

www.statsoft.com

- Products
- Solutions
- Buy
- Trials
- Support

Textbook Canonical Analysis

What can we help you find?

Search

Looking for info about statistics?

We wrote the book on it.
And you can read it for free!

- Elementary Concepts
- Statistics Glossary
- Basic Statistics
- ANOVA / MANOVA
- Association Rules
- Boosting Trees
- Canonical Analysis
- CHAID Analysis
- C & R Trees
- Classification Trees
- Cluster Analysis
- Correspondence Analysis
- Data Mining Techniques
- Discriminant Analysis
- Distribution Fitting
- Experimental Design
- Factor Analysis
- General Discrim. Analysis
- General Linear Models
- Generalized Additive Mod.
- Generalized Linear Mod.
- General Regression Mod.
- Graphical Techniques
- Ind. Components Analysis
- Linear Regression
- Log-Linear Analysis
- MARSplines
- Machine Learning
- Multidimensional Scaling
- Neural Networks
- Nonlinear Estimation
- Nonparametric Statistics
- Partial Least Squares
- Power Analysis
- Process Analysis
- Quality Control Charts
- Reliability / Item Analysis
- SEPATH (Structural eq.)
- Survival Analysis
- Text Mining
- Time Series / Forecasting









How to Assess the Relationship Between Variables, Canonical Analysis

- General Purpose
- Computational Methods and Results
- Assumptions
- General Ideas
- Sum Scores
- Canonical Roots/Variates
- Number of Roots
- Extraction of Roots

General Purpose of Canonical Analysis

There are several measures of correlation to express the relationship between two or more variables. For example, the standard Pearson product moment correlation coefficient (r) measures the extent to which two variables are related; there are various nonparametric measures of relationships that are based on the similarity of ranks in two variables; [Multiple Regression](#) allows one to assess the relationship between a dependent variable and a set of independent variables; [Multiple Correspondence Analysis](#) is useful for exploring the relationships between a set of categorical variables.

Canonical Correlation is an additional procedure for assessing the relationship between variables. Specifically, this analysis allows us to investigate the relationship between *two sets* of variables. For example, an educational researcher may want to compute the (simultaneous) relationship between three measures of scholastic ability with five measures of success in school. A sociologist may want to investigate the relationship between two predictors of social mobility based on interviews, with actual subsequent social mobility as measured by four different indicators. A medical researcher may want to study the relationship of various risk factors to the development of a group of symptoms. In all of these cases, the researcher is interested in the relationship between two sets of variables, and *Canonical Correlation* would be the appropriate

-  [Variance Components](#)
-  [Statistical Advisor](#)
-  [Distribution Tables](#)
-  [References Cited](#)
-  [Send Comments](#)
-  [Business Solutions](#)
-  [Free Resources](#)
-  [About Textbook](#)

method of analysis.

In the following topics, the major concepts and statistics in canonical correlation analysis are introduced. It is beneficial to be familiar with the correlation coefficient as described in [Basic Statistics](#), and the basic ideas of multiple regression as described in the [overview](#) section of *Multiple Regression*.

[To index](#)

Computational Methods and Results

Following is a review of some of the computational issues involved in canonical correlation and the major results that are commonly reported.

Eigenvalues. When extracting the canonical roots, we compute the *eigenvalues*. These can be interpreted as the proportion of variance accounted for by the correlation between the respective canonical variates. Note that the proportion here is computed relative to the variance of the canonical variates, that is, of the weighted sum scores of the two sets of variables; the eigenvalues do *not* tell how much variability is explained in either set of variables. We compute as many eigenvalues as there are canonical roots, that is, as many as the minimum number of variables in either of the two sets.

Successive eigenvalues will be of smaller and smaller size. First, we compute the weights that maximize the correlation of the two sum scores. After this first root has been extracted, we will find the weights that produce the second largest correlation between sum scores, subject to the constraint that the next set of sum scores does not correlate with the previous one, and so on.

Canonical correlations. If the square root of the eigenvalues is taken, then the resulting numbers can be interpreted as correlation coefficients. Because the correlations pertain to the canonical variates, they are called *canonical correlations*. Like the eigenvalues, the correlations between successively extracted canonical variates are smaller and smaller. Therefore, as an overall index of the canonical correlation between two sets of variables, it is customary to report the largest correlation, that is, the one for the first root. However, the other canonical variates can also be correlated in a meaningful and interpretable manner (see below).

Significance of Roots. The significance test of the canonical correlations is straightforward in principle. Simply stated, the different canonical correlations are tested, one by one, beginning with the largest one. Only those roots that are statistically significant are then retained for subsequent interpretation. Actually, the nature of the significance test is somewhat different. First, evaluate the significance of all roots combined, then of the roots remaining after removing the first root, the second root, etc.

Some authors have criticized this sequential testing procedure for the significance of canonical roots (e.g., Harris, 1976). However, this procedure was "rehabilitated" in a subsequent Monte Carlo study by Mendoza, Markos, and Gonter (1978).

In short, the results of that study showed that this testing procedure will detect strong canonical correlations most of the time, even with samples of relatively small size (e.g., $n = 50$). Weaker canonical correlations (e.g., $R = .3$) require larger sample sizes ($n > 200$) to be detected at least 50% of the time. Note that canonical correlations of small magnitude are often of little practical value, as they account for very little actual variability in the data. This issue, as well as the sample size issue, will be discussed shortly.

Canonical weights. After determining the number of significant canonical roots, the question arises as to how to interpret each (significant) root. Remember that each root actually represents two weighted sums, one for each set of variables. One way to interpret the "meaning" of each

canonical root would be to look at the weights for each set. These weights are called the *canonical weights*.

In general, the larger the weight (i.e., the absolute value of the weight), the greater is the respective variable's unique positive or negative contribution to the sum. To facilitate comparisons between weights, the canonical weights are usually reported for the standardized variables, that is, for the z transformed variables with a mean of 0 and a standard deviation of 1.

If you are familiar with [multiple regression](#), you may interpret the canonical weights in the same manner as you would interpret the beta weights in a multiple regression equation. In a sense, they represent the *partial correlations* of the variables with the respective canonical root. If you are familiar with [factor analysis](#), you may interpret the canonical weights in the same manner as you would interpret the *factor score coefficients*. To summarize, the canonical weights allow the user to understand the "make-up" of each canonical root, that is, it lets the user see how each variable in each set uniquely contributes to the respective weighted sum (canonical variate).

Canonical Scores. Canonical weights can also be used to compute actual values of the canonical variates; that is, we can simply use the weights to compute the respective sums. Again, remember that the canonical weights are customarily reported for the standardized (z transformed) variables.

Factor structure. Another way of interpreting the canonical roots is to look at the simple correlations between the canonical variates (or *factors*) and the variables in each set. These correlations are also called canonical factor *loadings*. The logic here is that variables that are highly correlated with a canonical variate have more in common with it. Therefore, we should weigh them more heavily when deriving a meaningful interpretation of the respective canonical variate. This method of interpreting canonical variates is identical to the manner in which factors are interpreted in [factor analysis](#).

Factor structure versus canonical weights. Sometimes, the canonical weights for a variable are nearly zero, but the respective loading for the variable is very high. The opposite pattern of results may also occur. At first, such a finding may seem contradictory; however, remember that the canonical weights pertain to the unique contribution of each variable, while the canonical factor loadings represent simple overall correlations. For example, suppose we included in your satisfaction survey two items that measured basically the same thing, namely: (1) "Are you satisfied with your supervisors?" and (2) "Are you satisfied with your bosses?" Obviously, these items are very redundant. When the program computes the weights for the weighted sums (canonical variates) in each set so that they correlate maximally, it only "needs" to include one of the items to capture the essence of what they measure. Once a large weight is assigned to the first item, the contribution of the second item is redundant; consequently, it will receive a zero or negligibly small canonical weight. Nevertheless, if we then look at the simple correlations between the respective sum score with the two items (i.e., the factor *loadings*), those may be substantial for *both*. To reiterate, the canonical weights pertain to the *unique contributions* of the respective variables with a particular weighted sum or canonical variate; the canonical factor loadings pertain to the *overall correlation* of the respective variables with the canonical variate.

Variance extracted. As discussed earlier, the canonical correlation coefficient refers to the correlation between the weighted sums of the two sets of variables. It tells nothing about how much variability (variance) each canonical root explains in the *variables*. However, we can infer the proportion of variance extracted from each set of variables by a particular root by looking at the canonical factor loadings. Remember that those loadings represent correlations between the canonical variates and the variables in the respective set. If we square those correlations, the resulting numbers reflect the *proportion* of variance accounted for in each variable. For each root, we can take the average of those proportions across variables to get an indication of how

much variability is explained, on the average, by the respective canonical variate in that set of variables. Put another way, we can compute in this manner the average proportion of *variance extracted* by each root.

Redundancy. The canonical correlations can be squared to compute the proportion of variance shared by the sum scores (canonical variates) in each set. If we multiply this proportion by the proportion of variance extracted, we arrive at a measure of *redundancy*, that is, of how redundant one set of variables is, given the other set of variables. In equation form, we can express the redundancy as:

$$\text{Redundancy}_{\text{left}} = [\sum(\text{loadings}_{\text{left}}^2)/p] * R_c^2$$

$$\text{Redundancy}_{\text{right}} = [\sum(\text{loadings}_{\text{right}}^2)/q] * R_c^2$$

In these equations, p denotes the number of variables in the first (left) set of variables, and q denotes the number of variables in the second (right) set of variables; R_c^2 is the respective squared canonical correlation.

Note that we can compute the redundancy of the first (*left*) set of variables given the second (*right*) set, and the redundancy of the second (*right*) set of variables, given the first (*left*) set. Because successively extracted canonical roots are uncorrelated, we could sum up the redundancies across all (or only the first significant) roots to arrive at a single index of redundancy (as proposed by Stewart and Love, 1968).

Practical significance. The measure of redundancy is also useful for assessing the *practical* significance of canonical roots. With large sample sizes (see below), canonical correlations of magnitude $R = .30$ may become statistically significant (see above). If we square this coefficient ($R\text{-square} = .09$) and use it in the redundancy formula shown above, it becomes clear that such canonical roots account for only very little variability in the variables. Of course, the final assessment of what does and does not constitute a finding of practical significance is subjective by nature. However, to maintain a realistic appraisal of how much actual variance (in the variables) is accounted for by a canonical root, it is important to always keep in mind the redundancy measure, that is, how much of the actual variability in one set of variables is explained by the other.

[To index](#)

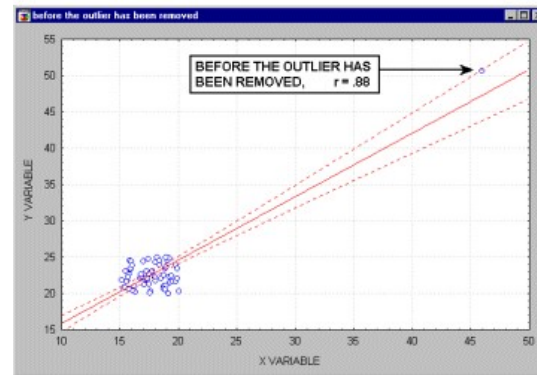
Assumptions

The following discussion provides only a list of the most important assumptions of canonical correlation analysis, and the major threats to the reliability and validity of results. Distributions. The tests of significance of the canonical correlations is based on the assumption that the distributions of the variables in the population (from which the sample was drawn) are multivariate normal. Little is known about the effects of violations of the multivariate normality assumption. However, with a sufficiently large sample size (see below) the results from canonical correlation analysis are usually quite robust.

Sample sizes. Stevens (1986) provides a very thorough discussion of the sample sizes that should be used in order to obtain reliable results. As mentioned earlier, if there are strong canonical correlations in the data (e.g., $R > .7$), then even relatively small samples (e.g., $n = 50$) will detect them most of the time. However, in order to arrive at reliable estimates of the canonical factor loadings (for interpretation), Stevens recommends that there should be at least 20 times as many cases as variables in the analysis, if one wants to interpret the most significant canonical root only. To arrive at reliable estimates for two canonical roots, Barcikowski and Stevens (1975) recommend, based on a Monte Carlo study, to include 40 to 60 times as many cases as variables.

Outliers. Outliers can greatly affect the magnitudes of correlation coefficients. Since canonical

correlation analysis is based on (computed from) correlation coefficients, they can also seriously affect the canonical correlations. Of course, the larger the sample size, the smaller is the impact of one or two outliers. However, it is a good idea to examine various scatterplots to detect possible outliers (as shown in the example animation below).



See also [Confidence Ellipse](#).

Matrix Ill-Conditioning. One assumption is that the variables in the two sets should not be completely redundant. For example, if we included the *same* variable twice in one of the sets, then it is not clear how to assign different weights to each of them. Computationally, such complete redundancies will "upset" the canonical correlation analysis. When there are perfect correlations in the correlation matrix, or if any of the multiple correlations between one variable and the others is perfect ($R = 1.0$), then the correlation matrix cannot be inverted, and the computations for the canonical analysis cannot be performed. Such correlation matrices are said to be *ill-conditioned*.

Once again, this assumption appears trivial on the surface; however, it often is "almost" violated when the analysis includes very many highly redundant measures, as is often the case when analyzing questionnaire responses.

[To index](#)

General Ideas

Suppose we conduct a study in which we measure satisfaction at work with three questionnaire items, and satisfaction in various other domains with an additional seven items. The general question that we may want to answer is how satisfaction at work relates to the satisfaction in those other domains.

Sum Scores

A first approach that we might take is simply to add up the responses to the work satisfaction items, and to correlate that sum with the responses to all other satisfaction items. If the correlation between the two sums is statistically significant, we could conclude that work satisfaction is related to satisfaction in other domains.

In a way this is a rather "crude" conclusion. We still know nothing about the particular domains of satisfaction that are related to work satisfaction. In fact, we could potentially have *lost* important information by simply adding up items. For example, suppose there were two items, one measuring satisfaction with one's relationship with the spouse, the other measuring satisfaction with one's financial situation. Adding the two together is, obviously, like adding "apples to oranges." Doing so implies that a person who is dissatisfied with her finances but happy with her spouse is comparable overall to a person who is satisfied financially but not happy in the relationship with her spouse. Most likely, people's psychological make-up is not that simple...

The problem then with simply correlating two sums is that one might lose important information in the process, and, in the worst case, actually "destroy" important relationships between variables by adding "apples to oranges."

Using a weighted sum. It seems reasonable to correlate some kind of a weighted sum instead, so that the "structure" of the variables in the two sets is reflected in the weights. For example, if satisfaction with one's spouse is only marginally related to work satisfaction, but financial satisfaction is strongly related to work satisfaction, then we could assign a smaller weight to the first item and a greater weight to the second item. We can express this general idea in the following equation:

$$a_1*y_1 + a_2*y_2 + \dots + a_p*y_p = b_1*x_1 + b_2*x_2 + \dots + b_q*x_q$$

If we have two sets of variables, the first one containing p variables and the second one containing q variables, then we would like to correlate the weighted sums on each side of the equation with each other.

Determining the weights. We have now formulated the general "model equation" for canonical correlation. The only problem that remains is how to determine the weights for the two sets of variables. It seems to make little sense to assign weights so that the two weighted sums do not correlate with each other. A reasonable approach to take is to impose the condition that the two weighted sums shall correlate maximally with each other.

[To index](#)

Canonical Roots/Variates

In the terminology of canonical correlation analysis, the weighted sums define a *canonical root* or *variate*. We can think of those canonical variates (weighted sums) as describing some underlying "latent" variables. For example, if for a set of diverse satisfaction items we were to obtain a weighted sum marked by large weights for all items having to do with work, we could conclude that the respective canonical variate measures satisfaction with work.

Number of Roots

So far we have pretended as if there is only one set of weights (weighted sum) that can be extracted from the two sets of variables. However, suppose that we had among our work satisfaction items particular questions regarding satisfaction with pay, and questions pertaining to satisfaction with one's social relationships with other employees. It is possible that the pay satisfaction items correlate with satisfaction with one's finances, and that the social relationship satisfaction items correlate with the reported satisfaction with one's spouse. If so, we should really derive two weighted sums to reflect this "complexity" in the structure of satisfaction.

In fact, the computations involved in canonical correlation analysis will lead to more than one set of weighted sums. To be precise, the number of roots extracted will be equal to the minimum number of variables in either set. For example, if we have three work satisfaction items and seven general satisfaction items, then three canonical roots will be extracted.

Extraction of Roots

As mentioned before, we can extract roots so that the resulting correlation between the canonical variates is maximal. When extracting more than one root, each successive root will explain a *unique* additional proportion of variability in the two sets of variables. Therefore, successively extracted canonical roots will be uncorrelated with each other, and account for less and less variability.

[To index](#)

Copyright © 2018 TIBCO Software Inc All Rights Reserved. | [Privacy Policy](#)