Chapter 401

Correlation Matrix

Introduction

This program calculates matrices of **Pearson product-moment correlations** and **Spearman-rank correlations**. It allows missing values to be deleted in a pair-wise or row-wise fashion.

When someone speaks of a correlation matrix, they usually mean a matrix of Pearson-type correlations. Unfortunately, these correlations are unduly influenced by outliers, unequal variances, nonnormality, and nonlinearities. One of the chief competitors of the Pearson correlation coefficient is the Spearman-rank correlation coefficient. The Spearman correlation is calculated by applying the Pearson correlation formula to the ranks of the data. In so doing, many of the distortions that infect the Pearson correlation are reduced considerably.

A matrix of differences can be displayed to compare the two types of correlation matrices. This allows you to determine which pairs of variables require further investigation.

Partial Correlation

This program lets you specify an optional set of *partial* variables. The linear influence of these variables is removed from the correlation matrix. This provides a statistical adjustment to the correlations among the remaining variables using multiple regression. Note that in the case of Spearman correlations, this adjustment occurs after the complete correlation matrix has been formed.

Heat Maps

Using heat maps to display the features of a correlation matrix was the topic of Friendly (2002) and Friendly and Kwan (2003). This program generates a heat map for various correlation matrices.

Plots of Eigenvectors

Friendly (2002) and Friendly and Kwan (2003) discuss the strengths of plotting the eigenvectors of a correlation matrix. They imply that such a plot is more informative than a heat map. This program generates a plot of the eigenvectors for various correlation matrices.

Another plot that is similar to the eigenvector plot is the map which is provided by a *metric multidimensional scaling* analysis (see the *Multidimensional Scaling* procedure for details).

Discussion

When there is more than one independent variable, the collection of all pair-wise correlations are succinctly represented in a matrix form. In regression analysis, the purpose of examining these correlations is two-fold: to find outliers and to identify collinearity. In the case of outliers, there should be major differences between the parametric measure, the Pearson correlation coefficient, and the nonparametric measure, the Spearman rank correlation coefficient. In the case of collinearity, high pair-wise correlations could be indicators of collinearity problems.

The Pearson correlation coefficient is unduly influenced by outliers, unequal variances, nonnormality, and nonlinearities. As a result of these problems, the Spearman correlation coefficient, which is based on the ranks of the data rather than the actual data, may be a better choice for examining the relationships between variables.

Finally, the patterns of missing values in multiple regression and correlation analysis can be very complex. As a result, missing values can be deleted in a pair-wise or a row-wise fashion. If there are only a few observations with missing values, it might be preferable to use the row-wise deletion, especially for large data sets. The row-wise deletion procedure omits the entire observation from the analysis.

On the other hand, if the pattern of missing values is randomly dispersed throughout the data and the use of the row-wise deletion would omit at least 25% of the observations, the pair-wise deletion procedure for missing values would be a safer way to capture the essence of the relationships among the variables. While this method appears to make full use of all your data, the resulting correlation matrix may have mathematical and interpretation difficulties. Mathematically, this correlation matrix may not have a positive determinant. Since each correlation may be based on a different set of rows, practical interpretations could be difficult, if not illogical.

The Spearman correlation coefficient measures the monotonic association between two variables in terms of ranks. It measures whether one variable increases or decreases with another even when the relationship between the two variables is not linear or bivariate normal. Computationally, each of the two variables is ranked separately, and the ordinary Pearson correlation coefficient is computed on the ranks. This nonparametric correlation coefficient is a good measure of the association between two variables when outliers, nonnormality, nonconstant variance, and nonlinearity may exist between the two variables being investigated.

Data Structure

The data are entered as two or more variables. An example of data appropriate for this procedure is shown in the table below. It is assumed that each row gives measurements on the same individual.

Test Scores

Test 1	Test 2	Test 3
45	54	78
87	92	58
55	77	88
44	46	53
73	45	
75	66	66
93	46	85
57	78	91
66	58	77
68	53	73
	45	68
54	65	65
	65	
59	66	72
	54	83
75	53	82

Example 1 – Creating a Correlation Matrix

This section presents an example of how to run an analysis of the data contained in the IQ dataset.

Setup

To run this example, complete the following steps:

1 Open the IQ example dataset

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

2 Specify the Correlation Matrix procedure options

- Find and open the **Correlation Matrix** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab	
Correlation Variables	**Test1, Test2, Test3, Test4, Test5, IQ**
Missing Value Removal	**Row-Wise**

Reports Tab	
Show Individual Tables	**Checked**
Pearson Correlations	**Checked**
Spearman Correlations	**Checked**
Difference	**Checked**
Show Combined Table	**Checked**
Pearson Correlations	**Checked**
Spearman Correlations	**Checked**
Pearson P-Value	**Checked**
Count	**Checked**
Add Cronbach's Alpha…	**Checked**

3 Run the procedure

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

Individual Reports

Pearson Correlation Report

Row-Wise Missing Value Deletion

Variables	Test1	Test2	Test3	Test4	Test5	IQ
Test1	1.0000	0.1000	-0.2608	0.7539	0.0140	0.2256
Test2	0.1000	1.0000	0.0572	0.7196	-0.2814	0.2407
Test3	-0.2608	0.0572	1.0000	-0.1409	0.3473	0.0741
Test4	0.7539	0.7196	-0.1409	1.0000	-0.1729	0.3714
Test5	0.0140	-0.2814	0.3473	-0.1729	1.0000	-0.0581
IQ	0.2256	0.2407	0.0741	0.3714	-0.0581	1.0000

Coefficient Alpha

Cronbach's Alpha	0.4519
Standardized Cronbach's Alpha	0.4785

Spearman Correlation Report

Row-Wise Missing Value Deletion

Variables	Test1	Test2	Test3	Test4	Test5	IQ
Test1	1.0000	0.0098	-0.3539	0.6517	0.0000	0.2202
Test2	0.0098	1.0000	0.0430	0.6971	-0.3118	0.2303
Test3	-0.3539	0.0430	1.0000	-0.2143	0.3982	0.1238
Test4	0.6517	0.6971	-0.2143	1.0000	-0.1577	0.3772
Test5	0.0000	-0.3118	0.3982	-0.1577	1.0000	-0.0125
IQ	0.2202	0.2303	0.1238	0.3772	-0.0125	1.0000

Difference (Pearson - Spearman) Report

Row-Wise Missing Value Deletion

Variables	Test1	Test2	Test3	Test4	Test5	IQ
Test1	0.0000	0.0902	0.0931	0.1022	0.0140	0.0054
Test2	0.0902	0.0000	0.0142	0.0225	0.0304	0.0104
Test3	0.0931	0.0142	0.0000	0.0734	-0.0509	-0.0497
Test4	0.1022	0.0225	0.0734	0.0000	-0.0152	-0.0058
Test5	0.0140	0.0304	-0.0509	-0.0152	0.0000	-0.0455
IQ	0.0054	0.0104	-0.0497	-0.0058	-0.0455	0.0000

The above tables display the Pearson Correlation Report, Spearman Correlation Report, and the Difference Report. Cronbach's Alpha is displayed at the bottom of the first report.

The Difference report displays the difference between the Pearson and the Spearman correlation coefficients. The report lets you find those variable pairs for which these two correlation coefficients are very different. A large difference indicates the presence of outliers, nonlinearity, nonnormality, and the like. You should investigate scatter plots of pairs of variables with large differences.

Reliability

Because of the central role of measurement in science, scientists of all disciplines are concerned with the accuracy of their measurements. Item analysis is a methodology for assessing the accuracy of measurements that are obtained in the social sciences where precise measurements are often hard to secure. The accuracy of a measurement may be broken down into two main categories: validity and reliability. The validity of an instrument refers to whether it accurately measures the attribute of interest. The reliability of an instrument concerns whether it produces identical results in repeated applications. An instrument may be reliable but not valid. However, it cannot be valid without being reliable.

The methods described here assess the reliability of an instrument. They do not assess its validity. This should be kept in mind when using the techniques of item analysis since they address reliability, not validity.

An instrument may be valid for one attribute but not for another. For example, a driver's license exam may accurately measure an individual's ability to drive. However, it does not accurately measure that individual's ability to do well in college. Hence the exam is reliable and valid for measuring driving ability. It is reliable and invalid for measuring success in college.

Several methods have been proposed for assessing the reliability of an instrument. These include the retest method, alternative-form method, split-halves method, and the internal consistency method. We will focus on internal consistency here.

Cronbach's Alpha

Cronbach's alpha (or *coefficient alpha*) is the most popular of the internal consistency coefficients. It is calculated as follows.

$$\alpha = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K}\sigma_{ii}}{\sum_{i=1}^{K}\sum_{j=1}^{K}\sigma_{ij}}\right]$$

where K is the number of items (questions) and σ_{ij} is the estimated covariance between items i and j. Note the σ_{ii} is the variance (not standard deviation) of item i.

If the data are standardized by subtracting the item means and dividing by the item standard deviations before the above formula is used, we obtain the standardized version of Cronbach's alpha. A little algebra will show that this is equivalent to the following calculations based directly on the correlation matrix of the items.

$$\alpha = \frac{K\bar{\rho}}{1 + \bar{\rho}(K-1)}$$

where K is the number of items (variables) and $\bar{\rho}$ is the average of all the correlations among the K items.

Cronbach's alpha has several interpretations. It is equal to the average value of alpha coefficients obtained for all possible combinations of dividing $2K$ items into two groups of K items each and calculating the two-half tests. Also, alpha estimates the expected correlation of one instrument with an alternative form containing the same number of items. Furthermore, alpha estimates the expected correlation between an actual test and a hypothetical test which may never be written.

Since Cronbach's alpha is supposed to be a correlation, it should range between -1 and 1. However, it is possible for alpha to be less than -1 when several of the covariances are relatively large, negative numbers. In most cases, alpha is positive, although negative values arise occasionally.

What value of alpha should be achieved? Carmines (1990) stipulates that as a rule, a value of at least 0.8 should be achieved for widely used instruments. An instrument's alpha value may be improved by either adding more items or by increasing the average correlation among the items.

Combined Report

Combined Correlation Report

Row-Wise Missing Value Deletion

Variables		Test1	Test2	Test3	Test4	Test5	IQ
Test1	Pearson Correlation	1.0000	0.1000	-0.2608	0.7539	0.0140	0.2256
	Spearman Correlation	1.0000	0.0098	-0.3539	0.6517	0.0000	0.2202
	Pearson P-Value		0.7228	0.3478	0.0012	0.9606	0.4187
	Count	15	15	15	15	15	15
Test2	Pearson Correlation	0.1000	1.0000	0.0572	0.7196	-0.2814	0.2407
	Spearman Correlation	0.0098	1.0000	0.0430	0.6971	-0.3118	0.2303
	Pearson P-Value	0.7228		0.8395	0.0025	0.3095	0.3876
	Count	15	15	15	15	15	15
Test3	Pearson Correlation	-0.2608	0.0572	1.0000	-0.1409	0.3473	0.0741
	Spearman Correlation	-0.3539	0.0430	1.0000	-0.2143	0.3982	0.1238
	Pearson P-Value	0.3478	0.8395		0.6164	0.2046	0.7931
	Count	15	15	15	15	15	15
Test4	Pearson Correlation	0.7539	0.7196	-0.1409	1.0000	-0.1729	0.3714
	Spearman Correlation	0.6517	0.6971	-0.2143	1.0000	-0.1577	0.3772
	Pearson P-Value	0.0012	0.0025	0.6164		0.5378	0.1729
	Count	15	15	15	15	15	15
Test5	Pearson Correlation	0.0140	-0.2814	0.3473	-0.1729	1.0000	-0.0581
	Spearman Correlation	0.0000	-0.3118	0.3982	-0.1577	1.0000	-0.0125
	Pearson P-Value	0.9606	0.3095	0.2046	0.5378		0.8371
	Count	15	15	15	15	15	15
IQ	Pearson Correlation	0.2256	0.2407	0.0741	0.3714	-0.0581	1.0000
	Spearman Correlation	0.2202	0.2303	0.1238	0.3772	-0.0125	1.0000
	Pearson P-Value	0.4187	0.3876	0.7931	0.1729	0.8371	
	Count	15	15	15	15	15	15

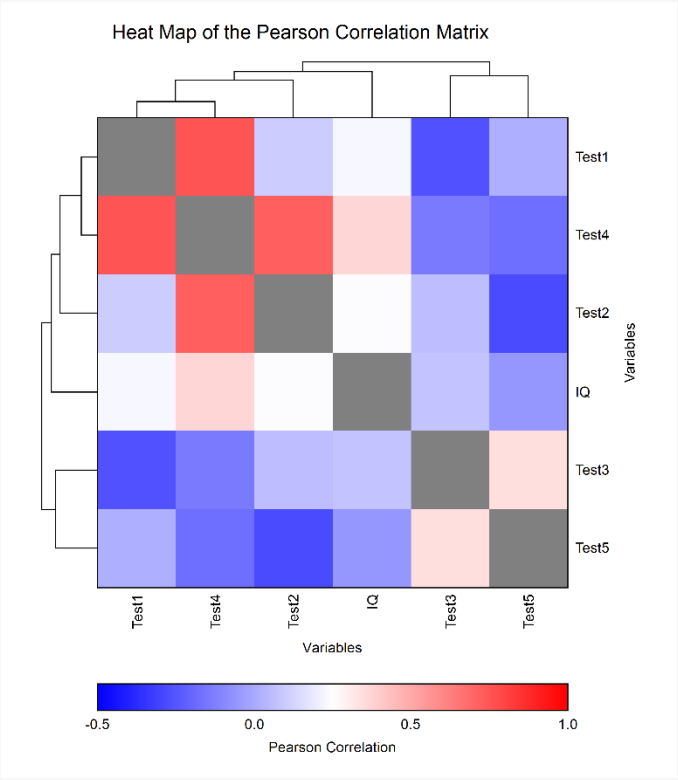
Coefficient Alpha

Cronbach's Alpha	0.4519
Standardized Cronbach's Alpha	0.4785

The above report displays the Pearson and Spearman correlations, the significance level of a test of the Pearson correlation (Pearson P-Value) and count for each pair of variables.

Heat Map of the Pearson Correlation Matrix

Heat Map of the Pearson Correlation Matrix



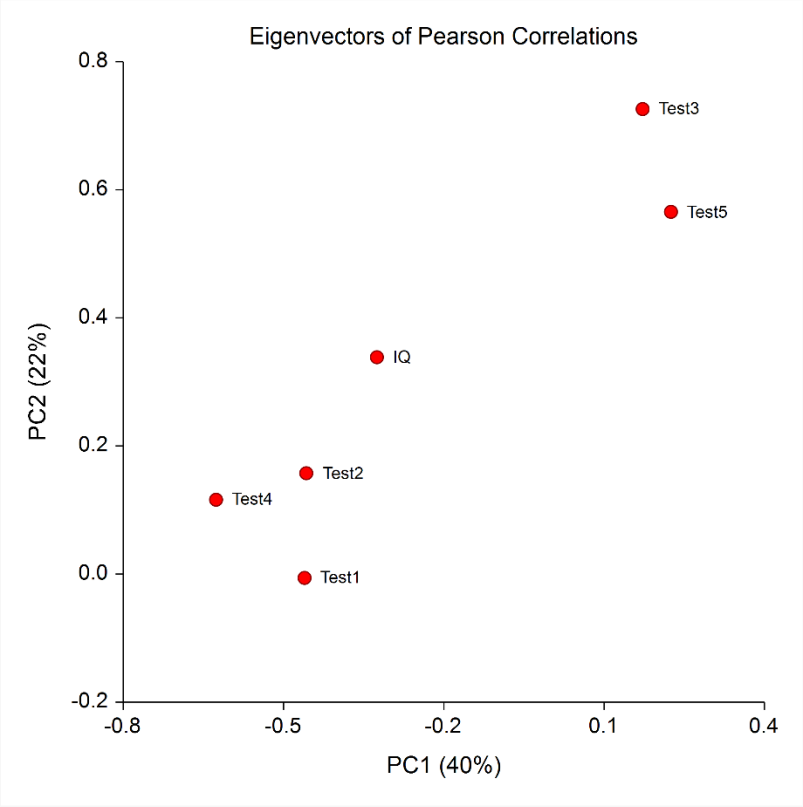
This report displays a heat map of the correlation matrix. Note that the rows and columns are sorted in the order suggested by the hierarchical clustering.

This plot allows you to discover various subsets of the variables that seem to be highly correlated within the subset. You can see that Test1, Test4, and Test2 seem to be highly related. Similarly, Test3 and Test5 seem to be related.

This plot was suggested by Friendly (2002) and Friendly and Kwan (2003).

Pearson Eigenvectors Plot(s)

Pearson Eigenvectors Plot(s)



This plot displays a scatter plot of PC1 (the first eigenvector) on the horizontal axis and PC2 (the second eigenvector) on the vertical axis. The number within the parentheses is the percentage of the sum of the eigenvalues that that is accounted for by the corresponding eigenvector. For example, in this plot, 40% of the variability in the correlation matrix is accounted for by the first eigenvector and 22% of the variability is accounted for by the second eigenvector. Thus, the two eigenvectors in this plot account for 62% of the variation among the correlations.

Note that this plot lets you see which variables to be clustered. In this case, Test3 and Test5 are related as are Test1, Test2, and Test4. The IQ variable seems to be by itself, although it is somewhat similar to the second three variables.

This is the same interpretation that we obtained from the heat map, but perhaps it is easier to see subtleties in this plot.

This plot was suggested by Friendly (2002) and Friendly and Kwan (2003).

Storing the Correlations on the Database

When you specify variables in either the Pearson Correlations or the Spearman Correlations boxes, the correlation matrix will be stored in those variables during the execution of the program.

Example 2 – Bartlett's Sphericity Test

This section presents an example of how to run Bartlett's Sphericity test of the data contained in the IQ dataset. Note that Bartlett's test is only available when Missing Value Removal is set to *Row Wise*.

Setup

To run this example, complete the following steps:

1 Open the IQ example dataset

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

2 Specify the Correlation Matrix procedure options

- Find and open the **Correlation Matrix** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab	
Correlation Variables	**Test1, Test2, Test3, Test4, Test5, IQ**
Missing Value Removal	**Row-Wise**
Reports Tab	
Show Individual Tables	**Checked**
Pearson Correlations	**Checked**
Eigenvectors Tab	
Pearson Eigenvector Plot(s)	**Checked**
Show the eigenvalue…	**Checked**
Eigenvalues and Eigenvectors of…	**Checked**

3 Run the procedure

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

Individual Reports

Pearson Correlation Report

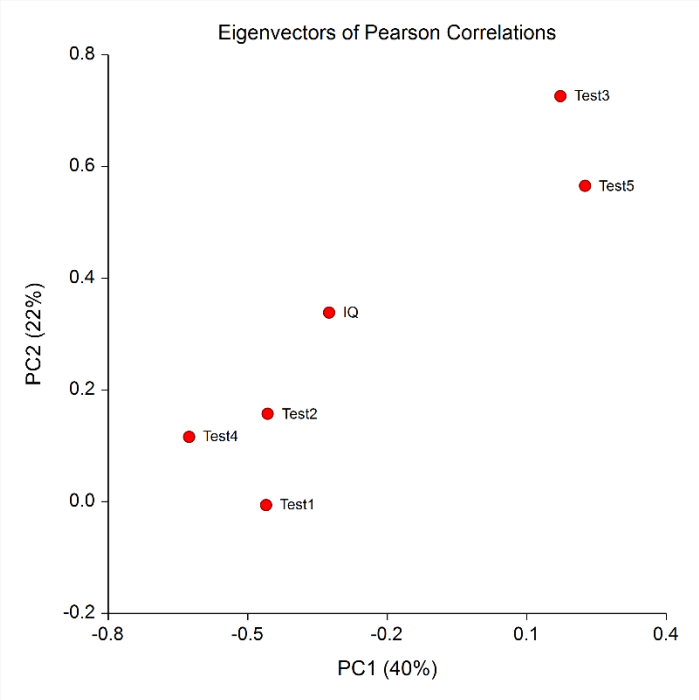
Row-Wise Missing Value Deletion

Variables	Test1	Test2	Test3	Test4	Test5	IQ
Test1	1.0000	0.1000	-0.2608	0.7539	0.0140	0.2256
Test2	0.1000	1.0000	0.0572	0.7196	-0.2814	0.2407
Test3	-0.2608	0.0572	1.0000	-0.1409	0.3473	0.0741
Test4	0.7539	0.7196	-0.1409	1.0000	-0.1729	0.3714
Test5	0.0140	-0.2814	0.3473	-0.1729	1.0000	-0.0581
IQ	0.2256	0.2407	0.0741	0.3714	-0.0581	1.0000

The above table displays the Pearson Correlation Report.

Pearson Eigenvector Plot

Pearson Eigenvectors Plot(s)



This plot displays a scatter plot of PC1 (the first eigenvector) on the horizontal axis and PC2 (the second eigenvector) on the vertical axis. The number within the parentheses is the percentage of the sum of the eigenvalues that that is accounted for by the corresponding eigenvector. For example, in this plot, 40% of the variability in the correlation matrix is accounted for by the first eigenvector and 22% of the variability is accounted for by the second eigenvector. Thus, the two eigenvectors in this plot account for 62% of the variation among the correlations.

Note that this plot lets you see which variables could be clustered. In this case, Test3 and Test5 are related as are Test1, Test2, and Test4. The IQ variable seems to be by itself, although it is somewhat similar to the second three variables.

Eigenvalues Report

Eigenvalues of Pearson Correlation Matrix

Row-Wise Missing Value Deletion

Eigenvector	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
PC1	2.374012	39.57	39.57	||||||||
PC2	1.297129	21.62	61.19	|||||
PC3	1.109029	18.48	79.67	||||
PC4	0.779485	12.99	92.66	|||
PC5	0.435845	7.26	99.92	||
PC6	0.004500	0.08	100.00	|

Matrix Summary Measures

Log(Det|R)	-5.254947

Bartlett Sphericity Test

Test Statistic	58.68
DF	15
Prob Level	0.000000

The above report displays the Pearson and Spearman correlations, the significance level of a test of the Pearson correlation (Pearson P-Value) and count for each pair of variables.

Eigenvector

This column gives the label of the eigenvector whose eigenvalue is displayed. Note that you can modify the label.

Eigenvalue

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

One rule-of-thumb is to retain those eigenvectors whose eigenvalues are greater than one. The sum of the eigenvalues is equal to the number of variables. Hence, in this example, the first eigenvector retains the information contained in 2.37 of the original variables.

Individual and Cumulative Percents

The first column gives the percentage of the total variation in the variables accounted for by this eigenvector. 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 eigenvectors to plot.

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 eigenvectors that make up the cliff. Cattell and Jaspers (1967) suggest keeping those that make up the cliff plus the first eigenvector of the rubble.

Log(Det|R|)

This is the log (base e) of the determinant of the correlation matrix.

Bartlett Test, DF, Prob Level

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 significance level (Prob Level) greater than 0.05, there is no evidence that any of the correlations are different from zero. 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 the Missing Value Removal option is set to *Row Wise*.

The formula for computing this test is:

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

Eigenvectors Report

Eigenvectors of Pearson Correlation Matrix

Row-Wise Missing Value Deletion

	Eigenvectors	
Variables	**PC1**	**PC2**
Test1	-0.4608	-0.0060
Test2	-0.4575	0.1575
Test3	0.1720	0.7261
Test4	-0.6263	0.1161
Test5	0.2251	0.5656
IQ	-0.3253	0.3386

The eigenvectors show the direction of each factor (principal component) after the correlation matrix is suitably scaled and rotated. These are the values that are plotted in the Eigenvector plots shown above.