

Canonical Correlation Analysis

In principal components analysis, we analyzed one set of variables and found linear combinations within that set that have maximum variance.

In canonical correlation, we analyze dimensions that sets of variables have in common.

Specifically, suppose you have data on two sets of variables, \mathbf{x} and \mathbf{y} , and you wish to find linear combinations $\mathbf{a}'\mathbf{x}$ and $\mathbf{b}'\mathbf{y}$ that are maximally correlated. These are *canonical variates*.

Another view of Canonical Variates

An alternate view of the first canonical variate in one set of variables is that it is the linear combination of those variables that has the highest multiple correlation with the variables in the other set.

What Are They Used For

Example.

The relationship between personality and achievement is of interest. Suppose the \mathbf{x} variables are a set of personality scale scores, and the \mathbf{y} variables are a set of academic achievement scores. Then the first canonical variate may isolate dimensions of personality and achievement that predict each other well.

Finding the Canonical Variates

Finding the canonical variates amounts to finding the linear weights \mathbf{a} and \mathbf{b} that generate them.

These linear weights can of course only be identified by shape. Multiplying \mathbf{a} or \mathbf{b} by a constant will not change the correlation between the canonical variates $u = \mathbf{a}'\mathbf{x}$ and $v = \mathbf{b}'\mathbf{y}$.

So we implement the unit variance restrictions

$$\mathbf{a}'_i \mathbf{S}_{xx} \mathbf{a}_i = \mathbf{b}'_i \mathbf{S}_{yy} \mathbf{b}_i = 1 \quad (1)$$

A straightforward application of matrix calculus leads to the result that the \mathbf{a}_i corresponding to the i th canonical variate is proportional to the eigenvectors of the quadruple product

$$\mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \mathbf{S}_{yy}^{-1} \mathbf{S}_{yx} \quad (2)$$

and that the squared *canonical correlation* r_i^2 is the corresponding eigenvalue. The canonical weights

for the \mathbf{y} variates are obtained from the eigenvectors of

$$\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy} \quad (3)$$

It is important to realize that textbooks, in general, are *very* confused (or at least very confusing) in their treatments of canonical correlation.

In particular, there are different meanings of the same term, depending on which book you read.

Raw Canonical Coefficients (Weights)

These are the linear weights used to produce the canonical variates from the raw scores. Note, however, our previous restriction that the canonical variates have unit variance as per Equation (1). Since the eigenvectors as output from most computer software are normalized to have “unit length”, i.e.,

$$\mathbf{a}'_i \mathbf{a}_i = \mathbf{1} \quad (4)$$

they will not generally satisfy Equation (1). What to do? In essence, we find out the variance of the variable created from the normalized eigenvectors, then restandardize each vector to produce a variance of 1. In a while, we will see how this is done using a symmetric power of a matrix.

This creates a paradox. Suppose that a variable in the \mathbf{y} set is perfectly reproduced from a linear combination of the variables in the \mathbf{x} set, but that the variable in the \mathbf{y} set does not have unit variance. Then the “raw canonical weights” *after* they are corrected will not produce a variable equal

to the variable in the \mathbf{y} set. It *will* be perfectly correlated with it. On the other hand, the canonical weights *before* being corrected *will* produce scores identical to the variable in the \mathbf{y} set.

So there are, in fact, 3 versions of canonical coefficients we may talk about.

1. *Completely raw*. Based on the eigenvectors in Equations (2) and (3).
2. *Partially standardized*. Rescaled so that the canonical variates computed from raw scores in \mathbf{x} and \mathbf{y} have unit variance.

3. *Completely standardized.* Based on standardized \mathbf{x} and \mathbf{y} (i.e., calculated from correlation matrices rather than covariance matrices), then rescaled so that the canonical variates computed from the standardized scores have unit variance.

Let's use score notation. The “partially standardized” canonical variates for the \mathbf{x} set are produced as

$$\mathbf{U} = \mathbf{XA}(\mathbf{A}'\mathbf{S}_{xx}\mathbf{A})^{-1/2} = \mathbf{XA}^* \quad (5)$$

and those for the \mathbf{y} set are

$$\mathbf{V} = \mathbf{YB}(\mathbf{B}'\mathbf{S}_{yy}\mathbf{B})^{-1/2} = \mathbf{YB}^* \quad (6)$$

where \mathbf{A} has in its columns the eigenvectors of the matrix in Equation (2), and \mathbf{B} has the eigenvectors in of the matrix in Equation (3). \mathbf{A}^* and \mathbf{B}^* are actually the “raw canonical weights” referred to by the SAS program.

If correlation matrices rather than covariance matrices are used in Equations (2) and (3), then the resulting \mathbf{A}^* and \mathbf{B}^* are the completely standardized weights referred to as “standardized canonical weights” by the SAS program.