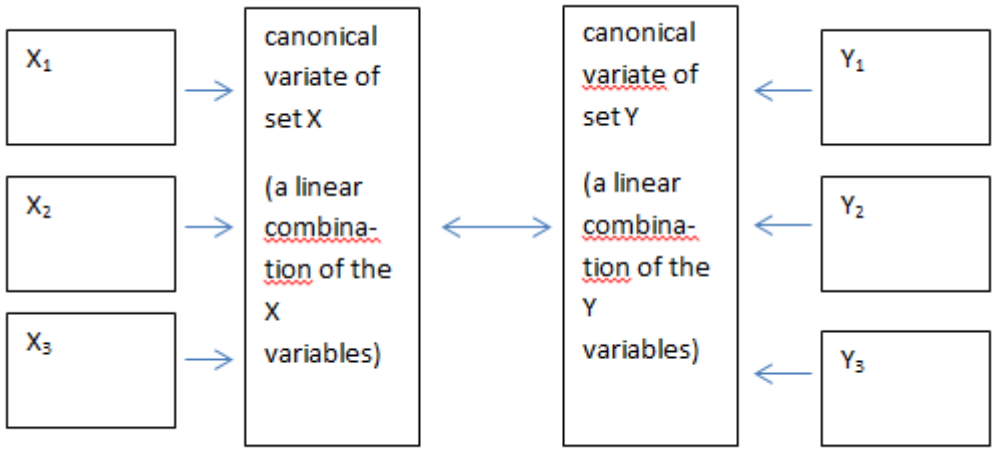
Canonical correlation

1. Introduction

We have two sets of variables, X and Y. Let X be an nxk and Y an nxm matrix. That is, we have n observations, k variables in set X and m in set Y. We would like to learn about the statistical relationship between the two sets of variables. If one of the sets had only a single variable, we could use a regression analysis. But this is not the case now, so we need to think of something else. The idea is that we create a linear combination for the two sets of variables each so that they have the highest possible correlation. The idea can be summarized in the block diagram below:



That is, we are going to create two canonical variates or canonical correlation variables (both are valid expressions):

$V_X = \sum_{j=1}^{k} a_j X_j$ (1.1)and (1.2) $V_y = \sum_{h=1}^{m} b_h Y_h$ that is, using matrix algebra:

$\mathbf{V_x} = \mathbf{Xa}$ (1.3)and $\mathbf{V}_Y = \mathbf{Yb}$ (1.4)

where **a** and **b** are the vectors of the coefficients that would maximize the correlation between the two canonical variates. The method was developed by Hotelling in 1935-36. Despite its age, it is not very popular, even though it may prove very useful in social sciences, whenever we have a reason to think that two sets of variables are linked through a single latent factor.

In case of single variables expressing the linear correlation coefficient would not be difficult:

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1.5)$$

In case of multiple variables, we can make use of the cross products:

$$\rho_{V_x V_y} = \frac{\mathbf{a}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XY}}\mathbf{b}}{\sqrt{\mathbf{a}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{a}\mathbf{b}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{b}}} \quad (1.6) \quad \text{where} \quad \boldsymbol{\Sigma}_{\mathbf{XY}} = \frac{1}{n}\mathbf{X}^{\mathbf{T}}\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}}\boldsymbol{\mu}_{\mathbf{X}}^{\mathbf{T}}, \boldsymbol{\Sigma}_{\mathbf{XX}} = \frac{1}{n}\mathbf{X}^{\mathbf{T}}\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{X}}^{\mathbf{T}},$$

$$\boldsymbol{\Sigma}_{\mathbf{YY}} = \frac{1}{n}\mathbf{Y}^{\mathbf{T}}\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}}\boldsymbol{\mu}_{\mathbf{Y}}^{\mathbf{T}}$$

Such cross products are of crucial importance in statistics, hence you should be aware what they mean.

If we have the following two matrices:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix} \text{ and } \mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ y_{31} & y_{32} \end{bmatrix}$$

then

$$\mathbf{X}^{\mathbf{T}}\mathbf{X} = \begin{bmatrix} \sum x_{\cdot 1}^2 & \sum x_{\cdot 1}x_{\cdot 2} & \sum x_{\cdot 1}x_{\cdot 3} \\ \sum x_{\cdot 1}x_{\cdot 2} & \sum x_{\cdot 2}^2 & \sum x_{\cdot 2}x_{\cdot 3} \\ \sum x_{\cdot 1}x_{\cdot 3} & \sum x_{\cdot 2}x_{\cdot 3} & \sum x_{\cdot 3}^2 \end{bmatrix} \text{then if we take the expectations:}$$

$$\boldsymbol{\Sigma}_{\mathbf{XX}} = \frac{1}{n}\mathbf{X}^{\mathbf{T}}\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{X}}^{\mathbf{T}} = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1x_2} & \sigma_{x_1x_3} \\ \sigma_{x_1x_2} & \sigma_{x_2}^2 & \sigma_{x_2x_3} \\ \sigma_{x_1x_3} & \sigma_{x_2x_3} & \sigma_{x_3}^2 \end{bmatrix}$$ which is called the variance-covariance or just simply covariance matrix. It is symmetric and positive semidefinite (all elements are larger than or equal to zero). From symmetry it follows that $\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{\mathbf{T}} = \mathbf{X}^{\mathbf{T}}\mathbf{X}$.

$$\mathbf{X}^{\mathbf{T}}\mathbf{Y} = \begin{bmatrix} \sum x_{\cdot 1}y_{\cdot 1} & \sum x_{\cdot 1}y_{\cdot 2} \\ \sum x_{\cdot 2}y_{\cdot 1} & \sum x_{\cdot 2}y_{\cdot 2} \\ \sum x_{\cdot 3}y_{\cdot 1} & \sum x_{\cdot 3}y_{\cdot 2} \end{bmatrix} \boldsymbol{\Sigma}_{\mathbf{XY}} = \frac{1}{n}\mathbf{X}^{\mathbf{T}}\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}}\boldsymbol{\mu}_{\mathbf{X}}^{\mathbf{T}} = \begin{bmatrix} \sigma_{x_1y_1} & \sigma_{x_1y_2} \\ \sigma_{x_2y_1} & \sigma_{x_2y_2} \\ \sigma_{x_3y_1} & \sigma_{x_3y_2} \end{bmatrix}$$ which is a covariance matrix among the two sets of variables.

Our optimization problem is the following:

$\max_{\mathbf{a},\mathbf{b}} \rho_{V_x V_y} = \frac{\mathbf{a}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XY}}\mathbf{b}}{\sqrt{\mathbf{a}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{a}\mathbf{b}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{b}}}$ (1.7), which reads as follows: we look for those vectors **a** and **b** that maximizes the correlation between the canonical variates.

2. Derivation

Fortunately, rescaling any of the variables will not affect the linear correlation and so we can find the solution much easier. For example, we can introduce two vectors such as: $\mathbf{c} = \boldsymbol{\Sigma}_{\mathbf{XX}}^{\frac{1}{2}}\mathbf{a}$ $\mathbf{d} = \boldsymbol{\Sigma}_{\mathbf{YY}}^{\frac{1}{2}}\mathbf{b}$. (1.7) becomes then:

$$\max_{\mathbf{c},\mathbf{d}} \rho_{V_x V_y} = \frac{\mathbf{c}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}\mathbf{d}}{\sqrt{\mathbf{c}^{\mathbf{T}}\mathbf{c}}\sqrt{\mathbf{d}^{\mathbf{T}}\mathbf{d}}} \text{ (2.1)}$$

If we assume that $\mathbf{c}^{\mathbf{T}}\mathbf{c} = \mathbf{a}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{a} = 1$(2.1)and $\mathbf{d}^{\mathbf{T}}\mathbf{d} = \mathbf{b}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{YY}}\mathbf{b} = 1$(2.2)then the problem simplifies into a conditional or constrained optimization problem:

$\max_{\mathbf{c},\mathbf{d}} \mathbf{c}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}\mathbf{d}$ subject to $\mathbf{c}^{\mathbf{T}}\mathbf{c} = \mathbf{d}^{\mathbf{T}}\mathbf{d} = 1$. (2.3)

So we have the following Lagrangian:

$$L = \mathbf{c}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}\mathbf{d} - \lambda_1(\mathbf{c}^{\mathbf{T}}\mathbf{c} - 1) - \lambda_2(\mathbf{d}^{\mathbf{T}}\mathbf{d} - 1) \text{ (2.4)}$$

The First Order Condition (FOC) requires that the first derivatives with respect to vectors a and b should be equal to zero:

$$\frac{\partial L}{\partial \mathbf{c}} = \boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}\mathbf{d} - 2\lambda_1\mathbf{c} = 0 \text{ (FOC 1) (2.5)}$$

$$\frac{\partial L}{\partial \mathbf{d}} = \boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\mathbf{c} - 2\lambda_2\mathbf{d} = 0 \text{ (FOC 2) (2.6)}$$

We need to find out the value of the two Lagrange multipliers. This is done by multiplying (2.5) by the transpose of vector **c** and (2.6) by the transpose of vector **d** we obtain:

$\mathbf{c}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}\mathbf{d} = 2\lambda_1\mathbf{c}^{\mathbf{T}}\mathbf{c}$ (2.7) and $\mathbf{d}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\mathbf{c} = 2\lambda_2\mathbf{d}^{\mathbf{T}}\mathbf{d}$ (2.8)

Since $\mathbf{c}^{\mathbf{T}}\mathbf{c} = \mathbf{d}^{\mathbf{T}}\mathbf{d} = 1$, it is straightforward that $\lambda_1 = \lambda_2 = \lambda = \frac{1}{2}\mathbf{d}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\mathbf{c} = \frac{\sigma_{V_X V_Y}}{2} = \frac{\rho_{V_X V_Y}}{2}$ (2.9). Note that what we obtain is that lambda equals half of the canonical covariance, which equals the canonical correlation if we assume that the terms in the denominator are unit.

We can use the two FOCs to arrive at the expressions for vectors **c** and **d**:

$$\mathbf{c} = \frac{\boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}\mathbf{d}}{\rho_{V_X V_Y}} \text{ (2.10)and } \mathbf{d} = \frac{\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\mathbf{c}}{\rho_{V_X V_Y}} \text{ (2.11)}$$

We can substitute these to the (2.6) and (2.5) respectively to obtain:

$$\left[\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}} - \rho_{V_X V_Y}^2\mathbf{I}_m\right]\mathbf{d} = 0 \text{ (2.12)} \left[\boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}} - \rho_{V_X V_Y}^2\mathbf{I}_k\right]\mathbf{c} = 0 \text{ (2.13)}$$

Where $\boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}$ (2.14) is a kxk matrix and $\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}$ (2.15) is an mxm matrix.

The above expressions (2.12) and (2.13) are called a general eigenvalue problem, where $\rho_{V_X V_Y}^2$ is vector of eigenvalues, and **c** and **d** are the respective eigenvectors. Nevertheless there may be a simple way to understand what they really mean, we can take some simple examples.

We could alternatively make use of a the general rule regarding quadratic, constrained optimization problems, involving matrices.

Let us define the following general problem:

$\max_{\mathbf{a}}\left\{\mathbf{a}^{\mathbf{T}}\mathbf{Q}\mathbf{a}\right\}$ subject to $\mathbf{a}^{\mathbf{T}}\mathbf{a} = \sum a_i^2 = 1$, where **Q** is a symmetric, quadratic, positive definite matrix.

The maximum of $\mathbf{a}^{\mathbf{T}}\mathbf{Q}\mathbf{a}$ will be the highest eigenvector of $\mathbf{Q}$ and vector a is going to equal the eigenvector of the highest eigenvalue of **Q**. If we rather wish to minimize the above objective function, then we should rather look for the smallest nonzero eigenvalue of **Q** as minimum with its respective eigenvector as the solution for **a.** In case of Principal Component analysis, where we only have a single vector of coefficients, we can simply use this.

If we have an asymmetric problem such as: $\max_{\mathbf{a}}\left\{\mathbf{a}^{\mathbf{T}}\mathbf{Q}\mathbf{b}\right\}$ subject to $\mathbf{a}^{\mathbf{T}}\mathbf{a} = \mathbf{b}^{\mathbf{T}}\mathbf{b} = 1$,where **Q** is a mxn matrix and **a** is an mx1 and **b** is an nx1 vector. Then the maximum (minimum) of $\mathbf{a}^{\mathbf{T}}\mathbf{Q}\mathbf{b}$ will be the highest (lowest) non-zero eigenvector of $\mathbf{Q}\mathbf{Q}^{\mathbf{T}}$ or $\mathbf{Q}^{\mathbf{T}}\mathbf{Q}$, which should be equal. Also the respective eigenvectors will be our estimates for **a** and **b**. $rank(\mathbf{Q}^{\mathbf{T}}\mathbf{Q}) = rank(\mathbf{Q}\mathbf{Q}^{\mathbf{T}}) = rank(\mathbf{Q}) = \min(m,n)$.

With canonical correlation $\mathbf{Q} = \boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}$, hence

$\mathbf{QQ^T} = \mathbf{\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}^{-\frac{1}{2}}\Sigma_{YY}^{-\frac{1}{2}}\Sigma_{XY}^{T}\Sigma_{XX}^{-\frac{1}{2}}} = \mathbf{\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{XY}^{T}\Sigma_{XX}^{-\frac{1}{2}}}$ and

$\mathbf{Q^TQ} = \mathbf{\Sigma_{YY}^{-\frac{1}{2}}\Sigma_{XY}^{T}\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}^{-\frac{1}{2}}} = \mathbf{\Sigma_{YY}^{-\frac{1}{2}}\Sigma_{XY}^{T}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-\frac{1}{2}}}$

Which is exactly as in (2.14) and (2.15).

Let us take a two-variate problem as special case 1. Then we have a single variable x and a single variable y with c=d=1. Then (2.12) simplifies into:

$\frac{\left(\sum xy\right)^2}{\sum y^2 \sum x^2} = \rho_{xy}^2$ or so we now that the Lagrange multipliers are going to equal $\lambda = \frac{\rho_{xy}}{2}$.

With (2.13) we obtain the same: $\frac{\left(\sum xy\right)^2}{\sum x^2 \sum y^2} = \rho_{xy}^2$.

But how many solutions can exist? The number of possible canonical correlations is given by the number of nonzero eigenvalues (2.12) and (2.13), that is, their rank. On order to find an answer we should remember a few rules on the rank of matrices.

Let A be an nxn matrix, B is an nxk matrix, C is an lxn matrix. Then:

1. A is invertible only if rank(A)=n
2. rank(AB)≤min(rank(A),rank(B))=n
3. If rank(B)=n then, rank(AB)=rank(A)=n
4. if rank(C)=n then rank(CA)=rank(A)=n
5. rank(A^TA)=rank(A)=rank(A^T)=n

Let us return to (2.14), using above rules:

$rank\left(\mathbf{\Sigma_{YY}^{-\frac{1}{2}}\Sigma_{XY}^{T}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-\frac{1}{2}}}\right) = rank\left(\mathbf{\Sigma_{XY}}\right) = \min(rank(X), rank(Y))$, hence the data matrix with the smallest number of columns will determine the number of nonzero eigenvalues. Obviously, since the eigenvalue reflect the canonical correlation between the canonical covariates which are created using X and Y with the eigenvectors as weights, we should choose the eigenvectors with the highest possible eigenvalue.

When estimating the weights or coefficients a and b, one should be aware that the eigenvectors of matrices (2.14) and (2.15) are estimated under the assumption that $\mathbf{a^T\Sigma_{XX}a} = \mathbf{c^Tc} = 1$and $\mathbf{b^T\Sigma_{YY}b} = \mathbf{d^Td} = 1$. If **X** has two columns, this would translate into the assumption that $a_1^2\sigma_{x_2}^2 + a_2^2\sigma_{x_2}^2 + 2a_1a_2\sigma_{x_1x_2} = 1$. If this is not true, the elements of the eigenvectors must be transformed as follows. Using (1.7) we obtain that

$\mathbf{a} = \boldsymbol{\Sigma}_{\mathbf{XX}}^{-\frac{1}{2}}\mathbf{c}$ (2.16) $\mathbf{b} = \boldsymbol{\Sigma}_{\mathbf{YY}}^{-\frac{1}{2}}\mathbf{d}$ (2.17) hence the correct values of vectors **a** and **b** are obtained if we first cary out the correction as in (2.16) and (2.17)

3. Example

We use the dataset mmreg.xls. This dataset contains data on 600 individuals. For each individuals we observe their scores in writing, reading, mathematics and science and also their gender.

We believe that the skills in reading and writing and the achievements in math and sciences are all aspects of skills in general. We also add gender to correct for possible differences among women and men.

The X set of variables has writing and reading scores, and the Y set of variables has math and science scores and a female dummy (1 is the individual is female, 0 otherwise).

1. canonical correlation in R step by step:

We can use the powerful matrix algebra capabilities of R to estimate the canonical correlation.

First we download the data:

mm <- read.csv("http://www.ats.ucla.edu/stat/data/mmreg.csv")

Then we assign the column names:

colnames(mm) <- c("Control", "Concept", "Motivation", "Read", "Write", "Math", "Science", "female")

Now our data is referred to as mm.

We create our data matrices X and Y:

X<-as.matrix(mm[,4:5])

Y<-as.matrix(mm[6:8])

That is, the new matrices should include the 4^{th} and 5^{th} and the 6^{th} to 8^{th} columns of the mm dataset.

Now we create the covariance matrices:

EX<-var(X)

EY<-var(Y)

EXY<-var(X,Y)

Let us look at the covariance matrix of X:

EX

Read Write

Read 102.07026 61.76924

Write 61.76924 94.60393

We will need to raise EX and EY to the power of -0.5. This is done by diagonalization. If a square matrix has eigenvectors and eigenvalues then it can be expressed as follows:

Matrix diagonalization and its use:

Let A be a square matrix which has eigenvalues and eigenvectors. A can be expressed then as $A = E\Lambda E^{-1}$, where E is the vector of A's eigenvalues (column 1 of E has the eigenvector associated with eigenvalue 1, and so on), and Λ is a diagonal matrix that has the eigenvalues of A in its diagonal.

$$A^k = \left(E\Lambda E^{-1}\right)_1 \left(E\Lambda E^{-1}\right)_2 ... \left(E\Lambda E^{-1}\right)_k = E\Lambda^k E^{-1}$$

So $A^{-\frac{1}{2}} = E\Lambda^{-\frac{1}{2}} E^{-1}$

We implement this procedure in R:

eex<-eigen(EX) this produces the eigenvectors and eigenvalues of EX and store then as object eex

eex

$values

[1] 160.21905 36.45514

$vectors

[,1] [,2]

[1,] -0.7281234 0.6854461

[2,] -0.6854461 -0.7281234

eey<-eigen(EY)

sqrEX=(eex$vectors)%*%diag(eex$values^-.5)%%solve(eex$vectors)* \

Above line is the diagonalization. Solve(A) returns the inverse of a matrix in R.

sqrEY=(eey$vectors)%*%diag(eey$values^-.5)%%solve(eey$vectors)*

Now we can estimate the eigenvalues and eigenvectors of the two matrices (2.14) and (2.15):

m1=sqrEX%%EXY%*%solve(EY)%*%t(EXY)%*%sqrEX*

m2=sqrEY%%t(EXY)%*%solve(EX)%*%(EXY)%*%sqrEY*

m1e<-eigen(m1)

m2e<-eigen(m2)

m1e

$values

[1] 0.6540663 0.1196161

$vectors

[,1] [,2]

[1,] -0.7386649 0.6740728

[2,] -0.6740728 -0.7386649

m2e

$values

[1] 6.540663e-01 1.196161e-01 7.759735e-17

$vectors

[,1] [,2] [,3]

[1,] 0.7111935 0.05691609 0.7006885

[2,] 0.6684659 -0.36329176 -0.6489780

[3,] 0.2176171 0.92993530 -0.2964172

and finally we correct as in (2.16) and (2.17):

sqrEX%%m1e$vectors*

[,1] [,2]

[1,] -0.05927735 0.1126200

[2,] -0.05227567 -0.1214192

sqrEY%%m2e$vectors*

[,1] [,2] [,3]

[1,] 0.06005211 0.01842777 0.1250149

[2,] 0.05490929 -0.03214626 -0.1211741

[3,] 0.44866702 1.87983468 -0.6178687

We found hence two non-zero eigenvectors, as expected, indicating the presence of two possible set of weights. The highest eigenvalue is 0.654, that is the canonical correlation is 0.809.

As a result, we should use the first eigenvectors as weights for the first canonical correlation:

$$V_{Y,1} = 0.06 math_i + 0.055 sci_i + 0.448\, female_i$$

and

$$V_{X,1} = -0.059 read_i - 0.052 write_i$$

There is also a second possible relationship as given by the second eigenvalue 0.1196 that translates to a canonical correlation of 0.346, which is quite low.

$$V_{Y,2} = .018 math_i - 0.032 sci_i + 1.88\, female_i$$

$$V_{X,2} = 0.113 read_i - 0.121 write_i$$

But how to interpret these results? There is no straightforward explanation, even though we may build a theory to explain these results.

A common way to interpret the results is to look at the correlation between our estimated canonical variates (also known as scores) and the original variables.

The two lines below gives us all canonical variates (the third is included too, but we should remember that only two exists)

scoreX<-X%%sqrEX%*%m1e$vectors*

scoreY<-Y%%sqrEY%*%m2e$vectors*

while the following commands results in correlation matrices between the canonical variates and between the canonical variates and the original datasets X and Y (these are called the canonical loadings, and are often used for interpreting the results).

cor(scoreY,scoreX)

[,1] [,2]

[1,] -8.087436e-01 -1.491684e-17

[2,] -3.512039e-17 -3.458557e-01

[3,] 2.610235e-17 -1.053131e-16

cor(scoreY,Y)

Math Science female

[1,] 0.90076377 0.8692854 0.1227007

[2,] -0.07434578 -0.3287935 0.9716349

[3,] 0.42789874 -0.3691039 -0.2021639

cor(scoreX,X)

Read Write

[1,] -0.9184895 -0.8849063

[2,] 0.3954454 -0.4657691

We can observe that the correlation between the two canonical correlates are negative. Theory may suggests that the most important common factor between the individual performance in the two areas of studies (humanities and science) is diligence and general study skills. Then the first canonical correlation should be strong and positive. Fortunately, the sign of the coefficient in the same vector can changed without the loss of any generality. If we multiply $V_{X,1}$ by -1 and let $V_{Y,1}$ as it was, we obtain:

$V_{X,1} = 0.059 read_i + 0.052 write_i$ (a relationship between general skills and humanities results)

$V_{Y,1} = 0.06 math_i + 0.055 sci_i + 0.448\, female_i$ (a relationship between general skills and science results)

The canonical loadings suggests that Reading and Writing are almost equally important in determining the results in humanities studies, while Math and Science are equally important in determining the results in sciences, and gender is of very low importance (0.12 correlation is so low, that actually we could even consider removing this variable). The second canonical correlation is quite low, yet we may attempt to assign some interpretation to it. It seems logical that people has different ability to deal with humanities and sciences. Some are more into numbers and models, while others are better in reading and writing, even if they are generally have the same level of devotion. Does the second canonical correlation capture

this relationship, perhaps? The correlation between the second canonical variates for X and Y are negative. If the underlying latent factor is the individual's better ability toward one of the subjects, this sounds quite logical.

$$V_{Y,2} = .018 math_i - 0.032 sci_i + 1.88\, female_i$$

$$V_{X,2} = 0.113 read_i - 0.121 write_i$$

The canonical loadings for Y for canonical correlation number 2 is very low for math and medium at best for sciences. The female dummy has a high correlation with the Y variables though. Gender seems to be a major determinant of the second canonical variate.

In case of X we find medium correlation between the second canonical variate for X and the original X variables. They are of roughly equal importance. Obviously it is very difficult to say more about this relationship, nevertheless, even if the second canonical covariance captures the effect of bias in abilities toward any fields of studies, it does not seem to explain much of the relationship between the scores in the two fields.

2. canonical correlation in R with built-in function

There are two built-in procedures for canonical correlation in R. The base package contains the function cancor, while the package CCA contains the function cc. To use the latter you will need to install the package CCA.

First we need to download the data and assign the variables into set X and Y, just as we did in the previous section.

```
mm <- read.csv("http://www.ats.ucla.edu/stat/data/mmreg.csv")

colnames(mm) <- c("Control", "Concept", "Motivation", "Read", "Write", "Math", "Science",
"female")

X<-as.matrix(mm[,4:5])

Y<-as.matrix(mm[6:8])
```

Now we can try cancor:

```
 cancor(X,Y)

$cor

[1] 0.8087436 0.3458557

$xcoef

          [,1]        [,2]
```

Read 0.002422007 -0.004601527

Write 0.002135926 0.004961054

$ycoef

[,1] [,2] [,3]

Math 0.002453663 0.0007529382 -0.005107971

Science 0.002243533 -0.0013134606 0.004951039

female 0.018332037 0.0768079630 0.025245430

$xcenter

Read Write

51.90183 52.38483

$ycenter

Math Science female

51.84900 51.76333 0.54500

As you can see, the derived weights are different from what we obtained in section 1. Yet, the canonical correlation and the ratios of the coefficients are the same. The reason is that while the eigenvalues are unique for the matrices the eigenvectors are not. If **v** is an eigenvector of matrix **A** then c**v**, where c is an arbitrary scalar is also an eigenvector of **A**. The R command eigen will yield normalized eigenvectors so that the sum of the square of the elements of the eigenvector equals 1.

We can also try the other function, cc. For this we need to load library CCA:

library(CCA) and then use the command cc:

cc<-cc(X,Y)

cc$xcoef

[,1] [,2]

Read -0.05927735 -0.1126200

Write -0.05227567 0.1214192

cc$ycoef

[,1] [,2]

Math -0.06005211 0.01842777

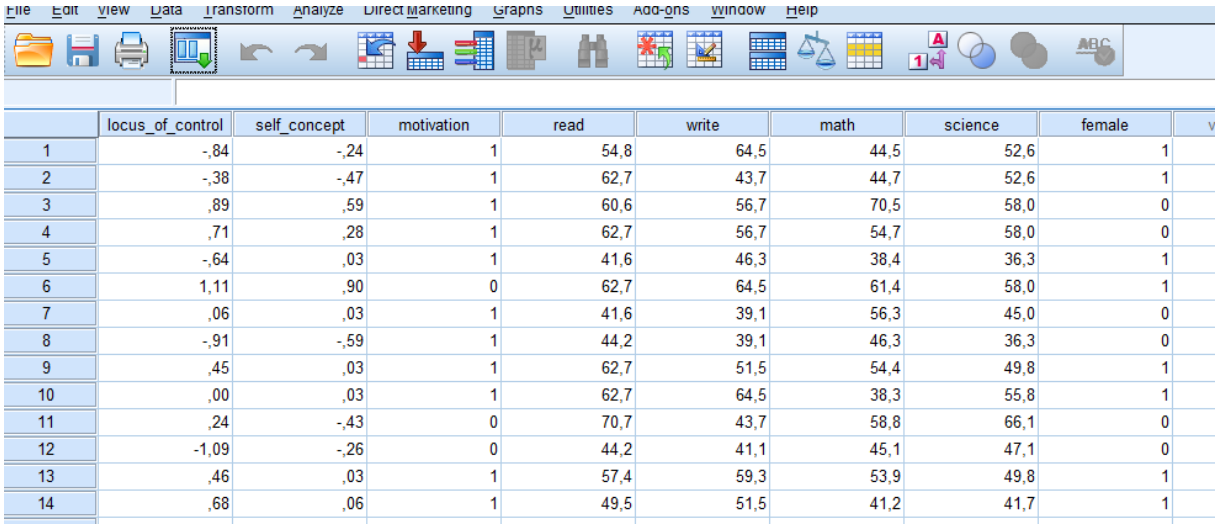
Science -0.05490929 -0.03214626

female -0.44866702 1.87983468

These estimates are the same what we obtained in section 1.

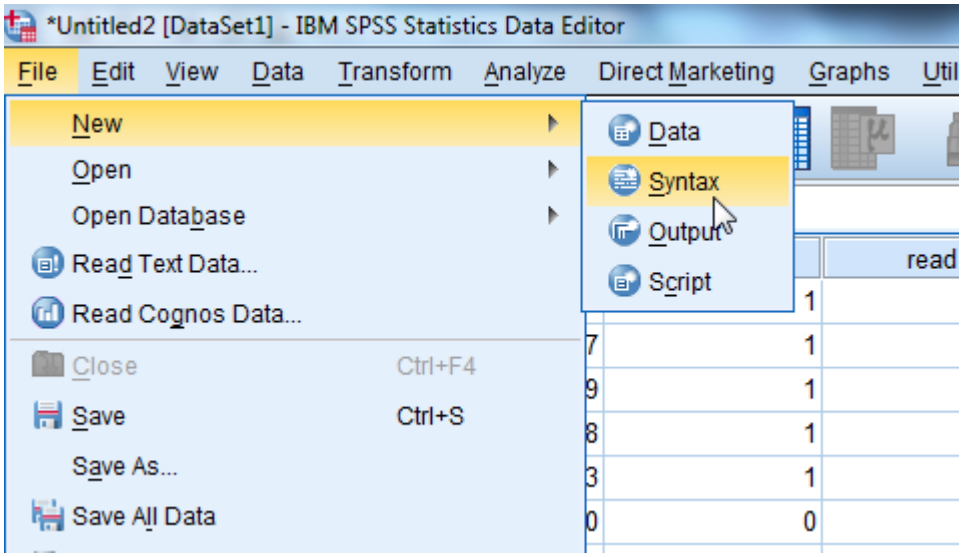
3. canonical correlation in SPSS

Canonical correlation is not accessible via the menu, but it can be quickly done with a syntax. First we need to open the file in SPSS.

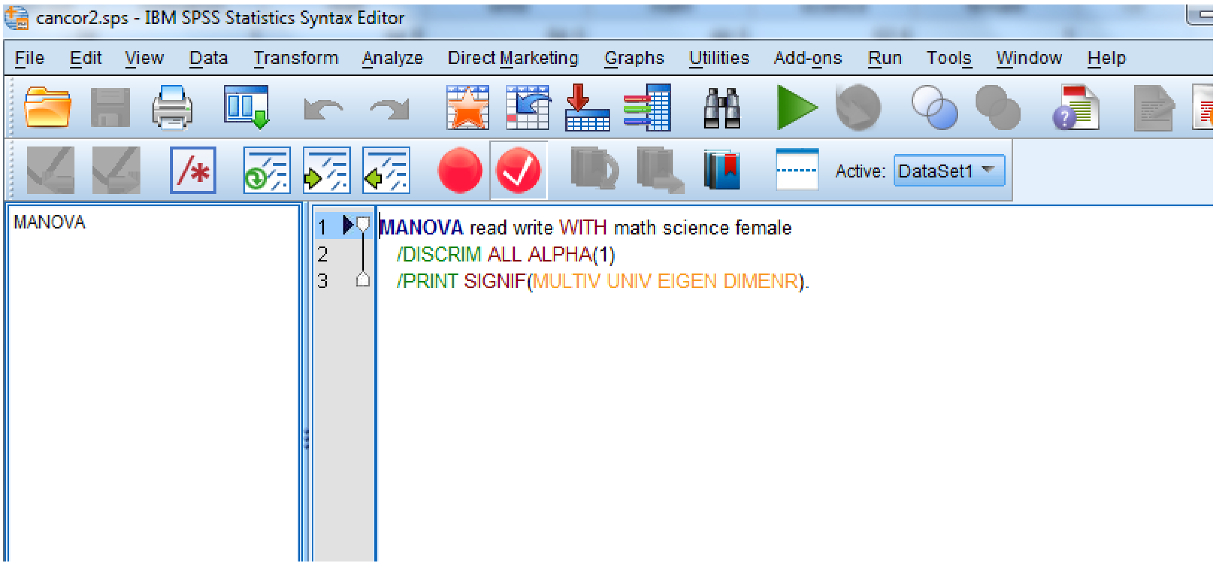


	locus_of_control	self_concept	motivation	read	write	math	science	female
1	-,84	-,24	1	54,8	64,5	44,5	52,6	1
2	-,38	-,47	1	62,7	43,7	44,7	52,6	1
3	,89	,59	1	60,6	56,7	70,5	58,0	0
4	,71	,28	1	62,7	56,7	54,7	58,0	0
5	-,64	,03	1	41,6	46,3	38,4	36,3	1
6	1,11	,90	0	62,7	64,5	61,4	58,0	1
7	,06	,03	1	41,6	39,1	56,3	45,0	0
8	-,91	-,59	1	44,2	39,1	46,3	36,3	0
9	,45	,03	1	62,7	51,5	54,4	49,8	1
10	,00	,03	1	62,7	64,5	38,3	55,8	1
11	,24	-,43	0	70,7	43,7	58,8	66,1	0
12	-1,09	-,26	0	44,2	41,1	45,1	47,1	0
13	,46	,03	1	57,4	59,3	53,9	49,8	1
14	,68	,06	1	49,5	51,5	41,2	41,7	1

Then we create a new syntax:



and we write the following in the syntax:



After the command MANOVA you specify the first set of variables, after the WITH you specify the second set. Use the green arrow to estimate the canonical correlation.

- -

The default error term in MANOVA has been changed from WITHIN CELLS to WITHIN+RESIDUAL. Note that these are the same for all full factorial designs.

* * * * * * * * * * * * * * * * * A n a l y s i s o f V a r i a n c e * * * * * * * * * * * * * * * * *

600 cases accepted.
0 cases rejected because of out-of-range factor values.
0 cases rejected because of missing data.
1 non-empty cell.

1 design will be processed.

- -

* * * * * * * * * * * * * * * * * A n a l y s i s o f V a r i a n c e -- Design 1 * * * * * * * * * * * * * * * * * *

EFFECT .. WITHIN CELLS Regression
Multivariate Tests of Significance (S = 2, M = 0, N = 296 1/2)

Test Name	Value	Approx. F	Hypoth. DF	Error DF	Sig. of F
Pillais	,77368	125,33858	6,00	1192,00	,000
Hotellings	2,02660	200,63292	6,00	1188,00	,000
Wilks	,30455	161,05437	6,00	1190,00	,000
Roys	,65407				

Note.. F statistic for WILKS' Lambda is exact.

- -

Eigenvalues and Canonical Correlations

Root No.	**Eigenvalue**	**Pct.**	**Cum. Pct.**	**Canon Cor.**	**Sq. Cor**
1	1,89073	93,29574	93,29574	,80874	,65407
2	,13587	6,70426	100,00000	,34586	,11962

We received the same canonical correlations as before.

- -

Dimension Reduction Analysis

Roots	Wilks L.	F	Hypoth. DF	Error DF	Sig. of F
1 TO 2	,30455	161,05437	6,00	1190,00	,000
2 TO 2	,88038	40,48871	2,00	596,00	,000

The above test suggests that both canonical correlations are significantly above zero.

EFFECT .. WITHIN CELLS Regression (Cont.)
Univariate F-tests with (3;596) D. F.

Variable	Sq. Mul. R	Adj. R-sq.	Hypoth. MS	Error MS	F	Sig. of F
read	,57049	,56833	11626,61396	44,06082	263,87650	,000
write	,53812	,53580	10164,72469	43,91540	231,46151	,000

Raw canonical coefficients for DEPENDENT variables
Function No.

Variable	1	2
read	,05928	-,11262
write	,05228	,12142

The raw canonical coefficients are as above. Observe that here both eigenvectors are multiplied by -1 relative to our results. This can be done without any loss of generality or explanatory power. One unit increase in reading increases the canonical variate by .059 units.

Standardized canonical coefficients for DEPENDENT variables
Function No.

Variable	1	2
read	,59888	-1,13780
write	,50846	1,18098

The standardized canonical coefficients are obtained by dividing the raw coefficients by the standard deviation of the X set of variables. They can be interpreted as standardized regression coefficients: one standard deviation increase in reading result will increase the canonical variate by 0.5989 standard deviations.

Correlations between DEPENDENT and canonical variables
Function No.

Variable	1	2
read	,91849	-,39545
write	,88491	,46577

The loadings, that is the correlation between the canonical variates and the X variables.

Variance in dependent variables explained by canonical variables

CAN. VAR.	Pct Var DEP	Cum Pct DEP	Pct Var COV	Cum Pct COV
1	81,33410	81,33410	53,19789	53,19789
2	18,66590	100,00000	2,23274	55,43064

Raw canonical coefficients for COVARIATES
Function No.

COVARIATE	1	2
math	,06005	,01843
science	,05491	-,03215
female	,44867	1,87983

Standardized canonical coefficients for COVARIATES
CAN. VAR.

COVARIATE	1	2
math	,56537	,17349
science	,53296	-,31202
female	,22361	,93688

Correlations between COVARIATES and canonical variables
CAN. VAR.

Covariate	1	2
math	,90076	-,07435
science	,86929	-,32879
female	,12270	,97163

Variance in covariates explained by canonical variables

CAN. VAR.	Pct Var DEP	Cum Pct DEP	Pct Var COV	Cum Pct COV
1	34,49301	34,49301	52,73627	52,73627
2	4,21729	38,71031	35,25689	87,99316

Regression analysis for WITHIN CELLS error term
--- Individual Univariate ,9500 confidence intervals
Dependent variable .. read

COVARIATE	B	Beta	Std. Err.	t-Value	Sig. of t	Lower -95%	CL- Upper
math	,4252118194	,3962450713	,03795	11,20589	,000	,35069	,49973
science	,4564972995	,4385679372	,03712	12,29811	,000	,38360	,52940
female	,7696512759	,0379673721	,55029	1,39862	,162	-,31110	1,85040

Dependent variable .. write

COVARIATE	B	Beta	Std. Err.	t-Value	Sig. of t	Lower -95%	CL- Upper
math	,4468872368	,4325651569	,03788	11,79660	,000	,37249	,52129
science	,3318482170	,3311564295	,03706	8,95483	,000	,25907	,40463
female	6,0684769537	,3109505214	,54939	11,04594	,000	4,98951	7,14744