Chapter 400

Canonical Correlation

Introduction

Canonical correlation analysis is the study of the linear relations between two sets of variables. It is the multivariate extension of correlation analysis. Although we will present a brief introduction to the subject here, you will probably need a text that covers the subject in depth such as Tabachnick (1989).

Suppose you have given a group of students two tests of ten questions each and wish to determine the overall correlation between these two tests. Canonical correlation finds a weighted average of the questions from the first test and correlates this with a weighted average of the questions from the second test. The weights are constructed to maximize the correlation between these two averages. This correlation is called the first canonical correlation coefficient.

You can create another set of weighted averages unrelated to the first and calculate their correlation. This correlation is the second canonical correlation coefficient. This process continues until the number of canonical correlations equals the number of variables in the smallest group.

Discriminant analysis, MANOVA, and multiple regression are all special cases of canonical correlation. It provides the most general multivariate framework. Because of this generality, it is probably the least used of the multivariate procedures. Researchers would rather use the specific procedure designed for their data. However, there are instances when canonical correlation techniques are useful.

Variates and Variables

Canonical correlation terminology makes an important distinction between the words variables and variates. The term *variables* is reserved for referring to the original variables being analyzed. The term *variates* is used to refer to variables that are constructed as weighted averages of the original variables. Thus, a set of Y variates is constructed from the original Y variables. Likewise, a set of X variates is constructed from the original X variables.

Basic Issues

Some of the issues that must be dealt with during a canonical correlation analysis are:

1. Determining the number of canonical variate pairs to use. The number of pairs possible is equal to the smaller of the number of variables in each set.
2. The canonical variates themselves often need to be interpreted. As in factor analysis, you are dealing with mathematically constructed variates that are usually difficult to interpret. However, in this case, you must relate two constructed variates to each other.
3. The importance of each variate must be evaluated from two points of view. You have to determine the strength of the relationship between the variate and the variables from which it was created. You also need to study the strength of the relationship between the corresponding X and Y variates.

4. Do you have a large enough sample size? In social science work you will often need a minimum of ten cases per variable. In fields with more reliable data, you can get by with a little less.

Canonical Correlation Checklist

Tabachnick (1989) provides the following checklist for conducting a canonical correlation analysis. We suggest that you consider these issues and guidelines carefully.

Missing Data

You should begin by screening your data for outliers. Pay particular attention to patterns of missing values. The program ignores rows with missing values. If it appears that most of the missing values occur in one or two variables, you might want to leave these out of the analysis in order to obtain more data on the remaining variables.

Multivariate Normality and Outliers

Canonical correlation analysis does not make strong normality assumptions. However, as with all least squares procedures, outliers can cause severe problems. You should screen your data carefully for outliers using the various univariate normality tests and plots.

Linearity

Canonical correlation analysis assumes linear relations among the variables. You should study scatter plots of each pair of variables, watching carefully for curvilinear patterns and for outliers. The occurrence of curvilinear relationship will reduce the effectiveness of the analysis.

Multicollinearity and Singularity

Multicollinearity occurs when one variable is almost a weighted average of the others. Singularity occurs when this relationship is exact. Since inverse matrices are needed during the analysis, you must check for this. Try running a principal components analysis on each set of variables, separately. If you have eigenvalues at or near zero, you have multicollinearity problems. You must omit the offending variables.

Technical Details

As the name suggests, canonical correlation analysis is based on the correlations between two sets of variables which we call **Y** and **X**.

The correlation matrix of all the variables is divided into four parts:

1. R_{xx}. The correlations among the **X** variables.
2. R_{yy}. The correlations among the **Y** variables.
3. R_{xy}. The correlations between the **X** and **Y** variables.
4. R_{yx}. The correlations between the **Y** and **X** variables.

Canonical correlation analysis may be defined using the singular value decomposition of a matrix ***C*** where:

$$C = R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy}$$

Define the singular value decomposition of ***C*** as:

$$C = U' \Lambda \hat{B}$$

The diagonal matrix Λ of the singular values of ***C*** is made up of the eigenvalues of ***C***. The *i*th eigenvalue λ_i of the matrix ***C*** is equal to the square of the *i*th canonical correlation which is called r_{ci}^2. Hence, the *i*th canonical correlation is the square root of the *i*th eigenvalue of ***C***.

Two sets of canonical coefficients (like regression coefficients) are used for each canonical correlation: one for the **X** variables and another for the **Y** variables. These coefficients are defined as follows:

$$B_y = R_{yy}^{-1/2} \hat{B}$$

$$B_x = \Lambda R_{xx}^{-1} R_{xy} B_y$$

The canonical scores for **X** and **Y** (denoted $\hat{X}$ and $\hat{Y}$) are calculated by multiplying the standardized data (subtract the mean and divide by the standard deviation) by these coefficient matrices. Thus, we have:

$$\hat{X} = Z_x B_x$$

and

$$\hat{Y} = Z_y B_y$$

where Z_x and Z_y represent the standardized versions of **X** and **Y**.

To aid in the interpretation of the canonical variates, loading matrices are computed. These are the correlations between the original variables and the constructed variates. They are computed as follows:

$$A_x = R_{xx} B_x$$

$$A_y = R_{yy} B_y$$

The *average squared loadings* are given by

$$pv_{xc} = 100 \sum_{i=1}^{k_x} \frac{a_{ixc}^2}{k_x}$$

$$pv_{yc} = 100 \sum_{i=1}^{k_y} \frac{a_{iyc}^2}{k_y}$$

The *redundancy indices* are given by:

$$rd = (pv)(r_c^2)$$

Data Structure

The data are entered in the standard columnar format in which each column represents a single variable.

Missing Values

Rows with missing values in any of the variables used in the analysis are ignored.

Example 1 – Canonical Correlation Analysis

This section presents an example of how to run a canonical correlation analysis using data contained on the Tests dataset. As an example, we will correlate variables Test1, Test2, and Test3 with variables Test4, Test5, and IQ.

Setup

To run this example, complete the following steps:

1 Open the Tests example dataset

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

2 Specify the Canonical Correlation procedure options

- Find and open the **Canonical Correlation** 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

Y Variables **Test4-IQ**
X Variables **Test1-Test3**

Reports Tab

Number of Correlations.................................... **3**
All Other 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.

Descriptive Statistics

Descriptive Statistics

Type	Variable	Mean	Standard Deviation	Non-Missing Rows
Y	Test4	65.53333	13.95332	15
Y	Test5	69.93333	16.15314	15
Y	IQ	104.3333	11.0173	15
X	Test1	67.93333	17.39239	15
X	Test2	61.4	19.39735	15
X	Test3	72.33334	14.73415	15

This report displays the descriptive statistics for each variable. You should check that the mean is reasonable and that the number of nonmissing rows is accurate.

Correlations

Correlations

	Variable					
Variable	**Test4**	**Test5**	**IQ**	**Test1**	**Test2**	**Test3**
Test4	1.000000	-0.172864	0.371404	0.753937	0.719623	-0.140941
Test5	-0.172864	1.000000	-0.058064	0.013967	-0.281449	0.347335
IQ	0.371404	-0.058064	1.000000	0.225648	0.240651	0.074070
Test1	0.753937	0.013967	0.225648	1.000000	0.100018	-0.260801
Test2	0.719623	-0.281449	0.240651	0.100018	1.000000	0.057232
Test3	-0.140941	0.347335	0.074070	-0.260801	0.057232	1.000000

This report presents the simple correlations among all variables specified.

Canonical Correlations

Canonical Correlations

				Degrees of Freedom (DF)			
Variate Number	**Canonical Correlation**	**R-Squared**	**F-Value**	**Numerator**	**Denominator**	**P-Value**	**Wilks' Lambda**
1	0.995600	0.991219	16.58	9	22	0.000000	0.006819
2	0.467461	0.218519	0.67	4	20	0.617695	0.776503
3	0.079810	0.006370	0.07	1	11	0.795498	0.993630

F-value tests whether this canonical correlation and those following are zero.

This report presents the canonical correlations plus supporting material to aid in their interpretation.

Variate Number

This is the sequence number of the canonical correlation. Remember that the first correlation will be the largest, the second will be the next to largest, and so on.

Canonical Correlation

The value of the canonical correlation coefficient. This coefficient has the same properties as any other correlation: it ranges between minus one and one, a value near zero indicates low correlation, and an absolute value near one indicates near perfect correlation.

R-Squared

The square of the canonical correlation coefficient. This gives the R-squared value of fitting the Y canonical variate to the corresponding X canonical variate.

F-Value

The value of the F approximation for testing the significance of the Wilks' lambda corresponding to this row and those below it. In this example, the first F-Value tests the significance of the first, second, and third canonical correlations while the second F-value tests the significance of only the second and third.

Numerator Degrees of Freedom (DF)

The numerator degrees of freedom of the above F-ratio.

Denominator Degrees of Freedom (DF)

The denominator degrees of freedom of the above F-ratio.

P-Value

This is the p-value for the above F statistic. A value near zero indicates a significant canonical correlation. A cutoff value of 0.05 or 0.01 is often used to determine significance.

Wilks' Lambda

The Wilks' lambda value for the canonical correlation on this report row. Wilks' lambda is the multivariate generalization of R-Squared. The Wilks' lambda statistic is interpreted just the opposite of R-Squared: a value near zero indicates high correlation while a value near one indicates low correlation.

Variation Explained

Variation Explained

Canonical Variate Number	Variation in these Variables	Explained by these Variates	Percent Explained		Canonical Correlation Squared
			Individual	Cumulative	
1	Y	Y	37.6	37.6	0.9912
2	Y	Y	32.1	69.7	0.2185
3	Y	Y	30.3	100.0	0.0064
1	Y	X	37.2	37.2	0.9912
2	Y	X	7.0	44.3	0.2185
3	Y	X	0.2	44.5	0.0064
1	X	Y	37.1	37.1	0.9912
2	X	Y	5.4	42.5	0.2185
3	X	Y	0.2	42.8	0.0064
1	X	X	37.4	37.4	0.9912
2	X	X	24.8	62.2	0.2185
3	X	X	37.8	100.0	0.0064

This report displays the percent of the variation in each set of variables explained by other sets of variables.

Canonical Variate Number

This is the sequence number of the canonical variable being reported on. Remember that the maximum number of variates is the minimum of the number of variables in each set.

Variation in these Variables

Each row of the report presents the results of how well a set of variables is explained by a particular canonical variate. This column designates which set of variables is being reported on.

Explained by these Variates

Each row of the report presents the results of how well a set of variables is explained by a particular canonical variate. This column designates which set of canonical variates is being reported on.

Individual Percent Explained

This column indicates the percentage of the variation in the designated set of variables that is explained by this canonical variate.

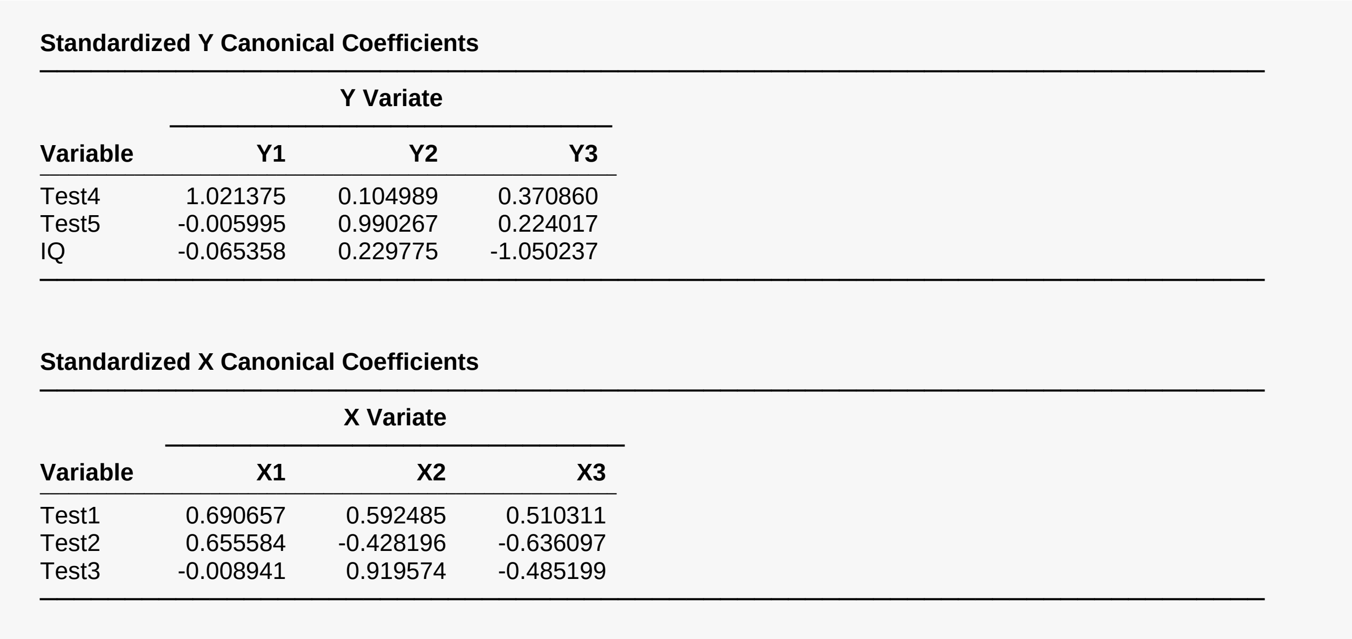
Cumulative Percent Explained

This column indicates the cumulative percentage of the variation in the designated set of variables that is explained by this canonical variate and those listed above it.

Canonical Correlation Squared

The square of the canonical correlation coefficient. This is repeated from an earlier report.

Standardized Canonical Coefficients



Standardized Y Canonical Coefficients

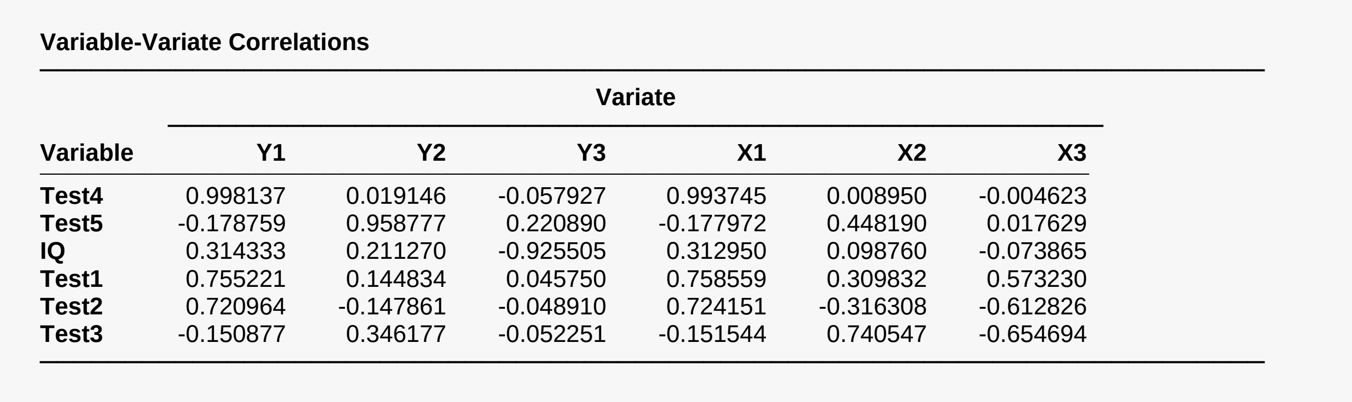
	Y Variate		
Variable	**Y1**	**Y2**	**Y3**
Test4	1.021375	0.104989	0.370860
Test5	-0.005995	0.990267	0.224017
IQ	-0.065358	0.229775	-1.050237

Standardized X Canonical Coefficients

	X Variate		
Variable	**X1**	**X2**	**X3**
Test1	0.690657	0.592485	0.510311
Test2	0.655584	-0.428196	-0.636097
Test3	-0.008941	0.919574	-0.485199

These coefficients are used to estimate the standardized scores for the X and Y variates. They aid the interpretation of the variates by showing the weight given each variable in the construction of the variate. They are analogous to standardized beta coefficients in multiple regression.

Variable-Variate Correlations



Variable-Variate Correlations

	Variate					
Variable	**Y1**	**Y2**	**Y3**	**X1**	**X2**	**X3**
Test4	0.998137	0.019146	-0.057927	0.993745	0.008950	-0.004623
Test5	-0.178759	0.958777	0.220890	-0.177972	0.448190	0.017629
IQ	0.314333	0.211270	-0.925505	0.312950	0.098760	-0.073865
Test1	0.755221	0.144834	0.045750	0.758559	0.309832	0.573230
Test2	0.720964	-0.147861	-0.048910	0.724151	-0.316308	-0.612826
Test3	-0.150877	0.346177	-0.052251	-0.151544	0.740547	-0.654694

This report shows the correlations between the variables and the variates. By determining which variables are highly correlated with a particular variate, it is hoped that you can determine its interpretation. For example, you can see that variate Y1 is highly correlated with Test4. Hence, we assume that Y1 has the same interpretation as Test4.

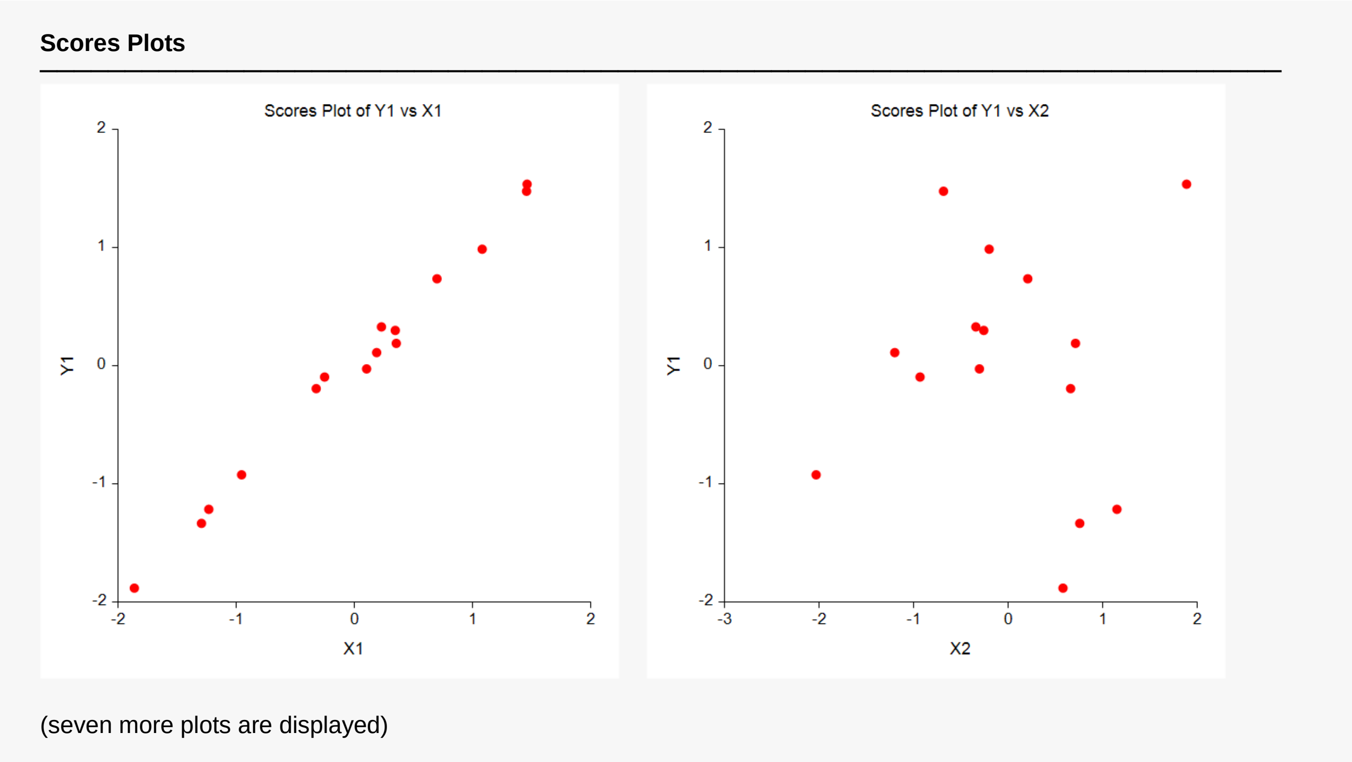
Scores

Scores

	Variate					
Row	**Y1**	**Y2**	**Y3**	**X1**	**X2**	**X3**
1	-0.193124	-0.348044	-0.308495	-0.323303	0.660431	1.582089
2	-1.214743	0.350598	0.877022	-1.232224	1.150186	1.517131
3	-0.026336	0.135325	0.250782	0.103271	-0.304012	-1.369888
4	1.536744	1.992049	-0.657871	1.461462	1.887123	-0.138798
5	0.189923	0.709643	0.455333	0.354314	0.711949	0.757851
6	0.986597	-0.677646	0.115011	1.081350	-0.201044	0.489839
7	0.299464	-0.490602	0.708912	0.345665	-0.258540	0.491428
8	-0.922687	0.503305	1.011073	-0.954587	-2.031644	0.963769
9	-1.881691	-0.288458	0.308479	-1.862181	0.579830	-0.951854
10	-1.333760	0.829021	-1.015632	-1.294283	0.756978	-1.297593
11	0.111861	-1.151067	-2.741954	0.188193	-1.199877	0.707092
12	0.329061	1.555086	-0.579356	0.228934	-0.342184	-0.612825
13	0.736439	-1.037650	0.634374	0.698925	0.206974	-0.929772
14	1.477329	-0.513679	1.201759	1.456751	-0.684236	-0.247278
15	-0.095076	-1.567882	-0.259437	-0.252288	-0.931936	-0.961191

This report provides the canonical scores of each set of variates for each row of non-missing data. These are the values that are plotted in the score plots shown next.

Scores Plots



Scores Plots

(seven more plots are displayed)

These reports show the relationship between each pair of canonical variates. The correlation coefficient of the data in the first plot (Y1 versus X1) is the first canonical correlation coefficient.