35459 Multivariate Statistics

Week 9 or 10 Seminar - Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a multivariate statistical technique to identify and quantify (measure the strength of) the association between two sets of variables. This technique is based on projecting the two sets of variables into indices that contain the maximum amount of information about the association between the two sets of variables.

The idea of CCA is to construct pairs of linear combinations of the variables from each set, known as canonical variables, and then to measure the correlation, known as canonical correlation, of each pair of canonical variables. The high-dimensional relationship between the original sets of variables is hopefully to be summarized by a few pairs of canonical variables.

Consider two sets of variables $\boldsymbol{x}$ and $\boldsymbol{y}$, where $\boldsymbol{x}$ contains p_x variables and $\boldsymbol{y}$ contains p_y variables. Then we define u_i as a linear combination of the variables in $\boldsymbol{x}, u_i = \boldsymbol{a}^T\boldsymbol{x}$, and each v_i is a linear combination of the variables $\boldsymbol{y}$, $v_i = \boldsymbol{b}^T\boldsymbol{y}$, with the coefficients (a_i, b_i), for $i = 1, \ldots, \min(p_x, p_y)$, being chosen so that the u_i and v_i satisfy the following:

- The u_i are mutually uncorrelated ($\text{cov}(u_i, u_j) = 0$ for $i \neq j$).
- The v_i are mutually uncorrelated ($\text{cov}(v_i, v_j) = 0$ for $i \neq j$).
- The correlation between u_j and v_j is R_i for $i = 1, \ldots, \min(p_x, p_y)$, where $R_1 > R_2 > \ldots > R_{\min(p_x,p_y)}$.
- The u_i are uncorrelated with all v_j except for v_i ($\text{cov}(u_i, v_j) = 0$ for $i \neq j$).

To test whether at least one canonical correlation is significant, we can use the test statistic

$$\Psi_0^2 = -\left(n - 1 - \frac{1}{2}(p_x + p_y + 1)\right) \sum_{i=1}^{\min(p_x,p_y)} \log(1 - \lambda_i),$$

which has $p_x \times p_y$ degrees of freedom. λ_i is the square of the i^{th} canonical correlation. This assumes that the data follow a multivariate normal distribution. To sequentially test individual correlations (if the previous test is significant), we use

$$\Psi_j^2 = -\left(n - 1 - \frac{1}{2}(p_x + p_y + 1)\right) \sum_{i=j+1}^{\min(p_x,p_y)} \log(1 - \lambda_i),$$

where we are testing the j^{th} cannonical correlation. This test statistic has a chi squared distribution with $(p_x - j) \times (p_y - j)$ degrees of freedom.

Example

Amitirptyline is prescribed by some physicians as an antidepressant. However, there are also conjectured side effects that seem to be related to the use of the drug, such as irregular heartbeat, abnormal blood pressure and irregular waves on and ECG. Data gathered from 17 patients who were admitted to the hospital after an amitriptyline overdose had the following response variables measured

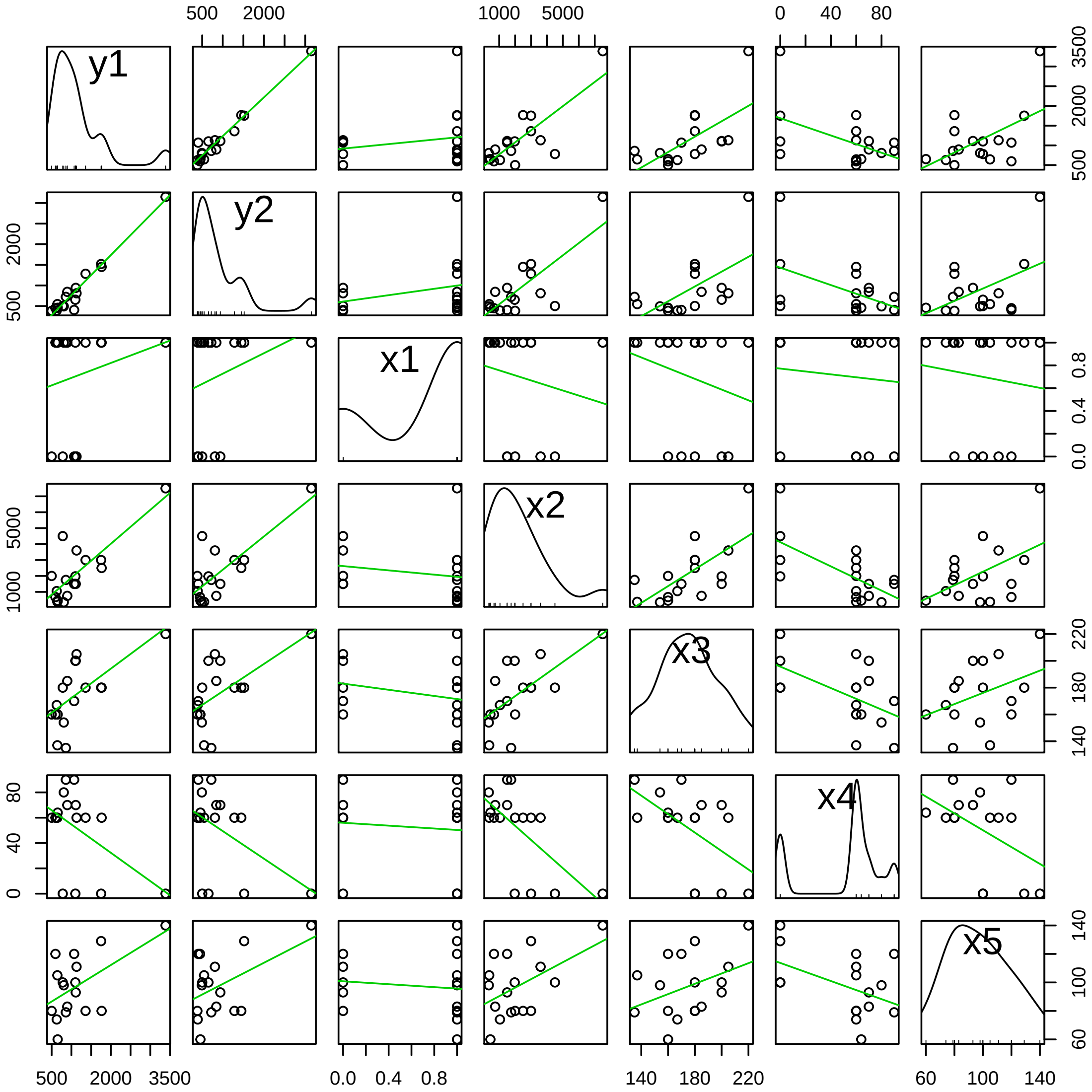
- Total TCAD plasma level (y_1)
- Amount of amitriptyline present in TCAD plasma level (y_2)

and the following predictive variables measured

- Gender (1 if male, 0 if female) (x_1)
- Amount of antidepressants taken at time of overdose (x_2)
- PR wave measurement (x_3)
- Diastolic blood pressure (x_4)
- QRS wave measurement (x_5)

```r
library(car)
```

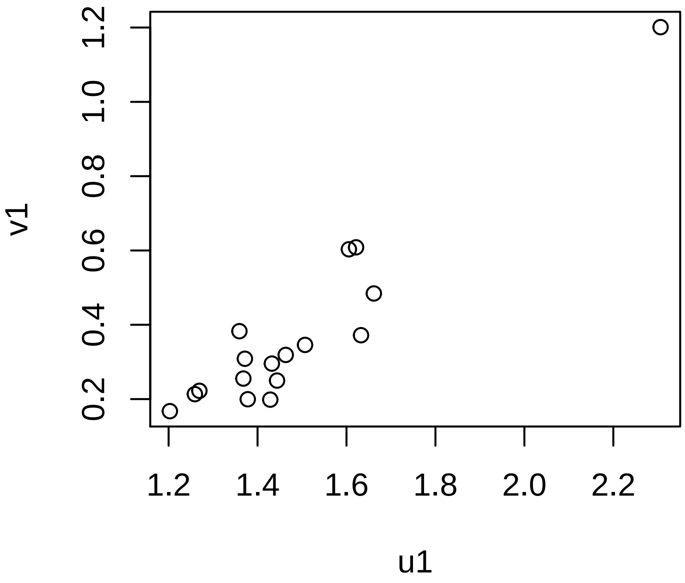
```r
y1<-c(3389, 1101, 1131, 596, 896, 1767, 807, 1111, 645, 628, 1360,
      652, 860, 500, 781, 1070, 1754)
y2<-c(3149, 653, 810, 448, 844, 1450, 493, 941, 547, 392, 1283, 458,
      722, 384, 501, 405, 1520)
x1<-c(1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1)
x2<-c(7500, 1975, 3600, 675, 750, 2500, 350, 1500, 375, 1050, 3000,
      450, 1750, 2000, 4500, 1500, 3000)
x3<-c(220, 200, 205, 160, 185, 180, 154, 200, 137, 167, 180, 160, 135,
      160, 180, 170, 180)
x4<-c(0, 0, 60, 60, 70, 60, 80, 70, 60, 60, 60, 64, 90, 60, 0, 90, 0)
x5<-c(140, 100, 111, 120, 83, 80, 98, 93, 105, 74, 80, 60, 79, 80, 100,
      120, 129)

scatterplotMatrix(cbind(y1,y2,x1,x2,x3,x4,x5),smoother=FALSE)
```



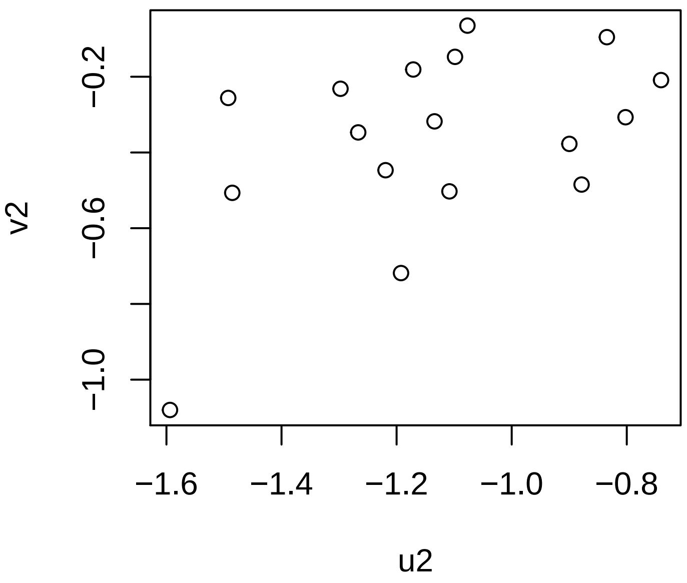
```
cor(cbind(x1,x2,x3,x4,x5,y1,y2))

##          x1      x2      x3       x4      x5      y1      y2
## x1  1.00000 -0.1726 -0.2326 -0.08489 -0.1113  0.1935  0.2645
## x2 -0.17262  1.0000  0.6696 -0.65181  0.5032  0.8075  0.7884
## x3 -0.23265  0.6696  1.0000 -0.54060  0.3882  0.6548  0.6062
## x4 -0.08489 -0.6518 -0.5406  1.00000 -0.4595 -0.4952 -0.4799
## x5 -0.11128  0.5032  0.3882 -0.45949  1.0000  0.5469  0.4733
## y1  0.19354  0.8075  0.6548 -0.49520  0.5469  1.0000  0.9761
## y2  0.26449  0.7884  0.6062 -0.47989  0.4733  0.9761  1.0000

ami_cancor<-cancor(cbind(x1,x2,x3,x4,x5),cbind(y1,y2))
ami_cancor
```

```
## $cor
## [1] 0.9437 0.4853
##
## $xcoef
##         [,1]       [,2]       [,3]       [,4]       [,5]
## x1 0.2709339  3.328e-01 -0.2022929 -1.855e-02  3.266e-01
## x2 0.0001123  8.781e-05  0.0001303 -8.323e-05 -5.177e-05
## x3 0.0037755 -4.207e-03 -0.0136155 -2.979e-03 -1.641e-03
## x4 0.0027828  2.020e-04 -0.0002260 -1.070e-02  3.045e-03
## x5 0.0025903 -8.571e-03  0.0024838  3.213e-04  1.004e-02
##
## $ycoef
##         [,1]      [,2]
## y1 0.0002430 -0.001639
## y2 0.0001199  0.001663
##
## $xcenter
##        x1        x2        x3        x4        x5
##    0.7059 2145.5882  174.8824   52.0000   97.1765
##
## $ycenter
##     y1     y2
## 1120.5  882.4
```

```
u1<-as.matrix(cbind(x1,x2,x3,x4,x5))%*%ami_cancor$xcoef[,1]
v1<-as.matrix(cbind(y1,y2))%*%ami_cancor$ycoef[,1]
plot(u1,v1)
```



```
u2<-as.matrix(cbind(x1,x2,x3,x4,x5))%*%ami_cancor$xcoef[,2]
v2<-as.matrix(cbind(y1,y2))%*%ami_cancor$ycoef[,2]
plot(u2,v2)
```



Test whether any canonical correlations are significant

```
n<-length(x1)
px<-2
py<-5
l1<-ami_cancor$cor[1]^2
l2<-ami_cancor$cor[2]^2
psi0<--(n-1-(px+py+1)/2)*(log(1-l1)+log(1-l2))
print(1-pchisq(psi0,px*py))

## [1] 0.0009366
```

Test whether the first canonical correlation is significant

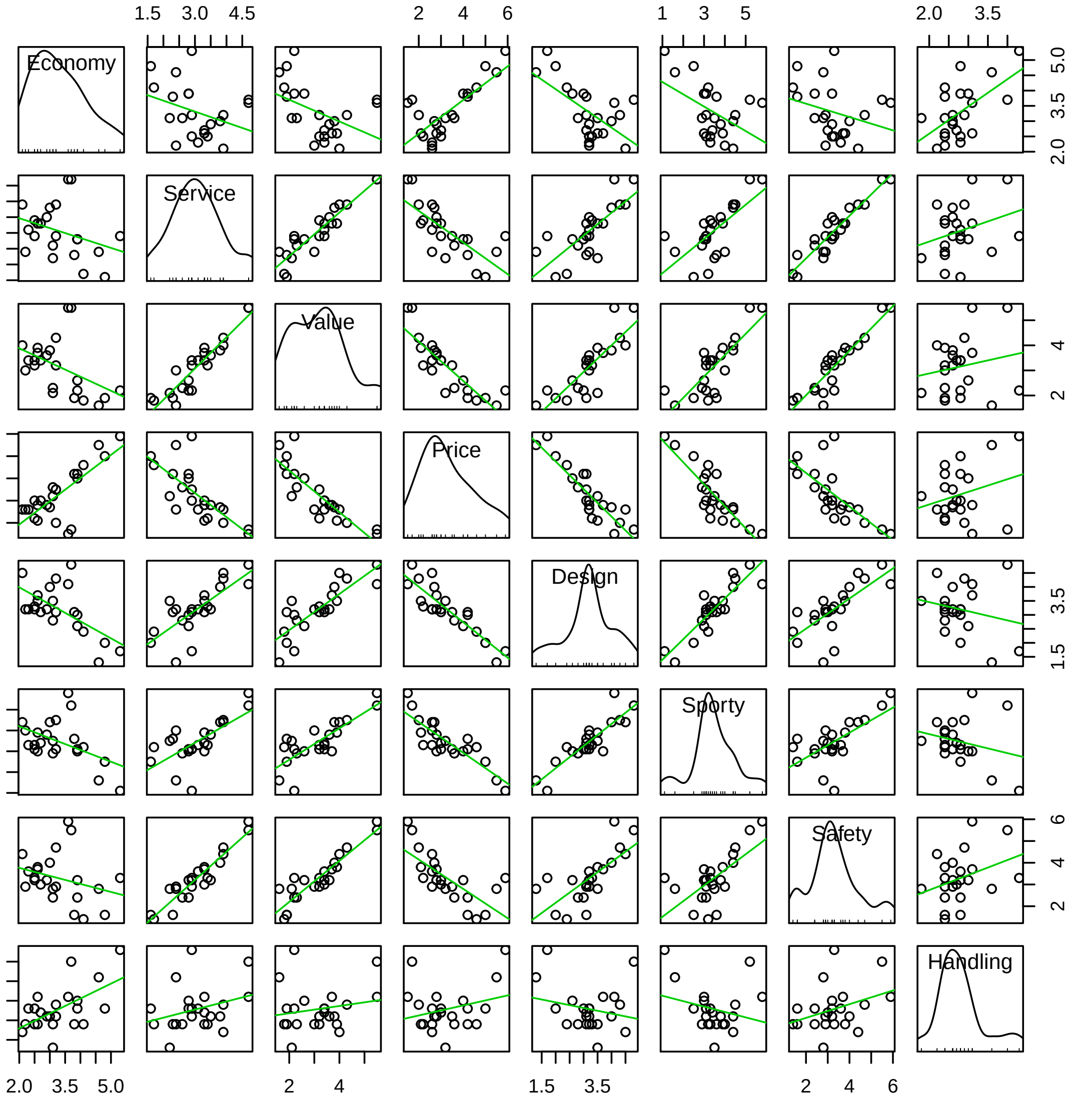
```
psi1<--(n-1-(px+py+1)/2)*(log(1-l2))
print(1-pchisq(psi1,(px-1)*(py-1)))

## [1] 0.5213
```

Car Ratings

The data set carmarks.csv contains average ratings of 24 types of car across 40 raters. The raters marked the cars between 1 and 6, with a 1 indicating a very good rating and a 6 indicating a very poor rating. The features of the car that were rated were:

- Economy
- Service
- Value retention
- Price (low is good)
- Design
- Sporty car
- Safety
- Handling

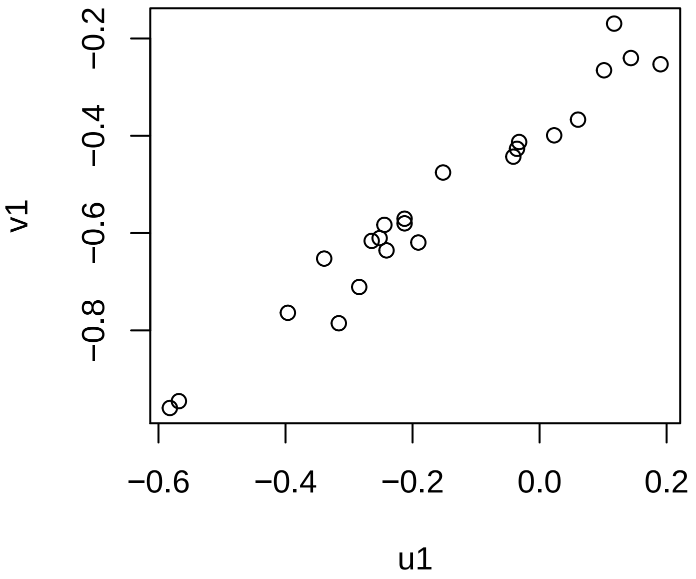
```
carratings<-read.csv("C:/Documents/carmarks.csv")
scatterplotMatrix(carratings[2:9],smoother=FALSE)
```



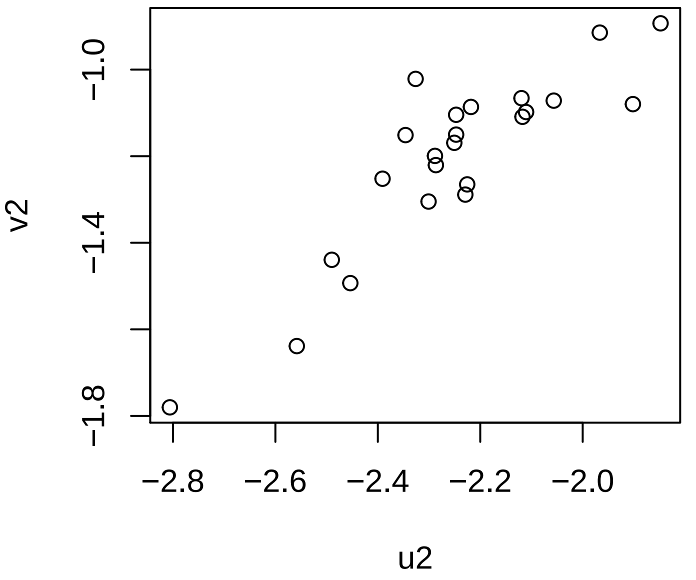
```
cor(carratings[2:9])

##            Economy Service   Value   Price  Design  Sporty  Safety Handling
## Economy     1.0000 -0.3346 -0.4480  0.7583 -0.6187 -0.4761 -0.2838   0.5825
## Service    -0.3346  1.0000  0.9277 -0.7369  0.7581  0.6877  0.9376   0.2974
## Value      -0.4480  0.9277  1.0000 -0.8576  0.8285  0.8015  0.9165   0.1802
## Price       0.7583 -0.7369 -0.8576  1.0000 -0.8874 -0.8558 -0.7140   0.2699
## Design     -0.6187  0.7581  0.8285 -0.8874  1.0000  0.8831  0.7123  -0.2171
## Sporty     -0.4761  0.6877  0.8015 -0.8558  0.8831  1.0000  0.6579  -0.2512
## Safety     -0.2838  0.9376  0.9165 -0.7140  0.7123  0.6579  1.0000   0.3478
## Handling    0.5825  0.2974  0.1802  0.2699 -0.2171 -0.2512  0.3478   1.0000

car_cancor<-cancor(carratings[4:5],carratings[c(2,3,6:9)])
```

```
car_cancor

## $cor
## [1] 0.9792 0.8851
##
## $xcoef
##            [,1]    [,2]
## Value -0.12521 -0.3596
## Price  0.07105 -0.3415
##
## $ycoef
##                [,1]      [,2]      [,3]      [,4]      [,5]    [,6]
## Economy   0.092212 -0.121093 -0.258100 -0.004307 -0.034429  0.3381
## Service  -0.040760 -0.116025  0.286369  0.013213 -0.731048  0.3336
## Design   -0.001013  0.002515 -0.478019 -0.516140 -0.009014 -0.0508
## Sporty   -0.097744  0.020398  0.016551  0.436492 -0.026011 -0.2598
## Safety   -0.047472  0.003025  0.007625 -0.006781  0.549570  0.2064
## Handling -0.080197 -0.195025  0.000854 -0.003304 -0.018228 -0.8114
##
## $xcenter
## Value Price
## 3.152 3.274
##
## $ycenter
##  Economy  Service   Design   Sporty   Safety Handling
##    3.291    3.061    3.200    3.465    3.287    2.787
```

```
u1<-as.matrix(carratings[4:5])%*%car_cancor$xcoef[,1]
v1<-as.matrix(carratings[c(2,3,6:9)])%*%car_cancor$ycoef[,1]
plot(u1,v1)
```



```
u2<-as.matrix(carratings[4:5])%*%car_cancor$xcoef[,2]
v2<-as.matrix(carratings[c(2,3,6:9)])%*%car_cancor$ycoef[,2]
plot(u2,v2)
```



Test whether any canonical correlations are significant

```
n<-nrow(carratings)
px<-2
py<-7
l1<-car_cancor$cor[1]^2
l2<-car_cancor$cor[2]^2
psi0<--(n-1-(px+py+1)/2)*(log(1-l1)+log(1-l2))
print(1-pchisq(psi0,px*py))

## [1] 2.556e-11
```

Test whether the first canonical correlation is significant

```
psi1<--(n-1-(px+py+1)/2)*(log(1-l2))
print(1-pchisq(psi1,(px-1)*(py-1)))

## [1] 0.0002219
```

Exercise: Head Size

In the data set headsize.csv, there are a set of four measurements taken from two brothers:

- x_1: Head length for older brother
- x_2: Head breadth for older brother
- x_3: Head length for younger brother
- x_4: Head breadth for younger brother

Find the two canonical variates for each of the older and younger brothers. What is the correlation between these canonical variates? Are any of the canonical correlations significant? Is the first canonical correlation significant?

Exercise: Glucose

The file glucose.csv contains data from O'Sullivan and Mahon (1966) (data also given in Rencher, 1995), giving measurements on blood glucose for 52 women. The ys represent fasting glucose measurements on three occasions and the xs are glucose measurements one hour after sugar intake. Investigate the relationship between the two sets of variables using canonical correlation analysis. Test whetehr any canonical correlations are significant, and then sequentially test the first two canonical correlations.

Exercises

Johnson and Wichern 10.7, 10.9, 10.10, 10.13.