IS507 ASSIGNMENT 6

1. Choose a technique that we have covered so far in this course, and try applying that technique to your data. You may choose any of

a) Model building and Multiple Regression
b) PCA
c) CFA
d) CCA
e) CA (correspondence analysis)

If you are working as a group, each member of your group should try a different technique, or the same technique with different aspects of the data

Answer:

The dataset chosen (part of my group #3) is the Breast Cancer Diagnostics Dataset. Since there are missing values, I replaced NA with 0's and removed 569 missing entries.
I performed CCA between the mean of variables and worst or largest (mean of the three largest values) of each variable.
CCA results in 10 canonical variates, and on performing F-test (Rao's Approximation) we can see that all 10 variates are statistically significant, hence we can reject null hypothesis (CC's are equal to zero).

```
> F.test.cca(c2)

        F Test for Canonical Correlations (Rao's F Approximation)

           Corr         F    Num df Den df    Pr(>F)
CV 1    0.98642 212.79441 100.00000 3942.5 < 2.2e-16 ***
CV 2    0.93368 152.58496  81.00000 3563.4 < 2.2e-16 ***
CV 3    0.90744 136.74453  64.00000 3184.6 < 2.2e-16 ***
CV 4    0.87696 125.35633  49.00000 2806.8 < 2.2e-16 ***
CV 5    0.83835 117.28844  36.00000 2431.2 < 2.2e-16 ***
CV 6    0.78872 112.97388  25.00000 2059.5 < 2.2e-16 ***
CV 7    0.72968 114.28559  16.00000 1696.2 < 2.2e-16 ***
CV 8    0.67413 125.21068   9.00000 1353.3 < 2.2e-16 ***
CV 9    0.61080 151.41995   4.00000 1114.0 < 2.2e-16 ***
CV 10   0.57501 275.62556   1.00000  558.0 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

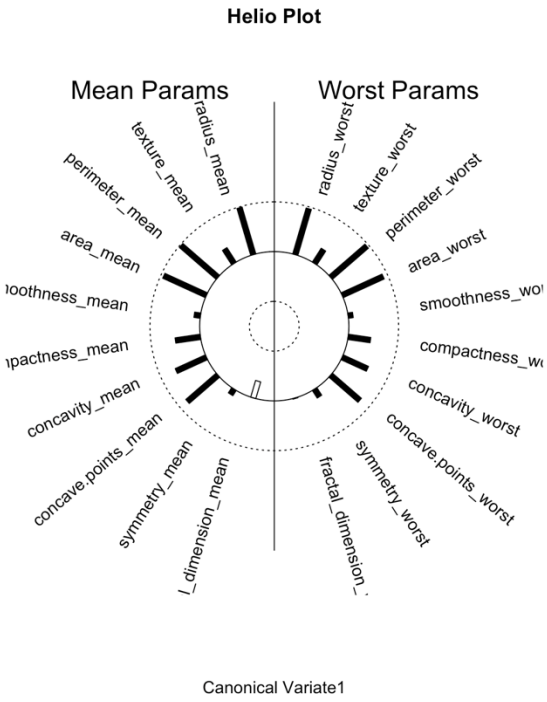
Correlation Values:

```
> c2

Canonical Correlation Analysis

Canonical Correlations:
     CV 1      CV 2      CV 3      CV 4      CV 5      CV 6      CV 7      CV 8      CV 9     CV 10
0.9864218 0.9336817 0.9074421 0.8769586 0.8383521 0.7887221 0.7296815 0.6741322 0.6108029 0.5750085
```

Helio Plot for CV1:



2. Paper Review: An academic paper from a conference or Journal will be posted to the Moodle. It contains a usage of Canonical Correlation Analysis. Review the paper and evaluate their usage of Canonical Correlation Analysis. In particular, address (The Association of Work Satisfaction and Burnout-CCA)

a. How suitable is the data for CC?

Answer: The research question they're trying to answer is the relationship or association between Burnout Symptoms and Work Satisfaction in endoscopic nursing staff in Germany. They used canonical correlation (CC) to determine the relation between burnout and work satisfaction. Since this is a question of association, the data used is appropriate for CC. Data collected contained 674 samples from endoscopy nursing staff (579 female and 95 male). Burnout scores were calculated using the Maslach burnout inventory (MBI) questionnaire.

b. How are they applying CC? What two groups are being correlated? Are they metric, ordinal, nominal?

Answer: They're using CC to get the relationship between Burnout and Work Satisfaction. The two groups being correlated are KAFA (general and facet-specific job satisfaction) and MBI (Maslach burnout inventory). Both of them have several sub-scales, all of which are metric.

c. What methods do they use to judge the quality of the correlation? Do they evaluate, and how do they evaluate the stability of the components?

Answer: To judge the quality of the correlation, they use statistical significance and variance of the functions. They got 3 functions but eliminated function 3 because they say it is significant only due to the large sample size.
Stability of components is discussed, and is measured using Cronbach's alpha, omega coefficient and average inter-item correlation. Other factors such as mean, median, min, max, deviation from normal distribution (Shapiro-Wilks test), skewness and kurtosis are also discussed for each scale.

d. How many correlates do they concentrate on in their analysis, and do they attempt to interpret the correlates in terms of the original variables?

Answer: The canonical correlation analysis (CCA) resulted in 3 canonical functions, with canonical correlations of 0.64, 0.32 and 0.17, all of which were statistically significant. They focus only on 2 out of the 3 functions because function 3 was only statistically significant due to the large sample size and a canonical value of 0.17 was less.
Yes, they do interpret the correlates in terms of the original variables. For example, it is explained that in function 1, the squared canonical correlation was 0.415 and the general job satisfaction scale contributed 0.069 (16.6% of the squared canonical correlation) to this squared coefficient of the burnout canonical variate – the original variables being job satisfaction and burnout.

e. What conclusions does CC allow them to draw?

Answer: using CC, they were able to say that there does exist some association between job satisfaction and burnout for endoscopic nursing staff, which could be helpful to decide the specific needs of employees. Also, they found some positive correlation between job satisfaction and work engagement, which means that if satisfaction was improved it could result in better patient care in the field of gastroenterology.

3. CCA in R: Perform the following Canonical Correlation Analysis on the Young People Survey from R PCA and FA Lab - Young People Survey. Perform a canonical correlation analysis describing the relationships between the music and spending variables using the data under the R PCA and FA Lab - Young People Survey in the content folder).

a. Test the null hypothesis that the canonical correlations are all equal to zero. Give your test statistic, d.f and p-value

Answer:

H0 – canonical correlations are all equal to zero between music and spending
HA – canonical correlations are NOT equal to zero between music and spending

Using F Test for Canonical Correlations (Rao's Approximation), the following values were seen as output:

```
> F.test.cca(c2)

        F Test for Canonical Correlations (Rao's F Approximation)

          Corr         F    Num df Den df    Pr(>F)
CV 1   0.46216   3.01086 133.00000 4371.8 < 2.2e-16 ***
CV 2   0.34532   2.14054 108.00000 3795.4 1.576e-10 ***
CV 3   0.27807   1.68923  85.00000 3204.6 9.906e-05 ***
CV 4   0.20622   1.38563  64.00000 2597.8   0.02405 *
CV 5   0.19818   1.31769  45.00000 1973.4   0.07792 .
CV 6   0.16573   1.15100  28.00000 1330.0   0.26829
CV 7   0.14084   1.03676  13.00000  666.0   0.41345
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test statistic here is F, degrees of freedom (d.f) and p-values are reported. From the p-values we can see that only the first 4 canonical variates are statistically significant against an alpha of 0.05. Hence, we can reject null hypothesis for CV 1-4.

Code:

```
library(yacca)
c2 = cca(music, spending)
c2

F.test.cca(c2)
```

b. How many significant canonical variates are there?

Answer: using chi-squared and F test for canonical correlations, we can see that only the first 4 canonical variates are significant as they have p-values < 0.05. Hence, CV 1, CV 2, CV 3 and CV 4 are significant.

Code:

```
library(yacca)
c2 = cca(music, spending)
c2

F.test.cca(c2)

#CV1
helio.plot(c2, cv=1, x.name="Music Values",
           y.name="Spending Values")

#CV2
helio.plot(c2, cv=2, x.name="Music Values",
           y.name="Spending Values")

#Function Names
ls(c2)

# Perform a chi-square test on C2
c2
ls(c2)
c2$chisq
c2$df
summary(c2)
round(pchisq(c2$chisq, c2$df, lower.tail=F), 3)
```

Chi-squared output:

```
> round(pchisq(c2$chisq, c2$df, lower.tail=F), 3)
 CV 1  CV 2  CV 3  CV 4  CV 5  CV 6  CV 7
0.000 0.000 0.000 0.024 0.078 0.268 0.413
>
```

c. Present the first two canonical correlations (cancor)?

Answer:

Code:

```
#Base Package
library(foreign)
library(CCA)
library(yacca)
library(MASS)
# This gives us the canonical correlates, but no significance tests
c = cancor(music, spending)
c
```

Output:

```
> c$cor
[1] 0.4621615 0.3453192 0.2780697 0.2062196 0.1981754 0.1657295 0.1408390
>
```

First two canonical correlations between music and spending: 0.4621615 and 0.3453192

d. What do you conclude from your analysis?

Answer:

From the above analysis, we can see that only the first 4 CVs are statistically significant. The correlation coefficients for each of the 4 CVs are not too high – it ranges from 0.46 to 0.20.

2. Answer the following questions regarding the canonical variates.

a. Give the formulae for the first canonical variate for the music and spending variables.

Answer:

```
#Breakdown of the Correlations
matcor(music, spending)

#Correlations between music and spending (Y)
cc_mm = cc(music, spending)
cc_mm$cor

#Funcrions for CCA
ls(cc_mm)

#XCoef Correlations
cc_mm$xcoef

#YCoef Correlations
cc_mm$ycoef
```

For spending:

$U1 = 0.37476949\ X_{Music} - 0.11590314\ X_{Slow.songs.or.fast.songs} + 0.08082942\ X_{Dance} - 0.05909564\ X_{Folk} - 0.01479423\ X_{Country} - 0.14498741\ X_{Classical.music} + 0.22436097\ X_{Musical} + 0.36687318\ X_{Pop} - 0.09733734\ X_{Rock} - 0.28268565\ X_{Metal.or.Hardrock} + 0.20224133\ X_{Punk} + 0.05800900\ X_{Hiphop..Rap} + -0.24815254\ X_{Reggae..Ska} + 0.01852198\ X_{Swing..Jazz} - 0.09779556\ X_{Rock.n.roll} - 0.08699481\ X_{Alternative} + 0.11950017\ X_{Latino} - 0.01727741\ X_{Techno..Trance} - 0.03241265\ X_{Opera}$

For Music:

$V1 = 0.095585476\ Y_{Finances} + 0.499907293\ Y_{Shopping.centres} - 0.006318212\ Y_{Branded.clothing} - 0.334534685\ Y_{Entertainment.spending} + 0.458637542\ Y_{Spending.on.looks} + 0.017301504\ Y_{Spending.on.gadgets} - 0.107647531\ Y_{Spending.on.healthy.eating}$

Output:

```
> cc_mm$xcoef
                              [,1]
Music                      0.37476949
Slow.songs.or.fast.songs -0.11590314
Dance                      0.08082942
Folk                      -0.05909564
Country                   -0.01479423
Classical.music           -0.14498741
Musical                    0.22436097
Pop                        0.36687318
Rock                      -0.09733734
Metal.or.Hardrock         -0.28268565
Punk                       0.20224133
Hiphop..Rap                0.05800900
Reggae..Ska               -0.24815254
Swing..Jazz                0.01852198
Rock.n.roll               -0.09779556
Alternative               -0.08699481
Latino                     0.11950017
Techno..Trance            -0.01727741
Opera                     -0.03241265
```

```
                                  [,1]
Finances                     0.095585476
Shopping.centres             0.499907293
Branded.clothing            -0.006318212
Entertainment.spending      -0.334534685
Spending.on.looks            0.458637542
Spending.on.gadgets          0.017301504
Spending.on.healthy.eating  -0.107647531
```

b. Give the correlations between first canonical variate for the music and spending variables.

Answer:

Code:

```
#Calculate Scores
loadings_mm = comput(music, spending, cc_mm)
ls(loadings_mm)

#Correlation X Scores
loadings_mm$corr.X.xscores

#Correlation Y Scores
loadings_mm$corr.Y.yscores

# A basic visualization of the canonical correlation
plt.cc(cc_mm)
```

Correlations between Music and canonical variate 1:

```
                              [,1]
Music                      0.21887538
Slow.songs.or.fast.songs -0.05544365
Dance                      0.48234591
Folk                      -0.10829069
Country                   -0.10460136
Classical.music           -0.24095739
Musical                    0.32409486
Pop                        0.72001733
Rock                      -0.36682287
Metal.or.Hardrock         -0.64973432
Punk                      -0.29381699
Hiphop..Rap                0.33768822
Reggae..Ska               -0.21347677
Swing..Jazz               -0.17948911
Rock.n.roll               -0.28412069
Alternative               -0.41896746
Latino                     0.34215227
Techno..Trance             0.12351345
Opera                     -0.13250016
```

Correlations between Spending and canonical variate 1:

	[,1]
Finances	0.121736658
Shopping.centres	0.856359153
Branded.clothing	0.282831849
Entertainment.spending	-0.155593524
Spending.on.looks	0.661313542
Spending.on.gadgets	0.097816064
Spending.on.healthy.eating	0.001913928

Code for 2(a) and 2(b):

```
#Breakdown of the Correlations
matcor(music, spending)

#Correlations between music and spending (Y)
cc_mm = cc(music, spending)
cc_mm$cor

#Funcrions for CCA
ls(cc_mm)

#XCoef Correlations
cc_mm$xcoef

#YCoef Correlations
cc_mm$ycoef

#Calculate Scores
loadings_mm = comput(music, spending, cc_mm)
ls(loadings_mm)

#Correlation X Scores
loadings_mm$corr.X.xscores

#Correlation Y Scores
loadings_mm$corr.Y.yscores

# A basic visualization of the canonical correlation
plt.cc(cc_mm)
```

c. What can you conclude from the above analyses?

Answer: We can see that for CV1, the value of the coefficients act as a weight to each of the individual variables in the canonical variable.

We can see from our significant canonical variates and their correlation coefficients with:

Music: none of the correlation coefficients are particularly large for any of the 4 variates, and hence does not yield any useful information about the data

Spending: Similar to music, none of the correlation coefficients are particularly large for any of the 4 variates, and hence does not yield any useful information about the data