

## Assignment 1 Solutions

1. (10 points)

- (a) (1 point) Use R to calculate the first two principal components based on the correlation matrix for this data set.

```
> uscr = read.table(file = url("http://www.maths.usyd.edu.au/u/UG/SM/STAT3014/"),
+   header = TRUE)
> colnames(uscr)

[1] "CR" "N.S" "AM" "YS" "PE" "Lab" "M.F" "Pop" "NW" "U" "MI"

> X = uscr[, 3:11]
> X.pc = princomp(X, cor = TRUE)
> summary(X.pc)
```

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.9203613	1.4146873	1.1292744	0.88152521	0.70102375
Proportion of Variance	0.4097542	0.2223711	0.1416956	0.08634297	0.05460381
Cumulative Proportion	0.4097542	0.6321253	0.7738209	0.86016391	0.91476772
	Comp.6	Comp.7	Comp.8	Comp.9	
Standard deviation	0.55145573	0.46791944	0.38563499	0.30874608	
Proportion of Variance	0.03378927	0.02432762	0.01652382	0.01059157	
Cumulative Proportion	0.94855699	0.97288461	0.98940843	1.00000000	

Together, the first three p.c.s explain 77.4 % of the total variability.

- (b) (2 point) The loadings for the first two principal components are:

```
> X.pc$loadings[, 1:2]

      Comp.1   Comp.2
AM  0.39640132 -0.1797704
YS -0.45762250 -0.1511717
PE -0.36012827  0.3673957
Lab -0.26599027 -0.3354231
M.F -0.20552320 -0.5314358
Pop -0.10464711  0.5802753
NW  0.39165213  0.1775546
U   -0.06370812 -0.1305172
MI  -0.47155437  0.1721560
```

The standard deviation for the first two principal components:

```
> X.pc$sdev[1:2]
```

```

  Comp.1  Comp.2
1.920361 1.414687

```

- (c) (2 points) The correlations of the first principal components with all the original variables.

```
> X.pc$loadings[, 1] * X.pc$sdev[1]
```

```

      AM      YS      PE      Lab      M.F      Pop      NW
0.7612337 -0.8788005 -0.6915764 -0.5107974 -0.3946788 -0.2009603 0.7521136
      U      MI
-0.1223426 -0.9055548

```

- (d) (2 points)

```
> loadings(X.pc)
```

Loadings:

```

      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
AM   0.396 -0.180 0.241 0.339 0.279 0.669 -0.155 0.283
YS  -0.458 -0.151 0.131      0.140 0.125 -0.743 -0.391
PE  -0.360 0.367      0.388 0.397 -0.126 0.233      -0.590
Lab -0.266 -0.335 0.478 0.203 -0.555 -0.183      0.425 -0.155
M.F -0.206 -0.531 -0.113 0.495      0.404 -0.431 0.249
Pop -0.105 0.580      0.308 -0.549 0.414      -0.249 0.148
NW  0.392 0.178 0.127 0.529      -0.555 -0.358      0.264
U      -0.131 -0.814 0.265 -0.182      -0.277 0.331 -0.165
MI  -0.472 0.172      0.301      0.470 0.657

```

```

      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
SS loadings      1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000
Proportion Var 0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111
Cumulative Var 0.111 0.222 0.333 0.444 0.556 0.667 0.778 0.889 1.000

```

```
> apply(X, 2, cor, X.pc$score[, 1:3])
```

```

      AM      YS      PE      Lab      M.F      Pop
[1,] 0.7612337 -0.8788005 -0.69157639 -0.5107974 -0.3946788 -0.20096026 0.7521
[2,] -0.2543189 -0.2138607 0.51974999 -0.4745188 -0.7518155 0.82090815 0.2511
[3,] 0.2723493 0.1481778 0.08126539 0.5393984 -0.1281401 0.01099317 0.1430
      U      MI
[1,] -0.1223426 -0.90555476
[2,] -0.1846410 0.24354696
[3,] -0.9188742 0.01521385

```

Thus pc1 is highly correlated with median income, years of schooling and police expenditure and is negatively correlated with adult males and non-whites.

PC2 is positively correlated with population and negatively correlated with males and females

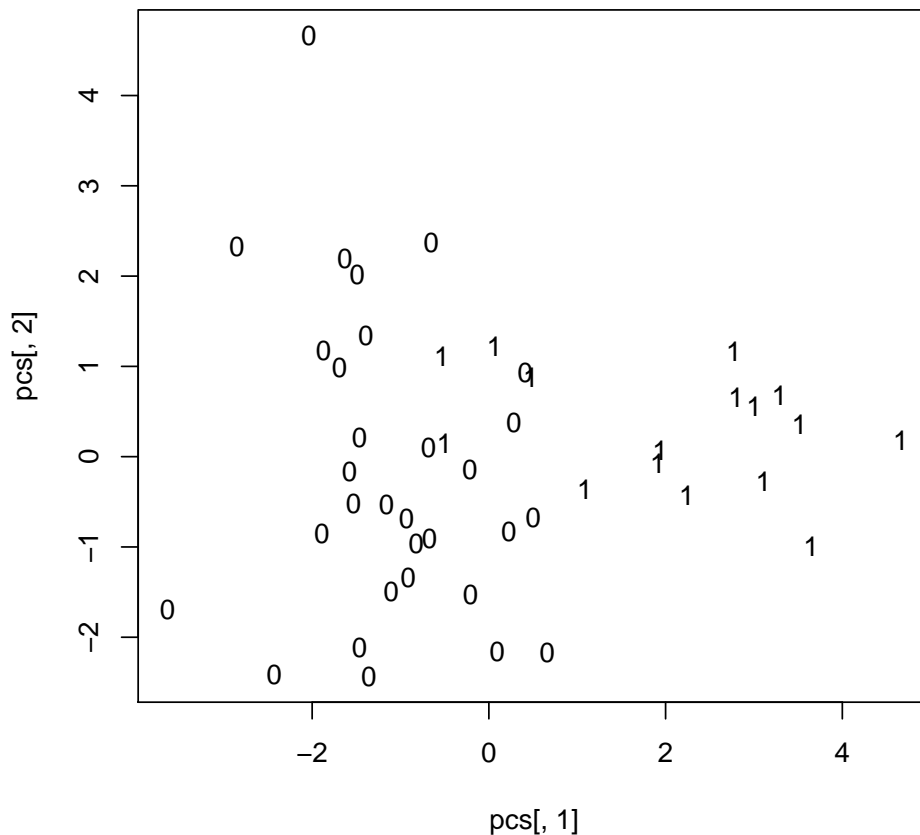
- (e) (3 points)

```

> pcs = X.pc$scores[, 1:2]
> plot(pcs[, 1], pcs[, 2], main = "p.c. scores coded with North (0) or South
+     type = "n")
> text(pcs[, 1], pcs[, 2], uscr[, 2])

```

**p.c. scores coded with North (0) or South (1)**



The graph shows that PC1 effectively separates the Northern and Southern states and the Southern states have less variables values for PC2 (less variability in population and M/F ratio).

- (6 points) Calculate the bivariate vector of differences of laboratory readings and test the hypothesis that the two laboratories produce the same results on average.

Null hypothesis:

$$H_0 : \mu_{private} - \mu_{state} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

```

> effluent = read.table(file = url("http://www.maths.usyd.edu.au/u/UG/SM/STAT3014/
+     header = TRUE)
> E1 = effluent[1:11, 2:3]

```

```

> E2 = effluent[12:22, 2:3]
> d = E1 - E2
> d

      BOD SS
1  -19 12
2  -22 10
3  -18 42
4  -27 15
5   -4 -1
6  -10 11
7  -14 -4
8   17 60
9    9 -2
10  4 10
11 -19 -7

> m = apply(d, 2, mean)
> m

      BOD      SS
-9.363636 13.272727

> v = var(d)
> t = 11 * t(m) %*% solve(v) %*% m
> t

      [,1]
[1,] 13.63931

> ta <- t * 9/(10 * 2)
> ta

      [,1]
[1,] 6.13769

> 1 - pf(ta, 2, 9)

      [,1]
[1,] 0.02082779

```

Using Hotellin's  $T^2$  statistics, the statistics is 13.639 with corresponding  $p$ -value for the test is 0.021. Therefore, we reject the claim that the two laboratories produce the same average results.

3. (12 points)

- (a) (4 points) Use the Mahalanobis statistic  $D^2 = 916.1616$ . If  $H_0$  is true then  $D^2 \sim \frac{98 \times 5}{94} F_{5,94}$ . The  $p$ -value is

$$p = P(F_{5,94} \geq 94 \times 916.1616/490) = P(F_{5,94} \geq 175.75) < 0.001.$$

Thus we reject the equal mean hypothesis.

The test is based on the assumption that we have two independent samples drawn from multivariate normal populations  $N_5(\boldsymbol{\mu}_i, \Sigma)$  with common covariance structure.

- (b) (3 points) If  $\boldsymbol{\mu}_{BM} = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)^T$  we want to test

$$H_0 : \mu_4 = \mu_3 + 5 \quad \text{and} \quad \mu_1 = \mu_5 + 5$$

that is x

$$\begin{pmatrix} 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & 0 & -1 \end{pmatrix} \boldsymbol{\mu}_{BM} = \begin{pmatrix} 5 \\ 2 \end{pmatrix}$$

or  $D\boldsymbol{\mu}_{BM} = \mathbf{r}$ . Use

$$T^2 = 50(D\bar{\mathbf{X}}_1 - \mathbf{r})^T (DV_1 D^T)^{-1} (D\bar{\mathbf{X}}_1 - \mathbf{r}),$$

where  $\bar{\mathbf{X}}_1$  is the sample mean vector for male blue crabs and  $V_1$  is the sample covariance matrix. If  $T^2 = 132.8$  the the  $p$ -value for the test is

$$p = P(F_{2,48} \geq \frac{132.8 \times 48}{49 \times 2} = 65.05).$$

- (c) (1 point) There is an error in the calculation of the  $W$  matrix. It should read  $W = 49 \times (V1 + V2 + V3 + V4)$ .
- (d) (2 points) Assume there is no error in the Routput. The MANOVA statistic is  $V = (200 - 4) \sum_{i=1}^5 \lambda_i = 196 \times 2147.25 = 420861.9$ . If the null hypothesis is true the exact distribution of the statistic is  $\chi_{15}^2$ .
- (e) (2 points) Let us denote the incorrect  $W$  as  $W_o$ , then the correct  $W = 49 * (V1 + V2 + V3 + V4) = 49 \times (4 \times W_o)$ .  
Therefore,  $W^{-1}B = \frac{W_o^{-1}B}{196}$  and the correct  $V = (200 - 4) \sum_{i=1}^5 \lambda_i = 196 \times \frac{1}{196} \times 2147.25 = 2147.25$ .

4. (12 points) Data on 9 mandible measurements (in mm)

- (a) (3 points) Test the hypothesis that there is no difference between Indian wolves' and prehistoric Thai dogs' mean mandible measurements.

```
> c4 = read.table(file = url("http://www.maths.usyd.edu.au/u/UG/SM/STAT3014/r/
+   header = FALSE)
> c5 = read.table(file = url("http://www.maths.usyd.edu.au/u/UG/SM/STAT3014/r/
+   header = FALSE)
> m1 = apply(c4, 2, mean)
```

```

> m2 = apply(c5, 2, mean)
> v1 = var(c4)
> v2 = var(c5)
> sp = (13 * v1 + 9 * v2)/22
> d = (m1 - m2)
> D = (14 * 10/24) * t(d) %>% solve(sp) %>% d
> D

```

```

      [,1]
[1,] 243.9862
> Da = (24 - 9 - 1)/(9 * 22) * D
> Da

```

```

      [,1]
[1,] 17.25155
> 1 - pf(Da, 9, 14)

```

```

      [,1]
[1,] 4.134174e-06

```

Thus the  $p$ -value for testing equality of mean vectors is very small and so there is strong evidence to argue in favour of these two species having different mean mandible measurements.

- (b) (2 points) Calculate Fisher's linear discriminant function based on these data.

```

> a <- solve(sp) %>% (m1 - m2)
> a

```

```

      [,1]
V1 -0.4553477
V2 -2.6826106
V3  0.8536265
V4 -1.4878339
V5  8.4619098
V6 -3.4104227
V7  2.2182099
V8  0.7179881
V9 -5.6847117
> t(a) %>% (m1 + m2)/2

```

```

      [,1]
[1,] 117.7814

```

Fisher's linear discriminant function is  $h(\mathbf{x}) = \mathbf{a}^T \mathbf{x}$ . Classify a dog as an Indian Wolf if the score is over 117.78.

- (c) (2 points) Build a classification tree.

```

> library(rpart)
> doglab = c(rep("IW", nrow(c4)), rep("TD", nrow(c5)))

```

```

> cdat = rbind(c4, c5)
> dogtree = rpart(factor(doglab) ~ ., dat = cdat)
> dogtree

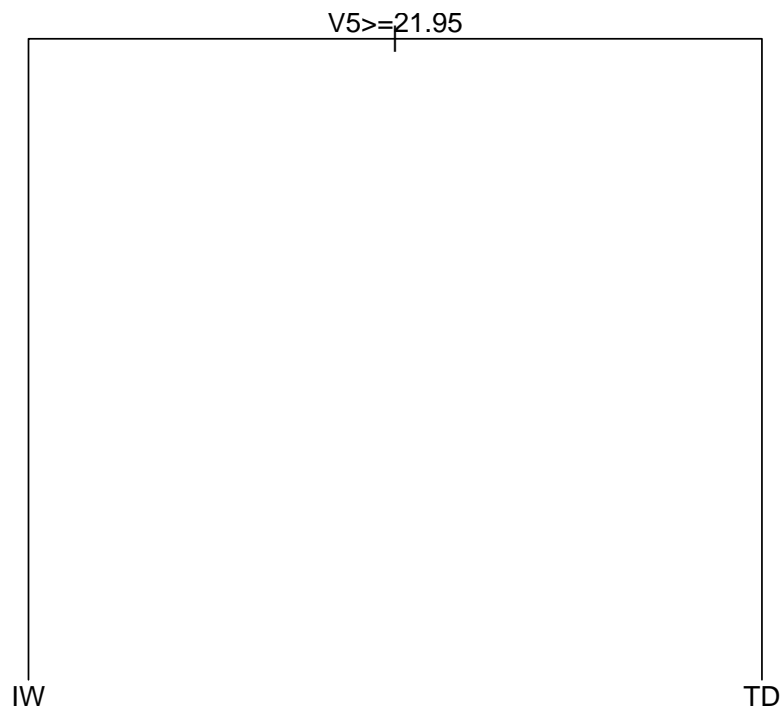
n= 24

node), split, n, loss, yval, (yprob)
  * denotes terminal node

1) root 24 10 IW (0.5833333 0.4166667)
  2) V5>=21.95 14 0 IW (1.0000000 0.0000000) *
  3) V5< 21.95 10 0 TD (0.0000000 1.0000000) *

> plot(dogtree)
> text(dogtree)

```



The discrimination rule from the classification tree only based on one variable (V5). We classify an observation to Thai Dog (TD) if  $V5 < 21.95$ , otherwise, we classify it as Inadian wolf (IW)

(d) (2 points) How would you classified a dog with the following 9 mandible measure-

ments. (131.1, 14.8, 20.9, 23.5, 20.8, 8.9, 37.7, 40.4, 6.5)

```
> newdog = data.frame(V1 = 131.1, V2 = 14.8, V3 = 20.9, V4 = 23.5,  
+ V5 = 20.8, V6 = 8.9, V7 = 37.7, V8 = 40.4, V9 = 6.5)  
> predict(dogtree, newdog)
```

```
   IW TD  
1  0  1
```

```
> as.numeric(newdog) %**% a
```

```
      [,1]  
[1,] 104.8155
```

Both rules (classification tree and FLDA) classify the new dog as a Thai Dog.

- (e) (3 points) What is the 5-fold cross validation error rate when using a *KNN* classifier with  $k = 3$

```
> library(class)  
> neworder = sample(1:24, 24)  
> neworder  
[1] 21 19 13 22  9  6  4  3  5 15  8 17 10 11  1  7 18 14 16 20  2 12 24 23  
> CVsub = matrix(c(c(0:4) * 5 + 1, c(1:5) * 5), nrow = 5)  
> CVsub[5, 2] = 24  
> CVsub  
      [,1] [,2]  
[1,]    1    5  
[2,]    6   10  
[3,]   11   15  
[4,]   16   20  
[5,]   21   24  
> res = c()  
> for (j in 1:5) {  
+   index = neworder[CVsub[j, 1]:CVsub[j, 2]]  
+   print(index)  
+   TS = cdat[index, ]  
+   LS = cdat[-index, ]  
+   disc.knn <- as.vector(knn(LS, TS, cl = doglab[-index], k = 3))  
+   res = c(res, sum(disc.knn != doglab[index]))  
+ }
```

```
[1] 21 19 13 22  9  
[1]  6  4  3  5 15  
[1]  8 17 10 11  1  
[1]  7 18 14 16 20  
[1]  2 12 24 23
```

The CV error rate is given below:

```
> sum(res)/24
```

```
[1] 0.04166667
```