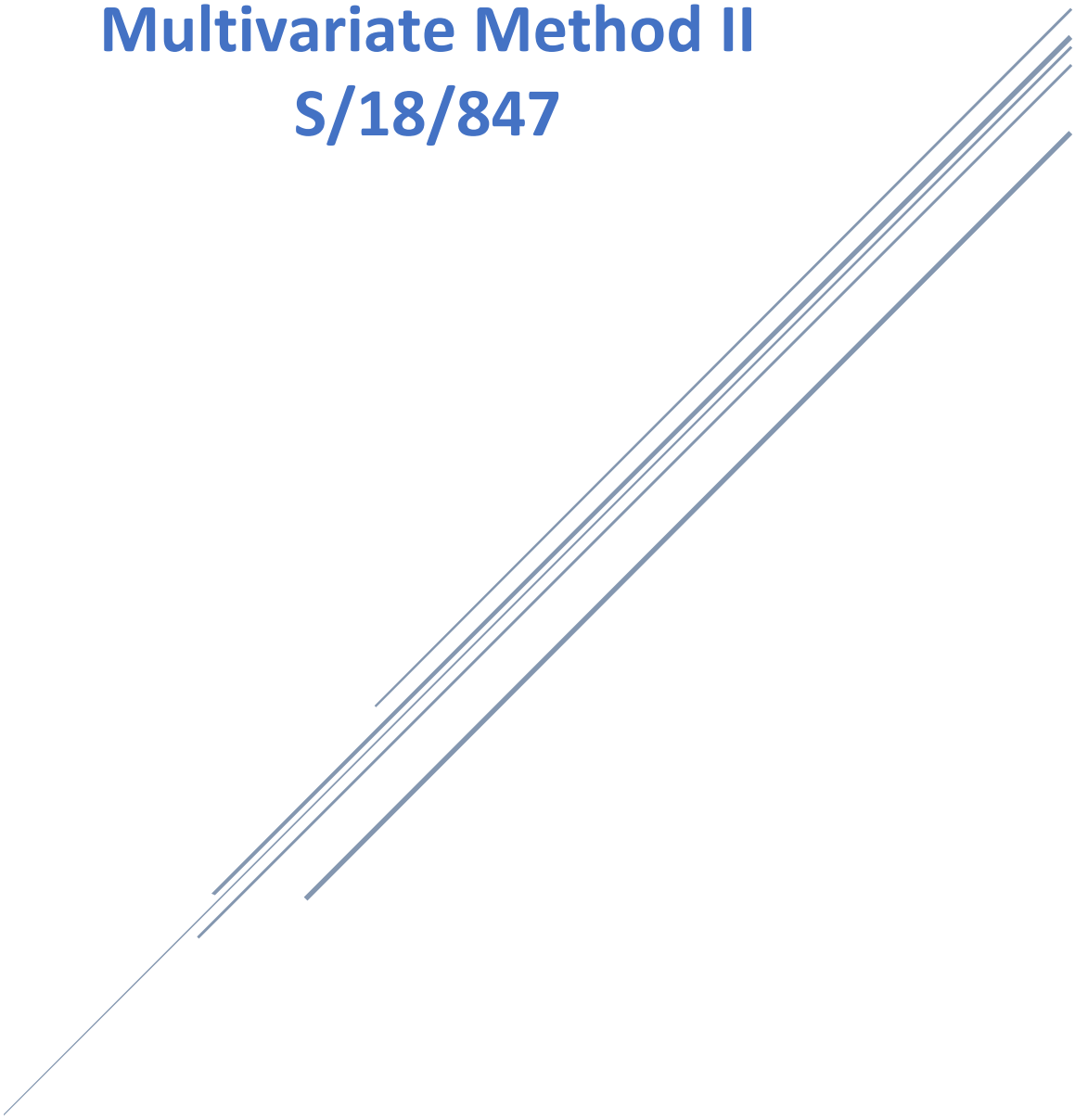
CANONICAL CORRELATION ANALYSIS

Report for Diabetes Data Set

ST405

Multivariate Method II

S/18/847



1. Introduction

Canonical correlation analysis is used to identify and measure the associations among two sets of variables. Canonical correlation is appropriate in the same situations where multiple regression would be, but where there are multiple intercorrelated outcome variables. Canonical correlation analysis determines a set of canonical variates, orthogonal linear combinations of the variables within each set that best explain the variability both within and between sets.

2. Methodology

For the Canonical Correlation analysis, I used "Diabetes Patient Data". This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. In the dataset, all patients here are females at least 21 years old of Pima Indian heritage. The data set contains 768 instances and 9 variables. The data set contains all numeric variables except the last variable "Outcome". It is a binary outcome variable (1-suffer from diabetes,0- do not suffer from diabetes). Including a binary outcome variable in CCA can be problematic because it may not meet the assumption of continuity and normality inherent in CCA. Therefore, it is better to exclude the "Outcome" variable from the dataset used in the CCA

Variable Description:

Pregnancies: To express the number of pregnancies
Glucose: To express the Glucose level in blood
BloodPressure: To express the blood pressure measurement
SkinThickness: To express the thickness of the skin
Insulin: To express the insulin level in blood
BMI: To express the body mass index
DiabetesPedigreeFunction: To express the Diabetes percentage
Age: To express the age
Outcome: To express the final result 1 is Yes and 0 is No

Description of the data

```
  Pregnancies        Glucose      BloodPressure    SkinThickness       Insulin
Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0
1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0
Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5
Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8
3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2
Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0
      BMI        DiabetesPedigreeFunction      Age            Outcome
Min.   : 0.00   Min.   :0.0780           Min.   :21.00   Min.   :0.000
1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00   1st Qu.:0.000
Median :32.00   Median :0.3725           Median :29.00   Median :0.000
Mean   :31.99   Mean   :0.4719           Mean   :33.24   Mean   :0.349
3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00   3rd Qu.:1.000
Max.   :67.10   Max.   :2.4200           Max.   :81.00   Max.   :1.000
```

The methodology involves several key steps. First, we load the necessary R packages required for our analysis. Next, we load the dataset and perform data preprocessing to clean and prepare the data for analysis. After preprocessing, we split the dataset into two separate sets. Both data sets are then standardized to ensure they are on a comparable scale. Finally, we apply Canonical Correlation Analysis to these two standardized data sets to explore the relationships between them.

03. Result and discussion

- Split the dataset into two sets.

Set 1: Biochemical Measurements

- Glucose
- Insulin
- Diabetes Pedigree Function

Set 2: Physiological Measurements

- Pregnancies
- Blood Pressure
- Skin Thickness
- BMI
- Age

Set 1: This contains 3 variables

A tibble: 6 × 3

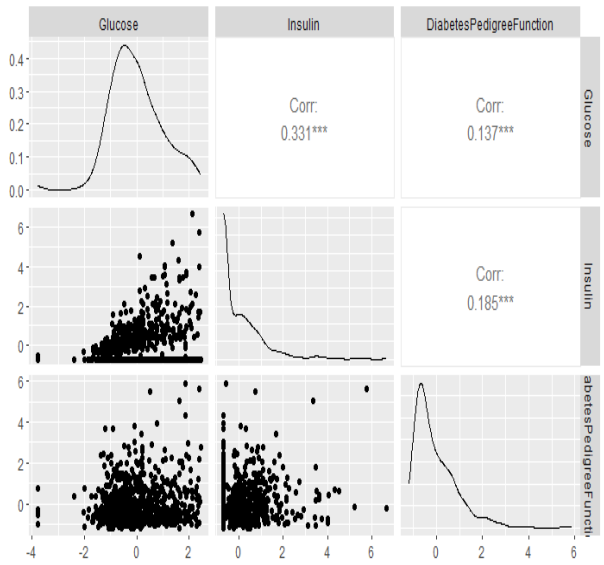
Glucose <dbl>	Insulin <dbl>	DiabetesPedigreeFunction <dbl>
148	0	0.627
85	0	0.351
183	0	0.672
89	94	0.167
137	168	2.288
116	0	0.201

Set 2: This contains 5 variables

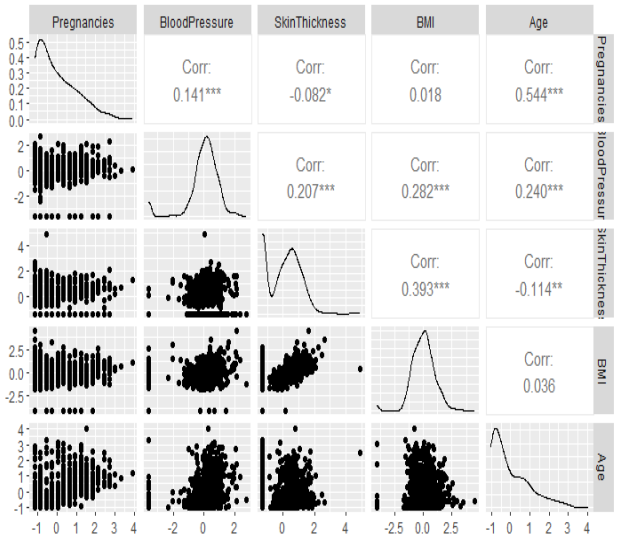
A tibble: 6 × 5

Pregnancies <dbl>	BloodPressure <dbl>	SkinThickness <dbl>	BMI <dbl>	Age <dbl>
6	72	35	33.6	50
1	66	29	26.6	31
8	64	0	23.3	32
1	66	23	28.1	21
0	40	35	43.1	33
5	74	0	25.6	30

3.1 Motivation of canonical correlation analysis



Pairwise scatterplot plots with variables in set 1



Pairwise scatterplot plots with variables in set 2

- The dimensions of those two sets are large, and the problem of interpretation arises. Canonical Correlation Analysis allows us to summarize the relationships into fewer statistics while preserving the main facts of the relationships. In a way, the motivation for canonical correlations is very similar to principal component analysis. It is another dimension-reduction technique.

3.2 Significance Tests for Canonical Correlations

This is carried out using **Wilk's lambda.**

```
Test of H0: The canonical correlations in the
current row and all that follow are zero

     CanR LR test stat approx F numDF  denDF Pr(> F)
1 0.46856      0.68431  20.6069    15 2098.4  <2e-16 ***
2 0.34813      0.87682  12.9248     8 1522.0  <2e-16 ***
3 0.04755      0.99774   0.5755     3  762.0  0.6312
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis:

H_0: All canonical variate pairs are uncorrelated. ($\rho 1 = \rho 2 = \rho 3 = 0$)
H_1: At least one $\rho i \neq 0$; I = 1,2,3

Here you can see first two canonical variate pairs are significantly correlated and depend on one another. The third canonical variate pair is not significantly correlated at a 1% significance level. This suggests that we would want to go ahead and summarize for only two canonical variate pairs. Each significant canonical correlation can be further tested using other multivariate tests such as Pillai's test, Lawley-Hotelling test, and Roy's largest root test.

3.3 Estimates of Canonical Correlation

By fitting the canonical correlation model, we can obtain 3 canonical variate pairs. (Because a small set (Set 1) has 3 variables)

```
cc_model <- cc(Biochemical_Measurements,Physiological_Measurements)

#Display the Canonical correlations
cc_model$cor
```

```
[1] 0.4685599 0.3481299 0.0475472
```

Here you can see canonical correlations of three canonical variate pairs.

After the significance test, we identified only two significant canonical variate pairs.

```
# Significant canonical correlations

cc_model$cor[1:2]

#Squared canonical correlations
cc_model$cor[1:2]^2
```

```
[1] 0.4685599 0.3481299
[1] 0.2195484 0.1211944
```

From the squared canonical correlation, we can conclude that 21.95% of the variation in the first canonical variate of the "Biochemical Measurements" set (U1) is explained by the variation in the first canonical variate of the "Physiological Measurements" set (V1) and 12.12% of the variation in the second canonical variate of the "Biochemical Measurements" set (U2) is explained by second canonical variate of the "Physiological Measurements" set (V2).

3.4 Canonical Coefficients

- **The estimated canonical coefficients for the "Biochemical Measurements" variables**

	[,1]	[,2]
Glucose	-0.3590690	-0.97702434
Insulin	0.9924974	0.08993511
DiabetesPedigreeFunction	0.2401284	-0.21319209

U_1 = -0.3591(Glucose) + 0.9925(Insulin) + 0.2401(DiabetesPedigreeFunction)

U_2 = -0.9770(Glucose) + 0.0899(Insulin) - 0.2132(DiabetesPedigreeFunction)

The magnitudes of the coefficients give the contribution of the individual variables to the corresponding canonical variable. "Insulin" contributes best to the first canonical variate of the "Biochemical Measurements" variable set. "Glucose" contributes the highest to the second canonical variate of the "Biochemical Measurements" variable set.

- **The estimated canonical coefficients for the "Physiological Measurements" variables**

	[,1]	[,2]
Pregnancies	-0.14638216	0.08232762
BloodPressure	-0.06361264	-0.08899644
SkinThickness	0.98596971	0.01625444
BMI	-0.04241753	-0.61028316
Age	-0.06521955	-0.77052306

V_1 = -0.1464(Pregnancies) -0.0636(BloodPressure) + 0.9860(SkinThickness) -0.0424(BMI) – 0.0652(Age)

V_2 = 0.0823(Pregnancies) -0.0890(BloodPressure) + 0.0163(SkinThickness) -0.6103(BMI) – 0.7705(Age)

"SkinThickness" contributes best to the first canonical variate of the "Physiological Measurements" variable set. "Age" contributes the highest to the second canonical variate of the "Physiological Measurements" variable set.

3.5 The correlation between the "Biochemical Measurements" variables and the canonical variables of the "Biochemical Measurements" set

- This value represents the correlation between the "Biochemical Measurements" variables and their corresponding canonical variates. These correlations help us understand how each original variable contributes to the canonical variates

```
                            [,1]       [,2]
Glucose                 0.0027807 -0.9765029
Insulin                 0.9179581 -0.2732645
DiabetesPedigreeFunction 0.3744973 -0.3307296
```

Only the "Insulin" variable has a high correlation with the first canonical variate(U1) of the "Biochemical Measurements" set. It is positively correlated with the first canonical variable. The other two correlations of the first canonical variable are small. Therefore, it means that the "Insulin" variable plays a major role in defining the first canonical relationship between the data set.

Only the "Glucose" variable has a high correlation with the second canonical variate (U2) of the "Biochemical Measurements" set. It is negatively correlated with the second canonical variable. The other two correlations of the second canonical variable are small. Therefore, it means that the "Glucose" variable plays a major role in defining the second canonical relationship between the data set.

Therefore, the best predictors of the "Biochemical Measurements" set are "Insulin" and "Glucose" as these indicators stand out most.

3.6 The correlation between the "Physiological Measurements" variables and the canonical variables of the "Physiological Measurements" set

```
                  [,1]       [,2]
Pregnancies   -0.27214714 -0.3617927
BloodPressure  0.09259188 -0.4305372
SkinThickness  0.97551472 -0.1606887
BMI            0.32176922 -0.6554512
Age           -0.27404692 -0.7709962
```

- This value represents the correlation between the "Physiological Measurements" variables and their corresponding canonical variates

Only the "SkinThickness" variable has a high correlation with the first canonical variate(V1) of the "Physiological Measurements" set. It is positively correlated with the first canonical variable. The other correlations of the first canonical variable are small. Therefore, it means that the "SkinThickness" variable plays a major role in defining the first canonical relationship between the data set.

The “Age” and “BMI” variables have a high correlation with the second canonical variate(V2) of the “Physiological Measurements” set. Both are negatively correlated with the second canonical variable. The “Blood Pressure” variable is moderately and negatively correlated also. The other correlations of the second canonical variable are small. Therefore, it means that the “Age” and “BMI” variables play a major role in defining the second canonical relationship between the data set.

Therefore, the best predictors of the “Physiological Measurements” set are “SkinThickness”, “Age” and “BMI” as these indicators stand out most.

- **Reinforcing the Results**

3.7 The correlation between the “Biochemical Measurement” variables and the canonical variates of the “Physiological Measurements” set.

```
                            [,1]        [,2]
Glucose                0.001302925 -0.33994988
Insulin                0.430118410 -0.09513155
DiabetesPedigreeFunction 0.175474415 -0.11513687
```

The correlation between “Insulin” and the first canonical variate of the “Physiological Measurements” set is moderately positive. This means the “Insulin” variable plays a notable role in defining the first canonical relation between two data sets. The correlation between “Glucose” and “DiabetesPedigreeFunction” variables and the first canonical variate of the “Physiological Measurements” set is very low and those give a negligible contribution to the first canonical variate of the “Physiological Measurements” set.

The correlation between the “Glucose” variable and the second canonical variate of the “Physiological Measurements” set is moderate and negative. This means it influences the second canonical relationship between two data sets. The “Insulin” and “DiabetesPedigreeFunction” variables have a negligible contribution to the second canonical variate of the “Physiological Measurements” set.

3.8 The correlation between the “Physiological Measurements” variables and the canonical variates of the “Biochemical Measurement” set.

```
                  [,1]        [,2]
Pregnancies   -0.12751725 -0.12595085
BloodPressure  0.04338485 -0.14988287
SkinThickness  0.45708713 -0.05594056
BMI            0.15076817 -0.22818216
Age           -0.12840741 -0.26840684
```

The correlation between “SkinThickness” and the first canonical variate from the “Biochemical Measurements” set is moderate and positive. This suggests that “SkinThickness” has a notable positive contribution to the first canonical variate, meaning it plays a significant role in defining the first canonical relationship between two data sets. Other variables “Pregnancies”, “Bloodpressure”, “BMI” and “Age” of the “Physiological Measurements” set do not significantly influence the first canonical relationship between the two datasets.

The correlation between “Age”, “BMI” and “BloodPressure” variables and the second canonical variate from the “Biochemical Measurements” set is low to moderate and negative. It means that it plays a somewhat significant inverse role in defining the second canonical relationship between two sets. The relationship between the “Pregnancies” and “SkinThickness” variables and the second canonical variable can be negotiable.

04. Conclusion and Recommendation

- In our canonical correlation analysis, we examined two sets of variables: “Biochemical Measurements” (3 Variables) and “Physiological Measurements” (5 Variables). We generated three canonical variate pairs, two of which we found significant using Wilks’lambda test. Therefore, for further analysis, we focused on these two significant canonical variate pairs.
- The first canonical correlation (0.4686) indicates a moderate linear relationship between the first pairs of canonical variates f the two datasets. The second canonical correlation (0.3481) indicates a weaker linear relationship between second pairs of canonical variates. The third canonical correlation (0.0475) is very low and not significant, indicating a negligible relationship for this pair.
- From the squared canonical correlation, we can conclude that 21.95% of the variation in the first canonical variate of the “Biochemical Measurements” set (U1) is explained by the variation in the first canonical variate of the “Physiological Measurements” set (V1) and 12.12% of the variation in the second canonical variate of the “Biochemical Measurements” set (U2) is explained by second canonical variate of the “Physiological Measurements” set (V2).
- “Insulin” contributes best to the first canonical variate of the “Biochemical Measurements” variable set. “Glucose” contributes the highest to the second canonical variate of the “Biochemical Measurements” variable set. “SkinThickness” contributes best to the first canonical variate of the “Physiological Measurements” variable set. “Age” contributes the highest to the second canonical variate of the “Physiological Measurements” variable set.
- The best predictors of the “Biochemical Measurements” set are “Insulin” and “Glucose” as these indicators stand out most. The best predictors of the “Physiological Measurements” set are “SkinThickness”, “Age” and “BMI” as these indicators stand out most.

05. References

- **The data set is taken from:**

https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

- https://stats.oarc.ucla.edu/r/dae/canonical-correlation-analysis/
- https://www.youtube.com/watch?v=oDzUAvwquxQ
- Rencher, A.C., & Christensen, W.F. (2012). Methods of Multivariate analysis (3rd ed.). John Wiley & Sons
- https://online.stat.psu.edu/stat505/book/export/html/682

06. Appendices

- **Part of the dataset**

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31.0	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0.0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38.0	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1
16	7	100	0	0	0	30.0	0.484	32	1
17	0	118	84	47	230	45.8	0.551	31	1
18	7	107	74	0	0	29.6	0.254	31	1
19	1	103	30	38	83	43.3	0.183	33	0

❖ **Codes**

```r
#Load required packages
library(tidyverse)
library(CCA)
library(CCP)
library(candisc)
library(skimr)
library(ggplot2)
library(GGally)

#Import the data set
Diabetes <- read_csv(file = "../Data/diabetes.csv")
head(Diabetes)
view(Diabetes)
summary(Diabetes)
any(is.na(Diabetes))

#Remove "Outcome" variable
Diabetes <- Diabetes[, -ncol(Diabetes)]
head(Diabetes)

#Split the variables into two sets
Biochemical_Measurements <- Diabetes[,c("Glucose","Insulin","DiabetesPedigreeFunction")]
head(Biochemical_Measurements)

Physiological_Measurements <- Diabetes[,c("Pregnancies", "BloodPressure", "SkinThickness", "BMI", "Age")]
head(Physiological_Measurements)

#Standardized the data sets
Biochemical_Measurements <- apply(Biochemical_Measurements,2,scale)
Physiological_Measurements <- apply(Physiological_Measurements,2,scale)

#Correlation between the variables of set "Biochemical_ Measurements"
ggpairs(Biochemical_Measurements)

#Correlation between the variables of set "Physiological_ Measurements"
ggpairs(Physiological_Measurements)

#Correlation within the set and between two set
matcor(Biochemical_Measurements,Physiological_Measurements)

#Test whether the canonical correlations are significant or not(Wilks' Lambda test)
Wilks(cancor(Biochemical_Measurements,Physiological_Measurements))

#tests of canonical dimensions
rho <- cc_model$cor

#Define number of observations,number of variables in first set, and number of variables
#in the second set
n <- dim(Biochemical_Measurements)[1]
p <- dim(Biochemical_Measurements)[2]
q <- dim(Physiological_Measurements)[2]

#Calculate p-values using the F-Approximations of the different test statistics:
p.asym(rho,n,p,q,tstat = "Wilks")
p.asym(rho,n,p,q,tstat = "Hotelling")
p.asym(rho,n,p,q,tstat = "Pillai")
p.asym(rho,n,p,q,tstat = "Roy")

cc_model <- cc(Biochemical_Measurements,Physiological_Measurements)

#Display the Canonical correlations
cc_model$cor

# Significant canonical correlations
cc_model$cor[1:2]

#Squared canonical correlations
cc_model$cor[1:2]^2
```

```r
#Canonical coefficients
cc_model$xcoef[,1:2]
cc_model$ycoef[,1:2]

#Compute canonical loadings
loadings <- comput(Biochemical_Measurements,Physiological_Measurements, cc_model)

#Display canonical loadings
loadings$corr.X.xscores[,1:2]
loadings$corr.Y.yscores[,1:2]
loadings$corr.X.yscores[,1:2]
loadings$corr.Y.xscores[,1:2]
```