# Canonical Correlation & Principal Components Analysis

**Aaron French and Sally Chess**

---

## Canonical Correlation

Canonical Correlation is one of the most general of the multivariate techniques. It is used to investigate the overall correlation between two sets of variables (***p'*** and ***q'***). The basic principle behind canonical correlation is determining how much variance in one set of variables is accounted for by the other set along one or more axes. If there is more than one axis, they must be orthogonal. Unlike many other techniques, there is no designation that one set of variables is independent and the other set dependent. Canonical correlation is a descriptive or exploratory technique rather than a hypothesis-testing procedure, and there are several ways data may be combined with this procedure.

### Jargon

Variables: These are what you have measured in your research.

Canonical Variates: These are linear combinations of variables within a single set.

Canonical Variate Pairs: These are linear pairings of canonical variates, one from each of the two sets of variables.

### Practical Issues

While a normal distribution of the variables is not strictly required when canonical correlation is used descriptively, it does enhance the analysis. Homoscedasticity implies that the relationship between two variables is constant over the full range of data and this increases the accuracy of canonical correlation. Linearity is an important assumption of canonical correlation; this technique finds linear relationships between variables within a set and between canonical variate pairs between sets. Nonlinear components of these relationships are not recognized and so are not 'captured' in the analysis. Transformation may be useful to increase linearity. Outliers have a disproportionate impact on the results of the analysis and each set of variables is inspected independently for univariate and multivariate outliers. In general, the pattern of missing data is more important than the amount. Because canonical correlation is very sensitive to small changes in the data set, the decision to eliminate cases or estimate missing data must be considered carefully. Multicollinearity occurs when variables are very highly correlated. Singularity occurs when one variable is a linear combination of

two or more variables in that set. Both multicollinearity and singularity should be eliminated before analysis proceeds. The number of cases required for accurate analysis depends on the reliability of the variables: the more reliable the variables, the fewer cases are needed per variable. In searching the literature, the number of cases recommended ranged from ten to sixty (cases per variable).

There are several types of questions that can be answered with Canonical Correlation.

1. How many reliable variate pairs are there in the data set?

2. How strong is the correlation between variates in a pair?

3. How are dimensions that relate the variables to be interpreted?

The general equations for performing a canonical correlation are relatively simple. First, a correlation matrix (**R**) is formed. **R** is the product of the inverse of the correlation matrix of ***q'*** ($\mathbf{R_{yy}}$), a correlation matrix between ***q'*** and ***p'*** ($\mathbf{R_{yx}}$), the inverse of correlation matrix of ***p'*** ($\mathbf{R_{xx}}$), and the other correlation matrix between ***q'*** and ***p'*** ($\mathbf{R_{xy}}$).

$$\mathbf{R} = \mathbf{R}_{\mathbf{yy}}^{-1}\mathbf{R}_{\mathbf{yx}}\mathbf{R}_{\mathbf{xx}}\mathbf{R}_{\mathbf{xy}}$$

Canonical analysis proceeds by solving the above equation for eigenvalues and eigenvectors of the matrix **R**. Eigenvalues consolidate the variance of the matrix, redistributing the original variance into a few composite variates. Eigenvectors, transformed into coefficients, are used to combine the original variables into these composites.

The eigenvalues are related to the canonical correlation by the following equation:

$$\lambda_i = r_{ci}^2$$

That is, each eigenvalue equals the squared canonical correlation for each pair of canonical variates.

The significance of one or more canonical correlations is tested as a chi-square variable using the following formula:

$$\chi^2 = -\left[N - 1 - \left(\frac{k_x + k_y + 1}{2}\right)\right]\ln \Lambda_m$$

$$\Lambda_m = \prod_{i=1}^{m} (1 - \lambda)$$

with,

- **N**= number of cases
- $\mathbf{k_x}$= number of variables in ***p'***
- $\mathbf{k_y}$= number of variables in ***q'***
- **df= $(\mathbf{k_x})(\mathbf{k_y})$**
- **m**= number of canonical correlations

Significant results indicate that the overlap in variability between variables in the two sets is significant; this is evidence of significance in the first canonical correlation. This process (of finding a canonical correlation and testing for significance) is then repeated with the first pair of variates removed to see if any of the other pairs are significant. Only pairs that test significant are interpreted.

Canonical coefficients (also referred to as canonical weights) are analogous to the beta values in regression. One set of canonical coefficients is required for each set of variables for each canonical correlation. To facilitate comparisons, these values are usually reported for standardized variables (z transformed variables). The coefficients reflect differences in the contribution of the variables to the canonical correlation.

For ***q'*** the equation is:

$$\mathbf{B_y} = \left(\mathbf{R_{yy}^{-1/2}}\right)' \hat{\mathbf{B}}_\mathbf{y}$$

For ***p'***:

$$\mathbf{B_x} = \mathbf{R_{xx}^{-1/2} R_{xy} B_y^*}$$

$\hat{\mathbf{B}}_\mathbf{y}$=normalized matrix of eigenvectors

$\mathbf{B}_y^*$ =the matrix formed by the coefficients for ***q'***, each divided by their corresponding canonical correlation

**R**= matrix of correlations

The two matrices of canonical coefficients are used to estimate scores on canonical variates:

$\mathbf{X = Z_x B_x}$

$\mathbf{Y = Z_y B_y}$

Scores on canonical variates (**X, Y**) are the product of scores of original variates and the canonical coefficients used to weight them. The sum of canonical scores for each variate is equal to zero.

A loading matrix is a matrix of correlations between the canonical coefficients and the variables in each set. It is created by multiplying the matrix of correlations between variables with the matrix of canonical coefficients. Loading matrices (**A**) are used to interpret the canonical variates.

$\mathbf{A_x = R_{xx} B_x}$

$\mathbf{A_y = R_{yy} B_y}$

How much variance does each canonical variate explain? This is the sum of the squared correlations divided by the number of variables in the set.

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

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

**a**= loading correlations

**k**= number of variables in the set

How strongly does a variate relate to the variables on the other side of the equation? Redundancy is the amount of variance that the canonical variates from one set of variables extract from the other set. It is the product of the

percentage of variance extracted from one set of variables and the squared canonical correlation for the pair.

Canonical Correlation is subject to several limitations. It is mathematically elegant but difficult to interpret because solutions are not unique. Procedures that maximize correlation between canonical variate pairs don't necessarily lead to solutions that make logical sense. It is the canonical variates that are actually being interpreted and they are interpreted in pairs. A variate is interpreted by considering the pattern of variables that are highly correlated (loaded) with it. Variables in one set of the solution can be very sensitive to the identity of the variables in the other set; solutions are based upon correlation within and between sets, so a change in a variable in one set will likely alter the composition of the other set. There is no implication of causation in solutions. The pairings of canonical variates must be independent of all other pairs. Only linear relationships are appropriate.

## Principal Components Analysis

PCA is a type of factor analysis that is most often used as an exploratory tool. It is used to simplify the description of a set of many related variables; PCA reduces the number of variables by finding new components that are combinations of the old variables. It can also be useful as a preliminary step in a complicated regression analysis. In this case, first run a PCA which decreases the number of "important" variables, and then a regression can be performed on the principal components. PCA is used to determine the relationships among a group of variables in situations where it is not appropriate to make *a priori* grouping decisions, i.e. the data is not split into dependent and independent groups. Groups are created by forming composite axes that maximize the overall distance between the data. In other words, PCA determines the net effect of each variable on the total variance of the data set, and then extracts the maximum variance possible from the data.

### Practical issues

Many of the issues that are relevant to canonical correlation also apply to PCA. Normality in the distribution of variables is not strictly required when PCA is used descriptively, but it does enhance the analysis. Linearity between pairs of variables is assumed and inspection of scatterplots is an easy way to determine if this assumption is met. Again, transformation can be useful if inspection of scatterplots reveals a lack of linearity. Outliers and missing data will affect the outcome of analysis and these issues need to be addressed through estimation or deletion. Tabachnick and Fidell (1996) suggest that the minimum number of cases needed for PCA to be accurate is 300, but PCA is routinely used for much smaller data sets.

To perform a PCA the data are arranged in a correlation matrix (**R**) and then diagonalized. A diagonalized matrix (**L**) has numbers in the positive diagonal, 0's everywhere else. The correlation matrix (**R**) is diagonalized with the equation:

**L = V'RV**

The diagonalized matrix **(L)**, also known as the eigenvalue matrix, is created by premultiplying the matrix **R** by the transpose of the eigenvector matrix **(V')** and postmultiplying by the eigenvector matrix **(V)**. Premultiplying the matrix of eigenvectors by its transpose **(V'V)** produces a matrix with ones in the positive diagonal and zeros everywhere else, so the equation above just redistributes the variance of the matrix **R**. Calculations of eigenvectors and eigenvalues are the same as for canonical correlation. Eigenvalues are measures of the variance in the matrix. In an example where there are 10 components, on average each component will have an eigenvalue of 1 and will explain 10% of the variation in the data. A component with an eigenvalue of 2 explains twice the variance of an "average" variable, or 20% in the example.

Rearrange the last equation to give: **R=VLV'**

The square root of the eigenvalue matrix (**L**) is taken: $\mathbf{R = (VL^{1/2})(L^{1/2}V')}$

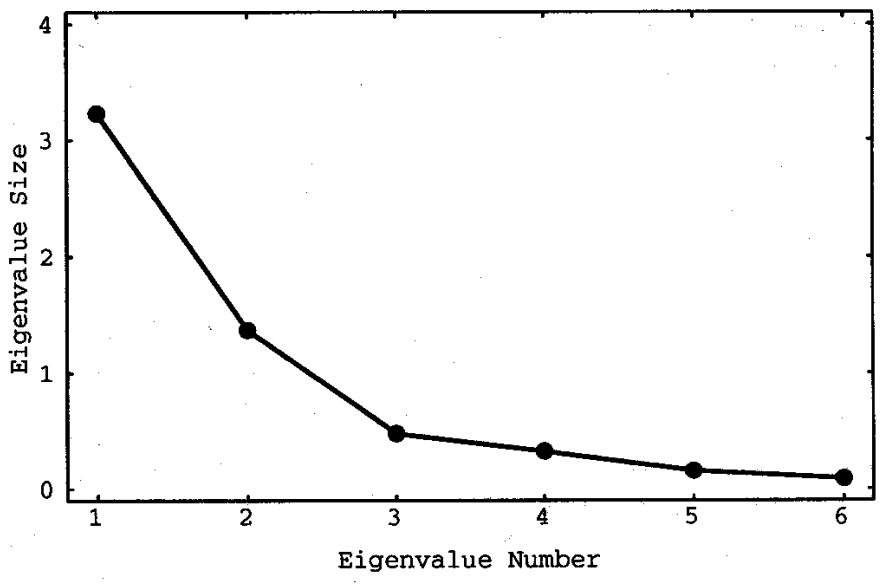
Let $\mathbf{VL^{1/2}=A}$ and $\mathbf{L^{1/2}V'=A'}$ to give **R=AA'**

The correlation matrix **(R)** is the product of the loading matrix (**A**) and its transpose **(A')**, each a combination of eigenvectors and square roots of eigenvalues. The loading matrix represents the correlation between each component and each variable, and is one of the commonly cited and referred to variables.

Score coefficients, which are similar to regression coefficients calculated in a multiple regression, are equal to the product of the inverse of the correlation matrix and the loading matrix: $\mathbf{B = R^{-1}A}$

After extraction of the principal components variables, which explain a small amount of variance in the data set may be discarded. There are no set rules as to how many principal components should be retained. One rule of thumb is to only accept components, which explain over 80% of the variance. Another more generous possibility is to accept all components, which explain a "non-random" percentage of the variance. A scree test can be useful in determining the total amount of variance explained by each component. A scree test is simply a graph with the principal components on the x-axis and eigenvalues on the y-axis. A decision about which principal components to included is made by drawing a line from the first component, and looking for an inflection point between the steep curve and the straight line at the bottom. Principal components to the left of the

inflection point would be retained. In the example below, two components would be retained.




(A. C. Rencher 1995)

The only remaining question now is “What do the various principal components mean?” Interpretation of principal components sometimes can be a bit ambiguous with a complex data set. However, components are usually interpreted by examining the correlations between the initial raw variables and each component. It may be possible to interpret each principal component as a combination of a small number of the original variables, with which they are most highly correlated. The principal components can also remain unnamed and simply be referred to by their component names (PC1, PC2, etc.).

There are several situations that are not appropriate for a PCA. PCA cannot be used when results are pooled across several samples, not for a repeated measures design. This is because the underlying structure of the data set may shift across samples or across time, and the principal components analysis does not allow for this. Additionally, PCA is very sensitive to the sizes of correlations between variables, and any non-linearity between pairs of variables.

There are several potential problems with performing a PCA. Most importantly, there are no criteria against which to test results. Therefore, there is no way of testing the significance of the groupings. There is no guarantee that the results will yield any biologically significant information.

**Reference Literature**

Tabachnick, B. G. and L. S. Fidell. 1996. Using Multivariate Statistics. 3rdEdition. HarperCollins College Publishers. A good reference book on multivariate methods.

Afifi, A. A., and V. Clark. 1996. Computer-aided Multivariate Analysis. 3rdEdition. Chapman and Hall.

Cooley, W.W. and D.R. Lohnes. 1971. Multivariate data Analysis. John Wiley & Sons, Inc.

Rencher, A. C. 1995. Methods of Multivariate Analysis. John Wiley & Sons, Inc.

Calhoon, R. E., and D. L. Jameson. 1970. Canonical correlation between variation in weather and variation in size in the Pacific Tree Frog, *Hyla regilla*, in Southern California. Copea 1: 124-144. A very good ecological paper describing the uses of Canonical Correlation.

Alisauskas, R.T.1998. Winter range expansion and relationships between landscape and morphometrics of midcontinent lesser snow geese. Auk 115: 851-862. A good ecological application of PCA.

**Links to sites with more information on PCA**

http://obelia.jde.aca.mmu.ac.uk/multivar/pca.htm

This is a great site for PCA (and other types of analysis not covered on this website) with background information, data sets, graphical explanation etc.

http://www.statsoftinc.com/textbook/stcanan.html

http://www2.chass.ncsu.edu/garson/pa765/canonic.htm

http://www.ivorix.com/en/products/tech/pca/pca.html

This site allows you to rotate a 3-D model of a PCA.

http://www.statsoftinc.com/textbook/stfacan.html

http://www.okstate.edu/artsci/botany/ordinate/PCA.htm

*This page was last updated on*