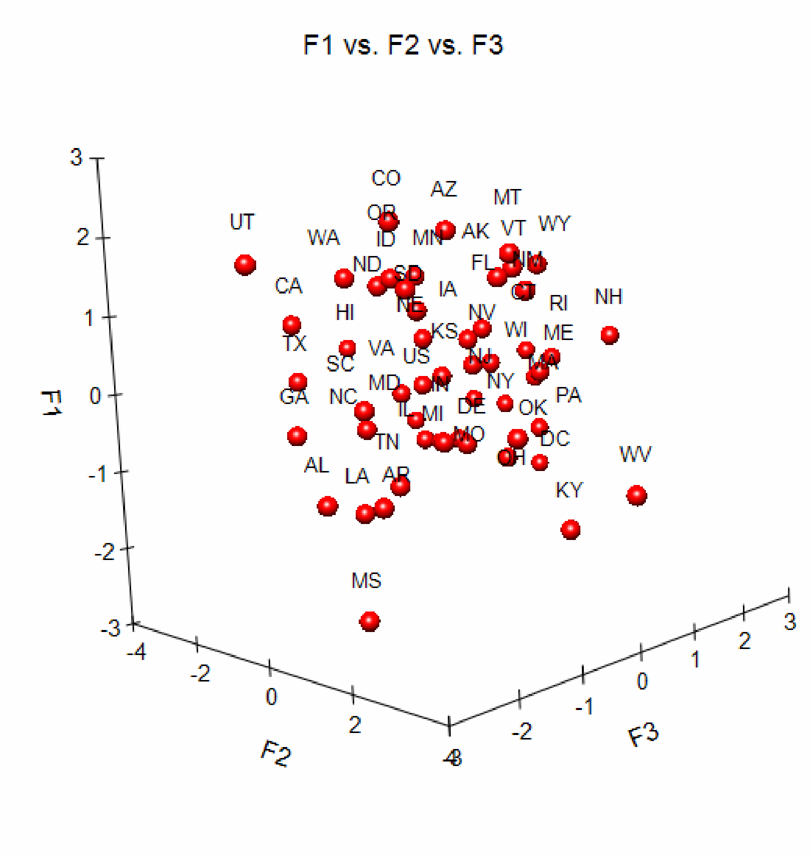
- Find and open the **3D Scatter Plot** procedure using the menus or the Procedure Navigator.
- Set **Variable Labels** to **Column Names** using the **Report Options** dropdown in the toolbar.
- The settings for this example are listed below and are stored in the **Example 2b** settings template. To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

Option	Value
Variables Tab	
X (Horizontal) Variable	**F2**
Y (Vertical) Variable	**F1**
Z (Depth) Variable	**F3**
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. 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 Factor Loadings

This example continues the analysis of the *Death Rates – States – 2016* dataset which was begun in Example 2. It will show how to store the factor 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 Factor Analysis procedure options**

- Find and open the **Factor 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**
Data Label Variable	**State**
Factor Rotation	**Varimax**
Number of Factors	**3**
Reports Tab	
Eigenvalue Summary	**Checked**
Loadings Report	**Checked**
Storage Tab	
Factor Scores	**F1, F2, F3** (We set these column names previously.)

3 **Run the procedure**

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

Eigenvalues

```
Eigenvalues after Varimax Rotation ───────────────────────────────────────
Robust Estimation:            None
Missing-Value Estimation:     None

                               Individual   Cumulative
Number      Eigenvalue            Percent      Percent     Scree Plot
1             2.803478              43.28        43.28     |||||||||
2             2.102399              32.46        75.74     |||||||
3             1.549342              23.92        99.66     |||||
4             0.308242               4.76       104.42     |
5             0.137108               2.12       106.54     |
6             0.039781               0.61       107.15     |
7             0.006154               0.10       107.24     |
8            -0.113017              -1.74       105.50     |
9            -0.140473              -2.17       103.33     |
10           -0.215772              -3.33       100.00     |
```

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

Factor Loadings

```
Factor Loadings after Varimax Rotation ───────────────────────────────────

                 Factors
Variables        Factor1       Factor2       Factor3
Alzheimers     -0.223149      0.198293     -0.661900
LowResDis      -0.348456      0.714621     -0.419938
Cancer         -0.779909      0.426796     -0.078043
Diabetes       -0.379745      0.533993     -0.389167
HeartDis       -0.760404      0.265673     -0.213001
FluPneum       -0.422246     -0.004266     -0.178371
Kidney         -0.732902      0.086983     -0.292631
Stroke         -0.613622      0.097195     -0.708299
Suicide         0.380809      0.658255     -0.330287
Accidents      -0.254164      0.751119      0.061192
```

This report shows again the factor loadings which are the correlations between each factor (Factor1, Factor2, Factor3) and each variable (Alzheimers, LowResDis, etc.). A quick scan of these values will give you an idea of which variables are most highly correlated with each factor.

Example 3b – Creating a 3D Scatter Plot of the Factor 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 Factor 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

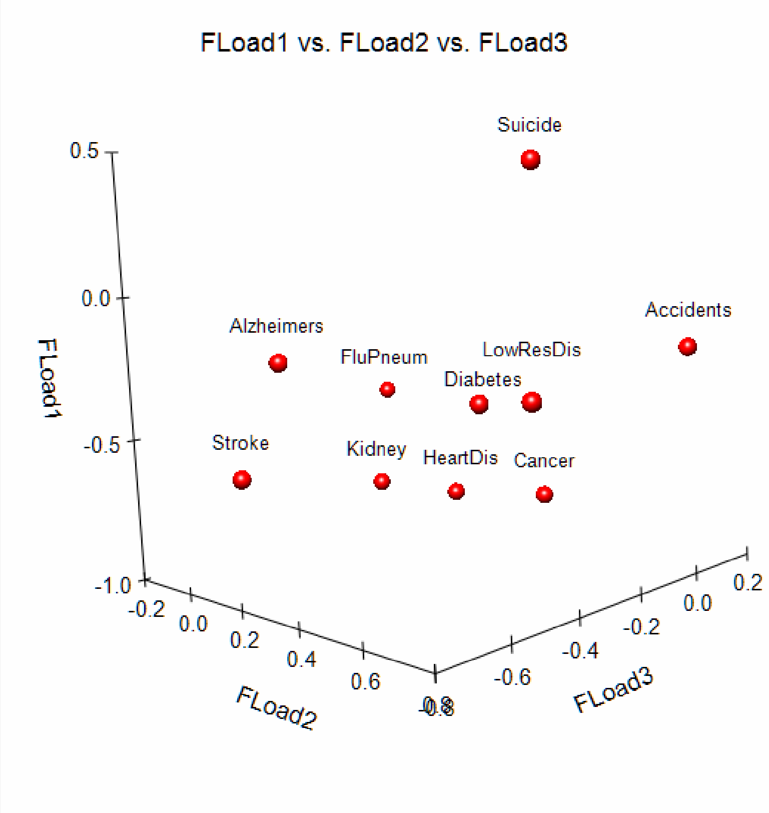
- Find and open the **3D Scatter Plot** procedure using the menus or the Procedure Navigator.
- Set **Variable Labels** to **Column Names** using the **Report Options** dropdown in the toolbar.
- The settings for this example are listed below and are stored in the **Example 3b** settings template. To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

Option	Value
Variables Tab	
X (Horizontal) Variable	**FLoad2**
Y (Vertical) Variable	**FLoad1**
Z (Depth) Variable	**FLoad3**
Data Label Variable	**FVariable**
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, and Kidney Disease have similar loadings.

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