Canonical Correlation Analysis- CCA

John Reddy Peasari

5/1/2020

Question: A survey is conducted among the 600 college senior students to see if there is any realationship between four different acedemic variables (test scores) and three psychological variables. Here, they are more intersted in the number of dimensions (canonical variates) required to understand the relationship between the two sets of variables

Loading the data

```
data = read.csv("CCA.csv")
head(data)
```

```
##   locus_of_control self_concept motivation read write math science female
## 1            -0.84        -0.24       1.00 54.8  64.5 44.5    52.6      1
## 2            -0.38        -0.47       0.67 62.7  43.7 44.7    52.6      1
## 3             0.89         0.59       0.67 60.6  56.7 70.5    58.0      0
## 4             0.71         0.28       0.67 62.7  56.7 54.7    58.0      0
## 5            -0.64         0.03       1.00 41.6  46.3 38.4    36.3      1
## 6             1.11         0.90       0.33 62.7  64.5 61.4    58.0      1
```

```
colnames(data) <- c("Control", "Concept", "Motivation", "Read", "Write", "Math","Science", "Sex")
```

It can be obsserved that, the dataset has 8 columns, first three columns are realted to psychological variables and the last four columns are realted to the academic variables. Now, in order to perform canoniacl correaltion analysis we have to split the data into two datatables where one has set of predictor variables and the other has outcome variables.

```
psych_var <- data[, 1:3]
acad_var <- data[, 4:8]
head(psych_var)
```

```
##   Control Concept Motivation
## 1   -0.84   -0.24       1.00
## 2   -0.38   -0.47       0.67
## 3    0.89    0.59       0.67
## 4    0.71    0.28       0.67
## 5   -0.64    0.03       1.00
## 6    1.11    0.90       0.33
```

```r
head(acad_var)
```

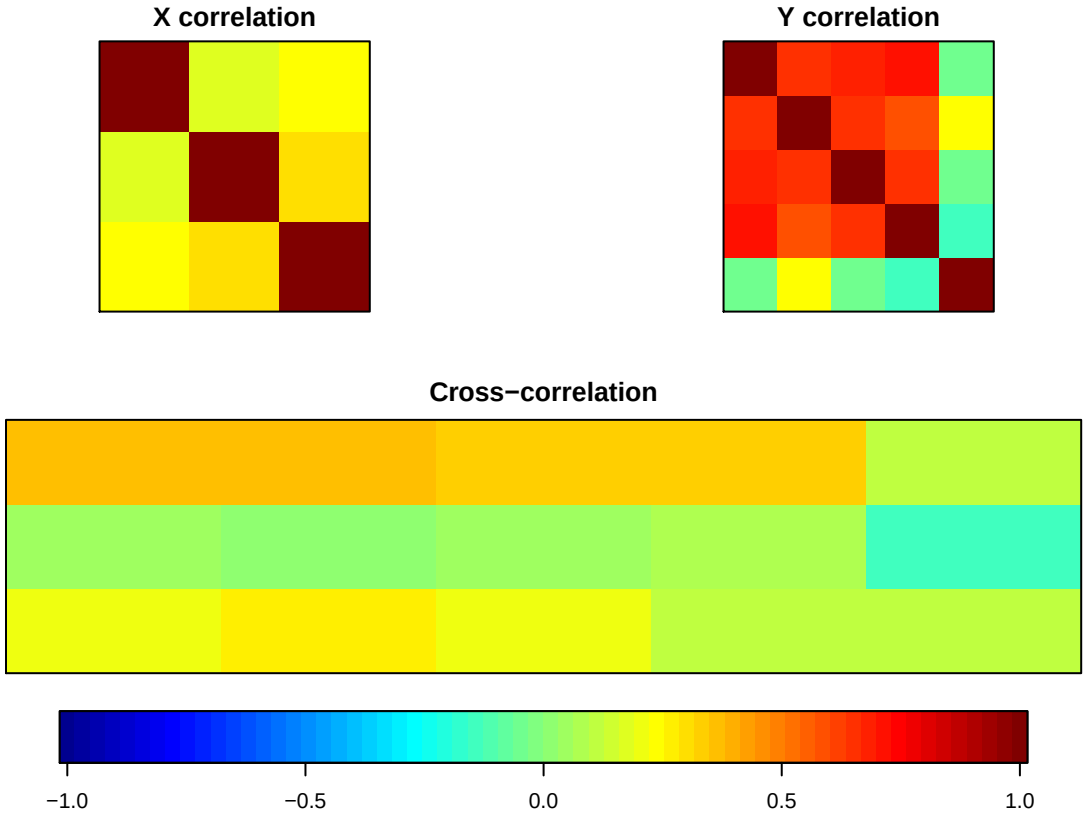
```
##   Read Write Math Science Sex
## 1 54.8  64.5 44.5    52.6   1
## 2 62.7  43.7 44.7    52.6   1
## 3 60.6  56.7 70.5    58.0   0
## 4 62.7  56.7 54.7    58.0   0
## 5 41.6  46.3 38.4    36.3   1
## 6 62.7  64.5 61.4    58.0   1
```

```r
## Understanding correlations within and between two variables
correl <- matcor(psych_var, acad_var)
correl
```

```
## $Xcor
##              Control   Concept Motivation
## Control    1.0000000 0.1711878  0.2451323
## Concept    0.1711878 1.0000000  0.2885707
## Motivation 0.2451323 0.2885707  1.0000000
##
## $Ycor
##                Read     Write       Math    Science         Sex
## Read     1.00000000 0.6285909  0.6792757  0.6906929 -0.04174278
## Write    0.62859089 1.0000000  0.6326664  0.5691498  0.24433183
## Math     0.67927568 0.6326664  1.0000000  0.6495261 -0.04821830
## Science  0.69069291 0.5691498  0.6495261  1.0000000 -0.13818587
## Sex     -0.04174278 0.2443318 -0.0482183 -0.1381859  1.00000000
##
## $XYcor
##              Control     Concept Motivation        Read      Write       Math
## Control    1.0000000  0.17118778 0.24513227  0.37356505 0.35887684  0.3372690
## Concept    0.1711878  1.00000000 0.28857075  0.06065584 0.01944856  0.0535977
## Motivation 0.2451323  0.28857075 1.00000000  0.21060992 0.25424818  0.1950135
## Read       0.3735650  0.06065584 0.21060992  1.00000000 0.62859089  0.6792757
## Write      0.3588768  0.01944856 0.25424818  0.62859089 1.00000000  0.6326664
## Math       0.3372690  0.05359770 0.19501347  0.67927568 0.63266640  1.0000000
## Science    0.3246269  0.06982633 0.11566948  0.69069291 0.56914983  0.6495261
## Sex        0.1134108 -0.12595132 0.09810277 -0.04174278 0.24433183 -0.0482183
##                Science         Sex
## Control     0.32462694  0.11341075
## Concept     0.06982633 -0.12595132
## Motivation  0.11566948  0.09810277
## Read        0.69069291 -0.04174278
## Write       0.56914983  0.24433183
## Math        0.64952612 -0.04821830
## Science     1.00000000 -0.13818587
## Sex        -0.13818587  1.00000000
```

THe function "matcor" is ued to understand correlations and displays all the correlations within X variable and Y variable and between X and Y as cross correlation.

```
img.matcor(correl, type = 2)
```



Correlation matrices for psychological variables (upper-left), academic variables (upper-right) and the bottom middle figure shows cross-correlation between psychological and academic variables. The strength of correaltion depends up on the intensity of the colour in the coloured bar from blue (negative correlation) to red (positive correlation). It looks like the observations between psychological and academic variables are not much correalated. Let us examine the real relationship by performing canoniacl correlation analysis using a package "CCA".

```
## Displaying the canonical correlation coefficients
CC1 <- cc(psych_var,acad_var)
CC1$cor
```

```
## [1] 0.4640861 0.1675092 0.1039911
```

Here, the value [0.4640861 0.1675092 0.1039911] are called canonical correlation coefficients or canonical variates. As our smallest data table is the psychological set that has only three observations (control,concept and motivation). So, the number of varaites will be equal to the number of observations in the smallest data table. Hence, there will be three canonical caorrelation coefficients.

```
## Displaying raw canonical coeffients
CC1[3:4]
```

```
## $xcoef
##                  [,1]       [,2]       [,3]
```

```
## Control    -1.2538339 -0.6214776 -0.6616896
## Concept     0.3513499 -1.1876866  0.8267210
## Motivation -1.2624204  2.0272641  2.0002283
##
## $ycoef
##                 [,1]         [,2]         [,3]
## Read    -0.044620600 -0.004910024  0.021380576
## Write   -0.035877112  0.042071478  0.091307329
## Math    -0.023417185  0.004229478  0.009398182
## Science -0.005025152 -0.085162184 -0.109835014
## Sex     -0.632119234  1.084642326 -1.794647036
```

Above displayed values are the raw canonical coefficients, which will define the linear realtionship between the variables in a given set and canonical variets.

These raw canonical values are initially used for finding the linear combination of observations within each set for three times (i.e.,) to calculate each canonical varaite. These values are similar to regression coefficients.

```
## Calculating canonial loadings
CC2 <- comput(psych_var,acad_var, CC1)
## Displaying canonical loadings
CC2[3:6]
```

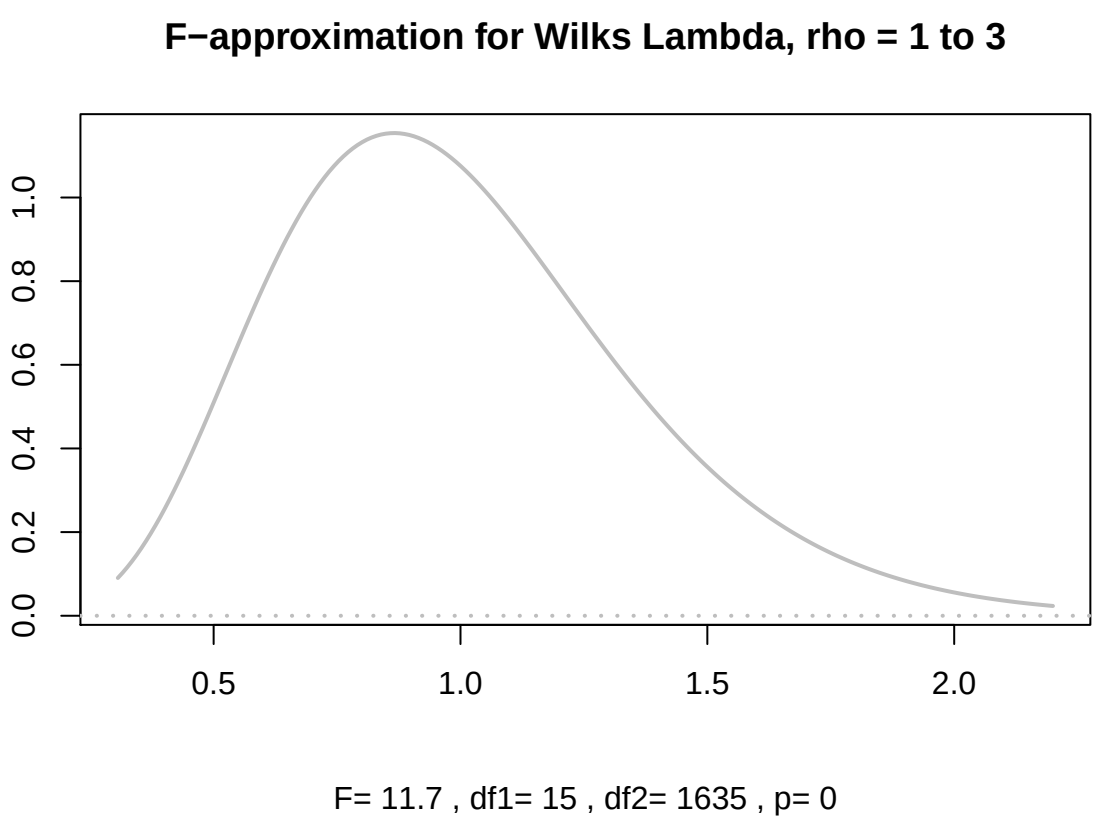
```
## $corr.X.xscores
##                   [,1]       [,2]       [,3]
## Control    -0.90404631 -0.3896883 -0.1756227
## Concept    -0.02084327 -0.7087386  0.7051632
## Motivation -0.56715106  0.3508882  0.7451289
##
## $corr.Y.xscores
##                [,1]        [,2]        [,3]
## Read    -0.3900402 -0.06010654  0.01407661
## Write   -0.4067914  0.01086075  0.02647207
## Math    -0.3545378 -0.04990916  0.01536585
## Science -0.3055607 -0.11336980 -0.02395489
## Sex     -0.1689796  0.12645737 -0.05650916
##
## $corr.X.yscores
##                    [,1]        [,2]        [,3]
## Control    -0.419555307 -0.06527635 -0.01826320
## Concept    -0.009673069 -0.11872021  0.07333073
## Motivation -0.263206910  0.05877699  0.07748681
##
## $corr.Y.yscores
##                [,1]        [,2]       [,3]
## Read    -0.8404480 -0.35882541  0.1353635
## Write   -0.8765429  0.06483674  0.2545608
## Math    -0.7639483 -0.29794884  0.1477611
## Science -0.6584139 -0.67679761 -0.2303551
## Sex     -0.3641127  0.75492811 -0.5434036
```

Next, canonical loadings of the observation/variables on the precalculated canonical domensions (variates). These values are the correlations between the canonical variates and the variables.

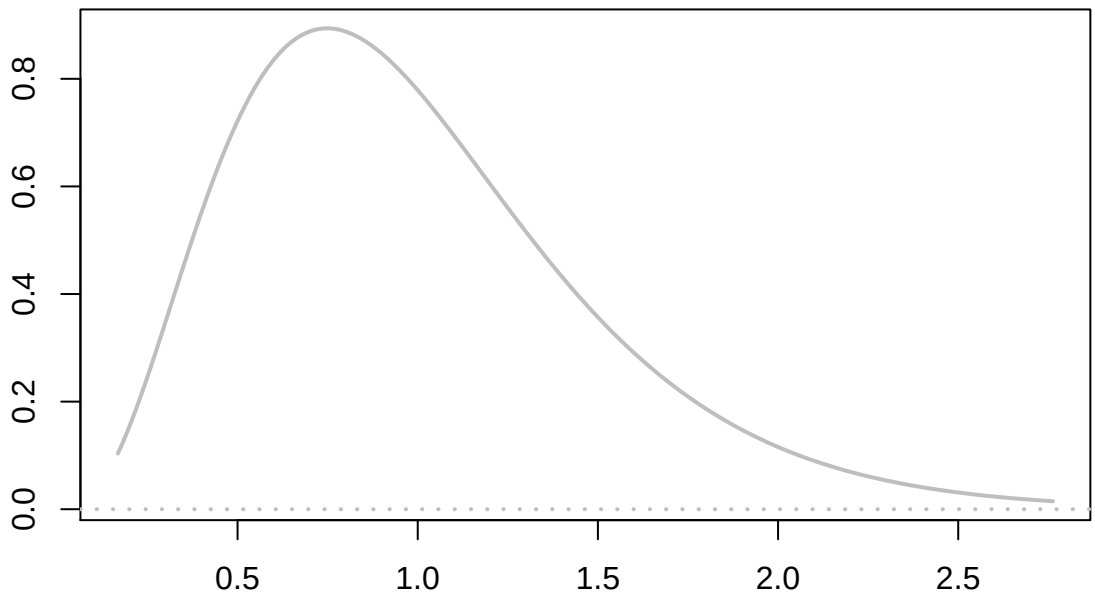
```r
## Testing Canonical variates or canonical dimensions
rho <- CC1$cor
## Define all parameters for the p.asym function to compute
N <- dim(psych_var)[1]
p <- length(psych_var)
q <- length(acad_var)
## Calculate p-values with the help of F approximations using various test statistics
# p.asym(rho,N,p,q,tstat="Wilks")

res1 <- p.asym(rho,N,p,q,tstat="Wilks")
```

```
## Wilks' Lambda, using F-approximation (Rao's F):
##               stat    approx df1      df2     p.value
## 1 to 3:  0.7543611 11.715733  15 1634.653 0.000000000
## 2 to 3:  0.9614300  2.944459   8 1186.000 0.002905057
## 3 to 3:  0.9891858  2.164612   3  594.000 0.091092180
```

```r
plt.asym(res1,rhostart=1)
```

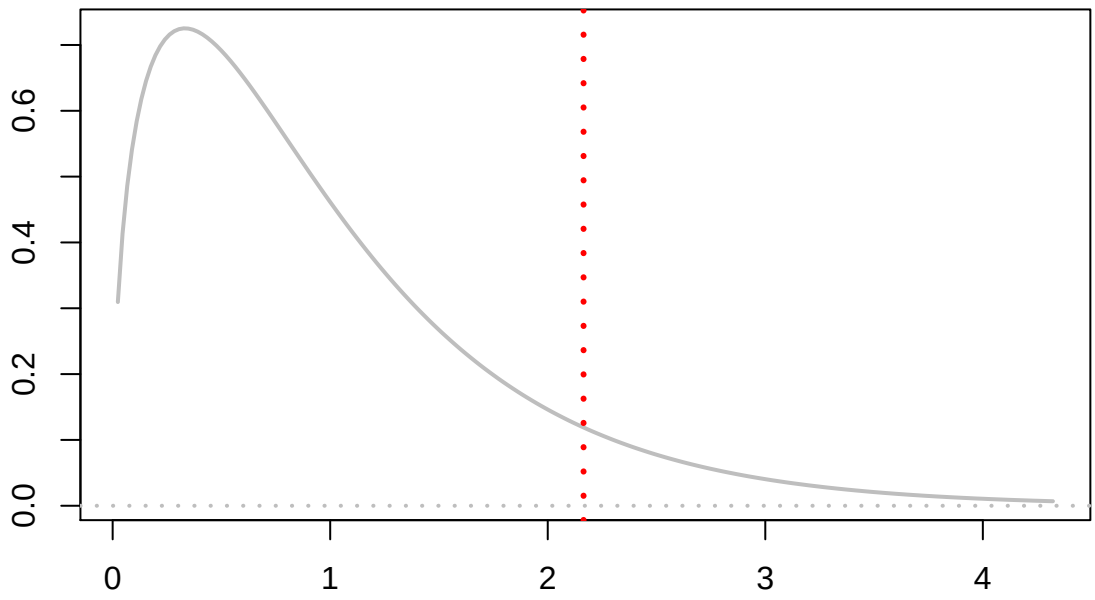



```r
plt.asym(res1,rhostart=2)
```



F= 2.94 , df1= 8 , df2= 1186 , p= 0.00291

```
plt.asym(res1,rhostart=3)
```



F= 2.16 , df1= 3 , df2= 594 , p= 0.0911

```
p.asym(rho,N,p,q,tstat="Hotelling")
```

```
##  Hotelling-Lawley Trace, using F-approximation:
##                 stat    approx df1  df2     p.value
## 1 to 3:  0.31429738 12.376333  15 1772 0.000000000
## 2 to 3:  0.03980175  2.948647   8 1778 0.002806614
## 3 to 3:  0.01093238  2.167041   3 1784 0.090013176
```

```
p.asym(rho,N,p,q,tstat="Pillai")
```

```
##  Pillai-Bartlett Trace, using F-approximation:
##                 stat    approx df1  df2     p.value
## 1 to 3:  0.25424936 11.000571  15 1782 0.000000000
## 2 to 3:  0.03887348  2.934093   8 1788 0.002932565
## 3 to 3:  0.01081416  2.163421   3 1794 0.090440474
```

```
p.asym(rho,N,p,q,tstat="Roy")
```

```
##  Roy's Largest Root, using F-approximation:
##                stat   approx df1 df2 p.value
## 1 to 1:  0.2153759 32.61008   5 594       0
##
##  F statistic for Roy's Greatest Root is an upper bound.
```

Result table analysis for test statistic

Stat: Gives the value of statistic i.e., it can be Wilks, Hottling, Pillai or Roy.

approx: Gives corresponding F - approximation for each significant statistic test.

df1: Numerator degrees of freedom for the F - approximation

df2: Denominator degrees of freedom for the F - approximation

p value: p-value

Next, we use p.asym function that can calculate F approximations and p value for each test statisic (Wilks, Hotelling, Pillai and Roy). It can be observed that all the test statistics (expect Roy) displayed same p-value for each canonical variate. The first test of every significant test (dimension) determine whether all the dimensions are significant or not (observed F value ~ 11.72) and followed by 2 and 3 dimensions combined (observed F value ~ 2.9) and finally dimension is tested itself whether is significant or not (observed F value ~ 2.16)

Test of dimensionlity can be inferred from test statistics results. Out of three dimensions ony two are satistically significant with a p value of 0 and 0.002 at a threshold p value of 0.05 in all the test statistics at correlations 0.46 and 0.16.

```
## Displaying standardized psycological variables (psych_var) canonical coefficients diagonal matrix
STD1 <- diag(sqrt(diag(cov(psych_var))))
STD1 %*% CC1$xcoef
```

```
##            [,1]       [,2]       [,3]
## [1,] -0.8404196 -0.4165639 -0.4435172
## [2,]  0.2478818 -0.8379278  0.5832620
## [3,] -0.4326685  0.6948029  0.6855370
```

```
## Displaying standardized academic variables (acad_var) canonical coefficients diagonal matrixjr
STD2 <- diag(sqrt(diag(cov(acad_var))))
STD2 %*% CC1$ycoef
```

```
##             [,1]        [,2]        [,3]
## [1,] -0.45080116 -0.04960589  0.21600760
## [2,] -0.34895712  0.40920634  0.88809662
## [3,] -0.22046662  0.03981942  0.08848141
## [4,] -0.04877502 -0.82659938 -1.06607828
## [5,] -0.31503962  0.54057096 -0.89442764
```

Computing standardized coefficients helps in evaluating comparions among the variables easily. As the third canonical dimension is not significant we only consider one and two dimensions. First standardized matrix indicates psycological variables and the second matrix indicates academic variables. It can be inferrend from the above results that, locus of control (-0.84)(in psycological table) influenced first canonical dimension mostly. Similarly, for motivation (0.69) and seld-concept (-0.84).

In case of academic variables, writing (-0.35), reading (-0.45) and gender (-0.32) influenced first dimension. Whereas, dominating variables in the second dimension are writing gender (0.54), Sciecne (-0.83) and writing (0.41).