Assignment 4

MTH 314-Multivariate Analysis

All questions are in reference to the chapter on discriminant analysis and principal component analysis. So, all terms have the usual meaning. The numerical-based questions have to be analyzed using any open-source software.

1. Consider two groups in a city π_1: riding –mower owners and π_2: those without riding mowers that is nonowners. In order to identify the best sales prospects for an intensive sales campaign a riding-mower manufacturer is interested in classifying families as prospective owners or nonowners on the basis of x_1=income and x_2= lot-size data. Random samples of n_1=12 current owners and $n_2=12$ current nonowners yield the values in following table:

x_1 (income in \$ 1000s)	x_2 (lot size in 1000 ft^2)	x_1 (income in \$ 1000 ft^2)	x_2 (lot size in 1000 ft^2)
60.0	18.4	75.0	19.6
85.5	16.8	52.8	20.8
64.8	21.6	64.8	17.2
61.5	20.8	43.2	20.4
87.0	23.6	84.0	17.0
110.1	19.2	49.2	17.6
108.0	17.6	59.4	16.0
82.8	22.4	66.0	18.4
69.0	20.0	47.4	16.4
93.0	20.8	33.0	18.8
51.0	22.0	51.0	14.0
81.0	20.0	63.0	14.8

(i) Plot the data and check if the classification rule is needed or not?

(ii) Obtain the linear discriminant function and set up the classification regions for the two populations.

2. A researcher has enough data available to estimate the density functions $p_1(\underset{\sim}{x})$ and $p_2(\underset{\sim}{x})$ associated with the populations π_1 & π_2, respectively. Suppose $C(2/1)=5$ units and $C(1/2)=10$ units. In addition, it is known that about 20% of all objects (for which the measurements $\underset{\sim}{x}$ can be recorded) belongs to π_2. Obtain the classification rule.

   Suppose the density evaluated at a new observation $\underset{\sim}{x}_0$ give $p_1(\underset{\sim}{x}_0)=.3$ and $p_2(\underset{\sim}{x}_0)=.4$. Where do we classify the new observation as π_1 and π_2,.

3. Consider the data matrices and descriptive statistics given below:

$$\underline{X}_1 = \begin{bmatrix} 2 & 4 & 3 \\ 12 & 10 & 8 \end{bmatrix}, \underline{\overline{x}}_1 = \begin{bmatrix} 3 \\ 10 \end{bmatrix}, \quad 2S_1 = \begin{bmatrix} 2 & -2 \\ -2 & 8 \end{bmatrix}$$

$$\underline{X}_2 = \begin{bmatrix} 5 & 3 & 4 \\ 7 & 9 & 5 \end{bmatrix}, \underline{\overline{x}}_2 = \begin{bmatrix} 4 \\ 7 \end{bmatrix}, \quad 2S_2 = \begin{bmatrix} 2 & -2 \\ -2 & 8 \end{bmatrix}, \quad n_1 = n_2 = 3.$$

Find out the confusion matrix and find apparent error rate.

4. Consider the two data sets

$$\underline{X}_1 = \begin{bmatrix} 3 & 2 & 4 \\ 7 & 4 & 7 \end{bmatrix}, \qquad \underline{X}_2 = \begin{bmatrix} 6 & 5 & 4 \\ 9 & 7 & 8 \end{bmatrix},$$

from two bivariate normal populations with same covariance matrix. Calculate the linear discriminant function. Further classify the observation $\underline{x}_0 = \begin{bmatrix} 2 \\ 7 \end{bmatrix}$ with equal priors and equal costs.

Assuming that the data on $\underline{X}_1$ and $\underline{X}_2$ is coming from the two bivariate normal population with different covariance matrices, find the classification rule and classify the new observation $\underline{x}_0 = \begin{pmatrix} 2 \\ 7 \end{pmatrix}$.

5. The following outcome is reported about the results of an experiment where subjects responded to "probe words" at five positions in a sentence. The variables are response times for the j^{th} probe word, $y_j, j = 1,2,...,5$. The data are given in following Table.

Table : <u>Response Times for Five Probe Word Positions</u>

Subject Number	y_1	y_2	y_3	y_4	y_5
1	51	36	50	35	42
2	27	20	26	17	27
3	37	22	41	37	30
4	42	36	32	34	27
5	27	18	33	14	29
6	43	32	43	35	40
7	41	22	36	25	38
8	38	21	31	20	16
9	36	23	27	25	28
10	26	31	31	32	36
11	29	20	25	26	25

Do a principal component analysis of the probe word data of above Table. Use both S (Sample Variance Covariance matrix) and R (Correlation matrix). Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either S or R?

6. Six hematology variables were measured were measured on 20 workers which are y_1 = hemoglobin concentration, y_2 = packed cell volume, y_3 = white blood cell count y_4 = lympocycte count , y_5 = neutruphil count and y_6 = serum lad concentration. The data is given in the following table.

Table: Hematology Data

Observation Number	y_1	y_2	y_3	y_4	y_5	y_6
1	13.4	39	4100	14	25	17
2	14.6	46	5000	15	30	20
3	13.5	42	4500	19	21	18
4	15.0	46	4600	23	16	18
5	14.6	44	5100	17	31	19
6	14.0	44	4900	20	24	19
7	16.4	49	4300	21	17	18
8	14.8	44	4400	16	26	29
9	15.2	46	4100	27	13	27
10	15.5	48	8400	34	42	36
11	15.2	47	5600	26	27	22
12	16.9	50	5100	28	17	23
13	14.8	44	4700	24	20	23
14	16.2	45	5600	26	25	19
15	14.7	43	4000	23	13	17
16	14.7	42	3400	9	22	13
17	16.5	45	5400	18	32	17
18	15.4	45	6900	28	36	24
19	15.1	45	4600	17	29	17
20	14.2	46	4200	14	25	28

Carry out a principal component analysis on the hematology data of Table. Use both S (Sample Variance Covariance matrix) and R (Correlation matrix). Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either S or R? Does the large variance of y_3 affect the pattern of the components of S?

7. Twenty engineer apprentices and 20 pilots were given 6 tests. The variables considered for this are y_1 = intelligence, y_2 = form relations, y_3 = dynamometer, y_4 = dotting, y_5 = sensory motor coordination, y_6 = perseveration. The data is presented in the following table:

Table : Comparison of Six on Engineer Apprentices and Pilots

Table : Comparison of Six Tests on Engineer Apprentices and Pilots

Engineer Apprentices						Pilots					
y_1	y_2	y_3	y_4	y_5	y_6	y_1	y_2	y_3	y_4	y_5	y_6
121	22	74	223	54	254	132	17	77	232	50	249
108	30	80	175	40	300	123	32	79	192	64	315
122	49	87	266	41	223	129	31	96	250	55	319
77	37	66	178	80	209	131	23	67	291	48	310
140	35	71	175	38	261	110	24	96	239	42	268
108	37	57	241	59	245	47	22	87	231	40	217
124	39	52	194	72	242	125	32	87	227	30	324
130	34	89	200	85	242	129	29	102	234	58	300
149	55	91	198	50	277	130	26	104	256	58	270
129	38	72	162	47	268	147	47	82	240	30	322
154	37	87	170	60	244	159	37	80	227	58	317
145	33	88	208	51	228	135	41	83	216	39	306
112	40	60	232	29	279	100	35	83	183	57	242
120	39	73	159	39	233	149	37	94	227	30	240
118	21	83	152	88	233	149	38	78	258	42	271
141	42	80	195	36	241	153	27	89	283	66	291
135	49	73	152	42	249	136	31	83	257	31	311
151	37	76	223	74	268	97	36	100	252	30	225
97	46	83	164	31	243	141	37	105	250	27	243
109	42	82	188	57	267	164	32	76	187	30	264

Carry out a principal component analysis on the engineer data of Table as follows:

(a) Use the pooled covariance matrix.

(b) Ignore groups and use a covariance matrix based on all 40 observations.

(c) Which of the approaches in (a) or (b) appears to be more successful?

8. Use the following data in R using the commands in **Blue Courier** font

```
library("factoextra")
data(decathlon2)
decathlonpc = decathlon2[1:23, 1:10]
```

Consider the data set `decathlon2` which is available in the package `factoextra` using the commands and store the data in `decathlon2[1:23, 1:10]` as `decathlonpc`

```
library("factoextra")
data(decathlon2)
decathlonpc = decathlon2[1:23, 1:10]
```

Conduct the principal component analysis and find (i) all the principle components (ii) variance of each principle component (iii) proportion of variance carried by each principle component (iv) construct biplot and find which of the variables contributes significantly more than other variables (v) construct a scree plot and find which of the principal components should be considered so that maximum variation is taken care

9. Consider the following data on five variables **x1**, **x2**, **x3**, **x4** and **x5**. Consider two sets of data created from these variables as set 1 created from the data frame 1 (**y1**) with data in **x2,x3** and set 2 (**y2**) created from data frame 2 in **x1,x4,x5**. Conduct the canonical correlation analysis with **y1** and **y2**. Find all the canonical correlations and associated canonical variables.

```
x1=c(21.65,25.50,20.27,15.59,23.25,19.65,7.82,21.68,16.15,2
2.84,25.53,13.81,24.19,20.54,21.20,18.92,12.88,15.60,11.06,
18.41,20.91,24.94,25.92,15.45,20.24,25.50,18.43,23.60,22.43
,30.06)

x2=c(70.27,63.79,64.39,81.77,81.27,75.74,81.45,87.69,86.69,
86.45,65.50,85.16,68.58,66.30,66.65,67.10,88.78,86.70,71.39
,81.29,71.44,65.93,66.86,82.15,61.01,71.44,64.52,65.30,69.8
9,68.68)

x3=c( 1.35, 7.04, 1.27, 3.35, 1.94, 5.97,-3.90, 0.83, 0.38,
4.82, 7.25, 4.32, 2.55,11.58, 1.81,-0.33,-1.79, 1.75, 6.43,
```

```
2.62, 6.12, 3.32,10.70, 5.64, 9.32, 6.30, 4.79,-0.64, 9.26,
7.25)

x4=c(2622.89,1807.45,2408.27, 489.89,1030.47,3283.34,
959.30, 597.31, 574.92, 763.91,2797.78,
580.89,1981.92,2512.07,2760.66,1173.08, 586.98,
530.05,2201.70, 388.23,1447.49,1691.73,1557.82,
500.34,2752.41, 898.08,2533.87,2044.42,1782.05,2640.01)

x5=c(9.12, 4.38,12.19, 6.29, 9.20, 6.62,11.84, 6.57, 9.87,
2.66, 8.94, 4.72, 7.09,13.23,10.33, 8.38, 3.02, 8.83, 6.38,
7.49,12.43, 2.30, 9.53,11.71, 7.11,12.76, 9.59,11.45,
4.23,25.63)
```