Published on *STAT 505* (https://onlinecourses.science.psu.edu/stat505)

Home > 13.1 - Setting the Stage for Canonical Correlation Analysis

13.1 - Setting the Stage for Canonical Correlation Analysis

What motivates canonical correlation analysis?

It is possible to create pairwise scatter plots 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 *p* and that of the second set is *q*, there will be *pq* such scatter plots, it may be difficult, if not outright impossible, 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 then compute all the correlations between the variables in the second set (e.g., health variables). But with *pq* a large number, problem of interpretation arises.

Canonical Correlation Analysis allows us to summarize the relationships into lesser number of statistics while preserving the main facets of the relationships. In a way the motivation for canonical correlation is very similar to principal component analysis. It is another dimension reduction technique.

Canonical Variates

Let's begin with the notation:

We have two sets of variables *X* and *Y*.

Suppose we have *p* variables in set 1: $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$

and suppose we have *q* variables in set 2: $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{pmatrix}$

We select X and Y based on the number of variables that exist in each set so that $p \le q$. This is done for computational convenience.

Just as done in principal components analysis we look at linear combinations of the data. We define a set of linear combinations named *U* and *V*. *U* will correspond to the linear combinations from the first set of variables, *X*, and *V* will correspond to the second set of variables, *Y*. Each member of *U* will be paired with a member of *V*. For example, U_1 below is a linear combination of the *p X* variables and V_1 is the corresponding linear combination of the *q Y* variables.

Similarly, U_2 is a linear combination of the *p X* variables, and V_2 is the corresponding linear combination of the *q Y* variables. And, so on....

$$\begin{array}{lcl} U_1 & = & a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ U_2 & = & a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ & \vdots & \\ U_p & = & a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \\ \\ V_1 & = & b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1q}Y_q \\ V_2 & = & b_{21}Y_1 + b_{22}Y_2 + \cdots + b_{2q}Y_q \\ & \vdots & \\ V_p & = & b_{p1}Y_1 + b_{p2}Y_2 + \cdots + b_{pq}Y_q \end{array}$$

Thus define

$$(U_i, V_i)$$

as the i^{th} *canonical variate pair*. (U_1, V_1) is the first canonical variate pair, similarly (U_2, V_2) would be the second canonical variate pair and so on... With $p \le q$ there are p canonical covariate pair.

We are to find linear combinations that maximize the correlations between the members of each canonical variate pair.

We can compute the variance of U_i variables using the following expression:

$$\text{var}(U_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} a_{ik} a_{il} cov(X_k, X_l)$$

The coeffcients a_{i1} through a_{ip} that appear in the double sum are the same coefficients that appear in the definition of U_i . The covariances between the *k*th and *l*th *X*-variables are multiplied by the corresponding coefficients a_{ik} and a_{il} for the variate $\mathrm{U_i}$.

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

$$\text{var}(V_j) = \sum_{k=1}^{p} \sum_{l=1}^{q} b_{jk} b_{jl} \text{cov}(Y_k, Y_l)$$

Then calculate the covariance between U_i and V_j as:

$$\text{cov}(U_i, V_j) = \sum_{k=1}^{p} \sum_{l=1}^{q} a_{ik} b_{jl} \text{cov}(X_k, Y_l)$$

The correlation between U_i and V_j 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 variances:

$$\frac{\text{cov}(U_i, V_j)}{\sqrt{\text{var}(U_i)\text{var}(V_j)}}$$

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

$$\rho_i^* = \frac{\text{cov}(U_i, V_i)}{\sqrt{\text{var}(U_i)\text{var}(V_i)}}$$

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.

Canonical Variates Defined

Let us look at each of the *p* canonical variates pair one by one.

First canonical variate pair: (U_1, V_1):

The coefficients $a_{11}, a_{12}, \dots, a_{1p}$ and $b_{11}, b_{12}, \dots, b_{1q}$ are to be selected so as to maximize the canonical correlation ρ_1^* 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 one.

$$\text{var}(U_1) = \text{var}(V_1) = 1$$

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

Second canonical variate pair: (U_2, V_2)

Similarly we want to find the coefficients $a_{21}, a_{22}, \dots, a_{2p}$ and $b_{21}, b_{22}, \dots, b_{2q}$ that maximize the canonical correlation ρ_2^* of the second canonical variate pair, (U_2, V_2). Again, we will maximize this canonical correlation subject to the constraints that the variances of the individual canonical variates are both equal to one. Furthermore, we require the additional constraints that (U_1, U_2), and (V_1, V_2) have to be uncorrelated. In addition, the combinations (U_1, V_2) and (U_2, V_1) must be uncorrelated. In summary, our constraints are:

$$\text{var}(U_2) = \text{var}(V_2) = 1,$$

$$\text{cov}(U_1, U_2) = \text{cov}(V_1, V_2) = 0,$$

$$\text{cov}(U_1, V_2) = \text{cov}(U_2, V_1) = 0.$$

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

This procedure is repeated for each pair of canonical variates. In general, ...

i^{th} canonical variate pair: (U_i, V_i)

We want to find the coefficients $a_{i1}, a_{i2}, \dots, a_{ip}$ and $b_{i1}, b_{i2}, \dots, b_{iq}$ that maximizes the canonical correlation ρ_i^* subject to the similar constraints that

$$\text{var}(U_i) = \text{var}(V_i) = 1,$$

$$\text{cov}(U_1, U_i) = \text{cov}(V_1, V_i) = 0,$$

$$\text{cov}(U_2, U_i) = \text{cov}(V_2, V_i) = 0,$$

$$\vdots$$

$$\text{cov}(U_{i-1}, U_i) = \text{cov}(V_{i-1}, V_i) = 0,$$

$$\text{cov}(U_1, V_i) = \text{cov}(U_i, V_1) = 0,$$

$$\text{cov}(U_2, V_i) = \text{cov}(U_i, V_2) = 0,$$

$$\vdots$$

$$\text{cov}(U_{i-1}, V_i) = \text{cov}(U_i, V_{i-1}) = 0.$$

Again, requiring all of the remaining correlations to be equal zero.

Next, let's see how this is carried out in SAS...

Source URL: https://onlinecourses.science.psu.edu/stat505/node/65