Homework3 Answer Key

1. (15 pts) Ex. 3.3 Find the principal components of the following correlation matrix given by MacDonnell (1902) from measurements of seven physical characteristics in each of 3 000 convicted criminals:

$$R = \begin{array}{l} \text{Head length} \\ \text{Head breadth} \\ \text{Face breadth} \\ \text{Left finger length} \\ \text{Left forearm length} \\ \text{Left foot length} \\ \text{Height} \end{array} \begin{pmatrix} 1.000 & & & & & & \\ 0.402 & 1.000 & & & & & \\ 0.396 & 0.618 & 1.000 & & & & \\ 0.301 & 0.150 & 0.321 & 1.000 & & & \\ 0.305 & 0.135 & 0.289 & 0.846 & 1.000 & & \\ 0.339 & 0.206 & 0.363 & 0.759 & 0.797 & 1.000 & \\ 0.340 & 0.183 & 0.345 & 0.661 & 0.800 & 0.736 & 1.000 \end{pmatrix}$$

How would you interpret the derived components?

- **Compute the eigenvalues and eigenvectors of correlation matrix R**

```
library(MVA)

R<-as.matrix(read.csv("D:/STAT 4400/Data/characteristics.csv", header=FALSE));R

eigen(R)

$values
[1] 3.7994745 1.5023283 0.6498074 0.3600569 0.3391625 0.2352531 0.1139173

$vectors
     [,1]       [,2]        [,3]        [,4]        [,5]         [,6]        [,7]
[1,] -0.2763037 -0.3647677  0.882274766 -0.08573946 -0.06740350  0.005384671 -0.01638732
[2,] -0.2118636 -0.6392041 -0.257527788  0.68707351  0.08129399  0.034955657  0.01762744
[3,] -0.2951449 -0.5123928 -0.381447691 -0.69856220 -0.10071831  0.033740772 -0.07462604
[4,] -0.4375581  0.2349399 -0.069924234  0.10160027 -0.61923662  0.318242311  0.50339046
[5,] -0.4557045  0.2766674 -0.036669136  0.11311530 -0.03907675  0.290305975 -0.78475748
[6,] -0.4502341  0.1784374 -0.059124621  0.05299938 -0.03440885 -0.870489463  0.01445146
[7,] -0.4356893  0.1795404 -0.006212105 -0.08162701  0.76976550  0.233030117  0.35269900
```

- **Compute the variances of PCs and the proportions of the principal components account for total variations of the original data in the PCA.**

$$P_1 = \frac{\lambda_1}{7} = \frac{3.7994745}{7} = .54$$

$$P_2 = \frac{\lambda_1 + \lambda_2}{7} = \frac{3.7994745 + 1.5023283}{7} = .76$$

$$P_3 = \frac{\lambda_1 + \lambda_2 + \lambda_3}{7} = \frac{3.7994745 + 1.5023283 + 0.6498074}{7} = .85$$

The three PCs account for 85% of variation of the original data. Therefore the three PCs are sufficient statistics.

- **Identify the three PCs**

$$Y_1 = -0.2763z_1 - .2119z_2 - 0.29519z_3 - 0.4376z_4 - 0.4557z_5 - 0.4502z_6 - 0.4357z_7$$

$$Y_2 = -0.3648z_1 - 0.63928z_2 - 0.5124z_3 + 0.235z_4 - 0.2767 + 0.1784z_6 + 0.1795z_7$$

$$Y_3 = 0.8823z_1 - 0.2575z_2 - 0.3814z_3 - 0.0699z_4 - 0.0367z_5 - 0.0591z_6 - 0.0062z_7$$

- **Interpretation of the three PCs**

$$Y_1 = -0.2763z_1 - .2119z\_2 - 0.29519z_3 - 0.4376z_4 - 0.4557z_5 - 0.4502z_6 - 0.4357z_7$$

- The first PC has large coefficients in negative values for height, foot length, forearm length, and finger length. Someone with a high score for PC1 would probably be short with short feet, arms, and fingers. Perhaps we can label the first component "shortness" index for height, feet, arms, and fingers. The first PC is a dominant variable in the PCA, which accounts for 54 % variation of original data.

$$Y_2 = -0.3648z_1 - 0.63928z_2 - 0.5124z_3 + 0.235z_4 - 0.2767 + 0.1784z_6 + 0.1795z_7$$

- The second PC has high negative loadings on head breadth and face breadth. Someone with a high score for PC2 would probably have a narrow face and head. Perhaps we can label the second component a "narrowness of head" index. The first PC is a moderate variable in the PCA, which accounts for 22 % variation of original data.

$$Y_3 = 0.8823z_1 - 0.2575z_2 - 0.3814z_3 - 0.0699z_4 - 0.0367z_5 - 0.0591z_6 - 0.0062z_7$$

- The third PC has a high positive coefficient for head length. Someone with a high score for PC3 would probably have a long head. We may label this is a "head length" index.

2. Ex. 3.4 Not all canonical correlations may be statistically significant. An approximate test proposed by Bartlett (1947) can be used to determine how many significant relationships exist. The test statistic for testing that at least one canonical correlation is significant is $\emptyset_0^2 = -\left\{n - \frac{1}{2}(q_1 + q_2 + 1)\right\}\sum_{i=1}^{s}\log(1-\lambda_i)$ where the λ_i are the eigenvalues of E1 and E2. Under the null hypothesis that all correlations are zero, $\emptyset_0^2$ has a chi-square distribution with $q1 \times q2$ degrees of freedom. Write R code to apply this test to the headsize data (Table 3.1) and the depression data (Table 3.3).

   a. Test to the headsize data

```
> (e2<-eigen(E2))
$values
[1] 0.621744734 0.002887956

$vectors
           [,1]       [,2]
[1,] -0.6837994 -0.7091095
[2,] -0.7296700  0.7050984

> n<-25
> q1<-2
> q2<-2
> k<-log(1-e1$values)
> m<-sum(k)
> phi2<--{n-0.5*(q1+q2+1)}*m
> (pvlaue<-1-pchisq(phi2,q1*q2))
[1] 0.0002060779
> phi2
[1] 21.93926
```

$H_0: \rho_1 = \rho_2 = 0, \quad v.s. \quad H_a:$ At least one of ρ is not zero

$\emptyset_0^2 = 21.93, df = 4$, which is chi-square distributed with 4 degrees of freedom. We reject null hypothesis if

$$\emptyset_0^2 > \chi_4(0.05)$$

$p - value = 0.00020 < 0.05$ Reject null hypothesis.

Conclusion: at least the correlation of the first canonical variates is significantly large then zero. For headsize data the $corr(\hat{u}_1, \hat{v}_1) = sqrt(0.6217)$

```
> (e2<-eigen(E2))
$values
[1] 0.621744734 0.002887956

$vectors
           [,1]       [,2]
[1,] -0.6837994 -0.7091095
[2,] -0.7296700  0.7050984

> n<-25
> q1<-2
> q2<-2
> k<-log(1-e1$values)
> phi2<--{n-0.5*(q1+q2+1)}*k[2,]
> (pvlaue<-1-pchisq(phi2,(q1-1)*(q2-1))
```

$H_0: \rho_1 \neq 0, \quad \rho_2 = 0, \quad v.s. \quad H_a: \rho_2 \neq 0$

We fail to reject null hypothesis because $p - value >$ 0.05.

b. Test to the depression data

```
library("MVA")
headsize<-read.csv("D:/STAT 4400/Data/data.csv", header=TRUE)
headsize<-as.matrix(headsize)
headsize.std<-sweep(headsize, 2, apply(headsize, 2, sd), FUN="/")
R<-cor(headsize.std)
r11<-R[1:2,1:2];
r12<-R[1:2,-(1:2)]
r21<-R[-(1:2),1:2]
r22<-R[-(1:2),-(1:2)]
(E1<-solve(r11)%*%r12%*%solve(r22)%*%r21)
(E2<-solve(r22)%*%r21%*%solve(r11)%*%r12)
(e1<-eigen(E1))
n<-294
q1<-2
q2<-4
k<-log(1-e1$values)
m<-sum(k)
phi2<--{n-0.5*(q1+q2+1)}*m
(pvlaue<-pchisq((phi2,q1*q2, lower.tail=f))
```

$H_0: \rho_1 = \rho_2 = 0,\ v.s.$ $H_a:$ At least one of ρ is not zero

$\emptyset_0^2 = 67.21, df = 8$, which is chi-square distributed with 8 degrees of freedom. We reject null hypothesis if

$$\emptyset_0^2 > \chi_8(0.05)$$

$p - value$ =1.755074e-11<0.05 Reject null hypothesis.

Conclusion: at least the correlation of the first canonical variates is significantly large then zero. For the depression data the $\boldsymbol{corr(\hat{u}_1, \hat{v}_1) = 0.3917}$.

$$H_0: \rho_1 \neq 0, \quad \rho_2 = 0, \quad v.s. \quad H_a: \rho_2 \neq 0$$

$\emptyset_0^2 = 18.984, df = 3$, which is chi-square distributed with 3 degrees of freedom. We reject null hypothesis if

$$\emptyset_0^2 > \chi_3(0.05)$$

$p - value$ =0.0002753796<0.05 Reject null hypothesis.

Conclusion: $\rho_2 \neq 0,$ **the correlation of the second canonical variates is significantly large then zero. For the depression data** the $\boldsymbol{corr(\hat{u}_2, \hat{v}_2) = 0.2519}$

3. Ex. 3.5 (15 pts) Repeat the regression analysis for the air pollution data described in the text after removing whatever cities you think should be regarded as outliers. For the results given in the text and the results from the outliers-removed data, produce scatterplots of Sulphur dioxide concentration against each of the principal component scores. Interpret your results.

   Multiple ways to do the question.

4. (30 pts) Construct PCA for heptathlon data (in textbook section 3.10.2) as the examples did in class. Interpret all the results you obtained from PCA package in R.

   (The example in textbook, skip)

5. (20 pts) Measurements of properties of pulp fibers and the paper made from them are contained in propertydata. There are $n = 62$ observations of the pulp fiber characteristics.

   Let the paper characteristics be
   $x_1 =$ breaking length, $x_2 =$ elastic modulus, $x_3 =$ stress at failue, $x_4 =$ burst strength.

   Let the pulp fiber characteristics be
   $y_1 =$ arithmetic fiber length, $y_2 =$ long fiber fraction, $y_3 =$ fine fiber fraction,
   $y_4 =$ zero span tensile.

   One approach to studying relationship between the two sets of variables is to use canonical correlation analysis which describes the relationship between the first set of variables and the second set of variables.

   a) Determine the number of significant canonical variate pairs;
   - $H_0: \rho_1^* = \rho_2^* = \rho_3^* = \rho_4^* = 0$, H_a : at least one of them is not zero.

     $\emptyset_0^2 = 170.86, df = 16$, which is chi-square distributed with 16 degrees of freedom. We reject null hypothesis if

     $$\emptyset_0^2 > \chi_{16}\,(0.05)$$

     p-value <0.05 Reject null hypothesis and conclude that at least $\rho_1 \neq 0$.

   - $H_0: \rho_1 \neq 0, \rho_2^* = \rho_3^* = \rho_4^* = 0$, H_a : at least one of them is not zero.

     $\emptyset_0^2 = 66.79, df = 9$, which is chi-square distributed with 9 degrees of freedom. We reject null hypothesis if

     $$\emptyset_0^2 > \chi_9\,(0.05)$$

     p-value <0.05 Reject null hypothesis and conclude that at least $\rho_1 \neq 0, \rho_2 \neq 0$.

```
> n<-nrow(propertydata)
> q1<-4
> q2<-4
> k<-as.matrix(log(1-e1$values))
> m<-sum(k[c(2:4),])
> phi2<--{n-1-0.5*(q1+q2+1)}*m  ( it is ok if you use <--{n-0.5*(q1+q2+1)}*m)
> (pvlaue<-1-pchisq(phi2,(q1-1)*(q2-1)))
[1] 6.451573e-11
```

   - $H_0: \rho_1 \neq 0, \rho_2 \neq 0$, $\rho_3^* = \rho_4^* = 0$, H_a : at least one of them is not zero.

$\emptyset_0^2 = 4.6032, df = 4$, which is chi-square distributed with 4 degrees of freedom. We reject null hypothesis if

$$\emptyset_0^2 > \chi_9\,(0.05)$$

p-value =0.33>0.05 we fail to reject the null hypothesis and conclude that there are sufficient evidence that $\rho_3^* = \rho_4^* = 0$.

```
k<-as.matrix(log(1-e1$values))
> m<-sum(k[c(3:4),])
> phi2<--{n-1-0.5*(q1+q2+1)}*m  ( it is ok if you use <--{n-0.5*(q1+q2+1)}*m)
> (pvlaue<-1-pchisq(phi2,(q1-2)*(q2-2)))
[1] 0.3304761
> phi2
[1] 4.603281
```

- The number of significant canonical variate pairs is 2.

b) Compute the canonical variates from the data;
R output by using $R_{11}^{-1}\ R_{12}\ R_{22}^{-1}\ R_{21}$ to solve for eigenvalues and eigenvectors.

```
> (e1<-eigen(E1))

$values
[1] 0.841493052 0.667369624 0.070429392 0.008405959

$vectors
            [,1]       [,2]       [,3]        [,4]
[1,]  0.57891084  0.6386846 -0.5241572 -0.63806138
[2,]  0.08149912  0.2819340  0.3240802 -0.07362493
[3,] -0.76847821 -0.1965904 -0.4333236  0.76160724
[4,] -0.26011801 -0.6884384  0.6576184 -0.08608986
```

$\widehat{U_1} = 0.579z_1 + 0.081z_2 - 0.768z_3 - 0.260z_4$
$\widehat{U_2} = 0.639z_1 + 0.282z_2 - 0.197z_3 - 0.688z_4$

```
> (e2<-eigen(E2))

$values
[1] 0.841493052 0.667369624 0.070429392 0.008405959

$vectors
           [,1]         [,2]       [,3]         [,4]
[1,] -0.1454230  0.347810919 -0.2688201  0.730320638
[2,]  0.5773914  0.506597476  0.0404533 -0.682498842
[3,]  0.2966067  0.002530949 -0.8715623  0.006958636
[4,]  0.7466564 -0.788910740 -0.4080178  0.027904710
```

$\widehat{V_1} = -.1454z'_1 + .5774z'_2 + 0.2966z'_3 + .7467z'_4$
$\widehat{V_2} = .3478z'_1 + .5066z'_2 + 0.0025z'_3 - .7889z'_4$

z and z' mean the standardized data for data set X and data set Y.

Based on the SAS output for eigenvectors regarding to the original variables

$\widehat{U_1} = -0.522X_1 - 0.2958X_2 + 1.366X_3 + .9760X_4$

$\widehat{U_2} = -1.213X_1 - 2.154X_2 + 0.736X_3 + 5.437X_4$

$\widehat{V_1} = -0.638Y_1 + .043Y_2 + 0.019Y_3 + 27.73Y_4$
$\widehat{V_2} = 2.76Y_1 + .067Y_2 + 0.0003Y_3 - 52.96Y_4$

c) Interpret each member of a canonical variate pair using its correlations with the member variables;

Correlations Between the paper characteristics and Their Canonical Variables				
	paper1	paper2	paper3	paper4
breaking_length	0.9351	-0.1261	-0.0534	-0.3270
elastic_modulus	0.8869	-0.4280	0.1306	-0.1148
stress_failue	0.9767	-0.1453	-0.0307	-0.1549
burst_strength	0.9518	0.0147	0.0127	-0.3061

- The first canonical variable for paper: All correlations are uniformly large $(0.9351, 0.8869, 0.9767, 0.9518)$. Therefore, this canonical variate as an overall measure of the paper characteristic.

- The second canonical variable for paper characteristic: none of the correlations is particularly large, and so, this canonical variable yields little information about the data.

- We had decided earlier not to look at the third fourth canonical variate pairs.

Correlations Between the fiber characteristics and Their Canonical Variables				
	fiber1	fiber2	fiber3	fiber4
length	0.8166	0.3683	0.1661	0.4122
fraction	0.9056	0.3848	0.1779	-0.0126
fine_fraction	-0.6496	0.0123	-0.7309	-0.2087
tensile	0.9395	-0.2307	0.1851	0.1730

- The first canonical variable for fiber: All correlations are large $(0.8166, 0.9056, -0.6496, 0.9395)$. Therefore, this canonical variate as an overall measure of the fiber characteristic.

- The second canonical variable for fiber characteristic: none of the correlations is particularly large, and so, this canonical variable yields little information about the data.

d) Use the results of canonical correlation analysis to describe the relationships between two sets of variables.

- 84.15% of the variation in u_1 is explained by the variation in v_1. 66.74% of the variation in u_2 is explained by v_2. Both pairs $(u_1, v_1), (u_2, v_2)$ have large canonical correlation and implies that the both canonical correlation are important.

```
title "Canonical Correlation Analysis - fibers and paper ";
data sales;
  infile "G:\Spring_2017\math 4400\Data\papertype.txt";
  input breaking_length elastic_modulus stress_failue burst_strength length
fraction fine_fraction tensile;
  run;

proc cancorr out=canout vprefix=paper vname="paper characteristics"
                        wprefix=fiber wname="fiber characteristics";
  var breaking_length elastic_modulus stress_failue burst_strength;
  with length fraction fine_fraction tensile;
  run;

proc gplot;
  axis1 length=3 in;
  axis2 length=4.5 in;
  plot paper1*fiber1 / vaxis=axis1 haxis=axis2;
  symbol v=J f=special h=2 i=r color=black;
  run;
```