Canonical Correlation Analysis (CCA)

**Introduction:** Canonical correlation is an extension of multiple correlation. Multiple correlation is the correlation between one response variable and several predictor variables . It is defined as

, where, , .

Canonical correlation analysis is a method for exploring the amount of (linear) relationship between two multivariate sets of variables (vectors), all measured on the same subject in the study.

Consider as an example of variables related to exercise and health. On one hand you have variables associated with exercise, observations such as the climbing rate on a stair stepper, how fast you can run, the amount of weight lifted on bench press, the number of push-ups per minute, number of calories burnt on a cardio machine for 15 minutes, etc. You may also have health variables such as blood pressure, cholesterol levels, glucose levels, body mass index, etc. So two types of variables are measured and the relationship between the exercise variables and the health variables is the point of interest.

As a second example consider variables measured on environmental health and environmental toxins. A number of environmental health variables such as frequencies of sensitive species, species diversity, total biomass, productivity of the environment, etc. may be measured on one hand; on the other a second set of variables such as environmental toxins which might include the concentrations of heavy metals, pesticides, dioxin, etc. are measured.

For a third example consider a group of sales representatives, on whom we have recorded several sales performance variables (Sales Growth, Sales Profitability, New Account Sales, etc.) along with several measures of intellectual and creative aptitude (Creativity, Mechanical Reasoning, Abstract Reasoning, Mathematics, etc.). We may wish to explore the relationships between the sales performance variables and the aptitude variables.

Other examples on two types of variables include – a set of price indices and a set of production indices, a set of student behaviors and a set of teacher behaviors, a set of ecological variables and a set of environmental variables, a set of academic achievement variables and a set of measures of job success, a set of closed-book exam scores and a set of open-book exam scores, a set of personality variables of freshmen students and the same variables on the same students as seniors.

One approach to studying relationship between the two sets of variables is to use canonical correlation analysis which describes the (linear) relationship between the first set of variables and the second set of variables.

Canonical correlation analysis focuses on the correlation between a linear combination of the variables in one set and a linear combination of variables in the another set. The idea is to first determine the pair of linear combinations having the largest correlation. Next, we determine the pair of linear combinations having the largest correlation among all pairs uncorrelated with the initially selected pair and so on. The pairs of linear combinations are called the **canonical variables** or **canonical variates** and their correlations are called **canonical correlations**.

**Motivations for Canonical Correlation Analysis**:It is possible to create pairwise *scatterplots* with variables in the first set (e.g., exercise variables), and variables in the second set (e.g., health variables). But if dimension of the first set is and that of the second set is , there will be  such scatter plots. It may be difficult to look at all of these graphs together and be able to interpret the results.

Similarly, you could compute all *correlations* between variables from the first set (e.g., exercise variables), and the variables in the second set (e.g., health variables). But with  a large number, problem of interpretation arises.

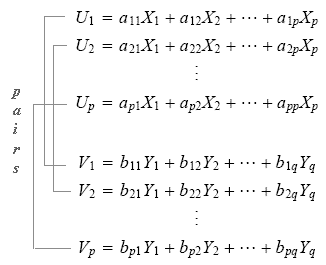
The main task of Canonical Correlation Analysis is to summarize the associations between the two sets of variables in terms of a *few* carefully chosen correlations rather than the correlations. In a way the motivation for canonical correlation is very similar to principal component analysis. It is another dimension reduction technique.

**Obtaining Canonical Variates and their Correlations**

We have two sets of variables and . Suppose we have variables in vector and variables in vector :

We select and based on the number of variables that exist in each set so that .  This is done for theoretical development. When writing program in R, we do not need to select and based on the number of variables. R will automatically handle this and find number of canonical correlations and canonical variates where

Now we define a set of linear combinations named and . Here will correspond to the linear combinations from the first set of variables, , and will correspond to the second set of variables, . Each member of will be paired with a member of . For example, below is a linear combination of the variables and is the corresponding linear combination of the variables. Similarly, *U*2 is a linear combination of the variables, and *V*2 is the corresponding linear combination of the variables. And, so on....



Thus define (*U*1, *V*1) as the first canonical variate pair, similarly (*U*2, *V*2) would be the second canonical variate pair. ( will be the *i*th *canonical variate pair*, and so on. With there are canonical covariate pairs.

We are to find linear combinations that **maximize the correlations** between the members of each canonical variate pair. In the analysis, our interest is in determining the number of dimensions (canonical variables) that are significant in explaining the association between the 2 sets of variables.

Let us define the following notations:

where is a matrix, is a matrix, is a matrix, and is a matrix. Note that

We can compute the variance of using the following expression:

The coefficients through that appear in the double sum are the same coefficients that appear in the definition of .

Similar calculations can be made for the variance of as shown below:

Now let us calculate the covariance between and as:

The correlation between and is calculated using the usual formula. We take the covariance between those two variables and divide it by the square root of the product of the individual variances:

The *canonical correlation* is a specific type of correlation. The canonical correlation for the *i*th canonical variate pair is simply the correlation between and :

This quantity is to be maximized. We want to find linear combinations of the *X*'s and linear combinations of the *Y*'s that maximize the above correlation.

*First canonical variate pair* (, ):

The coefficients   and  are to be selected so as to maximize the canonical correlation of the first canonical variate pair. This is subject to the constraint that variances of the two canonical variates in that pair are equal to 1.

This condition is required so that unique values for the coefficients are obtained.

*Second canonical variate pair* (, ):

Similarly, we want to find the coefficients   and  that maximize the canonical correlation of the second canonical variate pair, (, ). Again, we will maximize this canonical correlation subject to the constraints that the variances of the individual canonical variates are both equal to 1. Furthermore, we require the additional constraints that (, ), and (, ) have to be uncorrelated. In addition, the combinations (, ), and (, ) must be uncorrelated. In summary, our constraints are:

, (,

,

Basically we require that all of the remaining correlations equal zero.

The  *canonical variate pair* (, ):

We want to find the coefficients   and  that maximizes the canonical correlation subject to the similar constraints that

, (, ,

, (, ,

…..

, (,

, (,

, (,

…

, (, .

The above relations imply that is uncorrelated with , …, ;

is uncorrelated with , …, ; is uncorrelated with , …, ;

is uncorrelated with , …,

**How to find the coefficients of vectors and ?**

Following Result 10.1on page 541 of Johnson’s textbook, the square of the canonical correlations are the eigenvalues of the matrix . Furthermore, the components of the vector andare the elements of the eigenvectors associated with the eigenvalues of matrices and , respectively.

**Properties of Canonical Correlations**

Three interesting properties of canonical correlations are the following:

1. Canonical correlations are always nonnegative.
2. Canonical correlations are invariant to changes of scale on either the ’s or the ’s. For example, if the measurement scale is changed from inches to centimeters, the canonical correlations will not change (the corresponding eigenvectors will change). This property holds for simple and multiple correlations as well.
3. The first canonical correlation is the maximum correlation between linear functions of and .Therefore, exceeds (the absolute value of) the simple correlation between any Y and any X or the multiple correlation between any Y and all the X’s or between any X and all the Y’s.

for

for any

for any ,

**Tests of Significance:** Here we discuss the basic tests of significance associated with the canonical correlations. The very first thing to determine is if there is any relationship between the two sets of variables at all. Perhaps the two sets of variables are completely unrelated to one another and independent!

Here the null hypothesis is : the covariance matrix between and is a zero matrix: or , that is, the covariance of every with every is zero, and all corresponding correlations are zero. Hence, under : there is no linear association between ’s and ’s. For notational convenience, let us denote the null hypothesis by . The null hypothesis is equivalent to the statement that all canonical correlations are insignificant, that is,

This test is carried out using Wilks’ lambda, Pillai, Hotelling-Lawley and Roy’s statistics. The choice of test depends on the specific requirements of your analysis and the assumptions you are willing to make.

Below we present the test using Wilks’ lambda statistic. Let , and denote the sample covariance matrices of random vectors and , respectively. Then, Wilks’ lambda is defined as

, or equivalently,

If one or more is large, will be small. Thus, we reject for small values of

Alternately, we use F-approximation given by

where

We reject if . If Wilks’ lambda (or the F statistic) is significant (therefore, we reject ) and since the canonical correlations are ordered from largest to smallest, we can conclude that at least

.

If we reject the hypothesis and conclude that , then we may also wish to test that canonical correlations beyond the first one are significant. Therefore, in case of rejection, we move on and test the null hypothesis

To test significance of , we calculate

If this test rejects hypothesis, we conclude that at least is signiﬁcantly different from zero. We can continue in this manner, testing each in turn, until a test fails to reject the hypothesis (the test is insignificant).

In order to find the F approximation, at any step, the quantities and will be calculated by replacing and by , and , respectively, to obtain

where

,

In practice, these tests would be carried out successively until we find an insignificant result. Once an insignificant result is found we would stop. If this happens with the first canonical variate pair it suggests that there is no evidence of any linear relationship between the two sets of variables and the analysis may be stopped. If the first pair shows significance, then we move on to the second canonical variate pair. If this second pair is not significantly correlated then we would stop. If it was significant we would continue to the third pair, proceeding in this iterative manner through the pairs of canonical variates testing until we find insignificant results.

Notes:

* For testing the significance of canonical correlations, we need the assumption that the joint distribution of and is multivariate normal, that is,
* Why/when use Wilks' Lambda (Λ): Wilks' lambda is a commonly used test statistic for CCA because it is based on multivariate analysis of variance (MANOVA) and is relatively easy to compute. It tests the hypothesis that all canonical correlations are equal to zero.
* Why/when use Hotelling's T-squared statistic: Hotelling's T-squared statistic is used in multivariate hypothesis testing and is especially useful when the sample size is small. It is sensitive to deviations from multivariate normality.
* Why/when use Roy's Largest Root: Roy's Largest Root is a likelihood ratio test used in CCA. It is based on the maximum likelihood estimation of canonical correlations and is particularly useful for large sample sizes.

When choosing a significance test for CCA, it's essential to consider the assumptions of each test, the size of your dataset, and the nature of your variables. Each test has its own strengths and limitations, and researchers often use multiple tests to cross-validate their findings. It's important to consider the assumptions of each test and the specific research question when selecting an appropriate significance test for canonical correlation analysis. Additionally, it is good practice to report the results of multiple tests if feasible, as this can provide a more robust assessment of the significance of the canonical correlations in your analysis.

**Precautions when using Canonical Correlations**

* The sample size involved should be of a relatively large size. The technique works best with larger sample sizes.
* Even with large sample sizes, small deviations from zero will show up statistically significant for some canonical correlation values. From a practical point of view, the small canonical correlation can probably be ignored, since (1) they are small in magnitude and (2) the corresponding canonical variates explain very little of the sample variation in the variable sets and .

R programs: CCA\_ChemicalReaction