ST412-Multivariate Statistics with Advanced Topics Assignment 2

Name: Christis Katsouris
Student ID: 1154579
Department of Statistics
University of Warwick

January 23, 2012

1. Question 1

A data set of $n = 276$ measurements on skull and bone size of white fowl is available. Each measurement has six components: skull length, skull breadth, femur length, tibia length, humerus length and ulna length, recorded as a vector $(X_1, X_2, X_3, X_4, X_5, X_6)$ in six space dimensions. The sample correlation matrix is constructed and given by

$$\mathbf{R} = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{array} \begin{array}{c} \begin{array}{cccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \end{array} \\ \begin{pmatrix} 1 & 0.505 & 0.569 & 0.602 & 0.621 & 0.603 \\ 0.505 & 1 & 0.422 & 0.467 & 0.482 & 0.450 \\ 0.569 & 0.422 & 1 & 0.926 & 0.877 & 0.878 \\ 0.602 & 0.467 & 0.926 & 1 & 0.874 & 0.894 \\ 0.621 & 0.482 & 0.877 & 0.874 & 1 & 0.937 \\ 0.603 & 0.450 & 0.878 & 0.894 & 0.937 & 1 \end{pmatrix} \end{array}$$

(a) Give the principal component solution for the factor problem with two factors.

Solution **:**
The factor problem with 2 factors satisfies the following equation in index form.

$$\boxed{X_j - \mu_j = \sum_{k=1}^{m} \ell_{jk} F_k + \epsilon_j} \text{ where m=1,2 and j=1,...,6.}$$

The principal component factor analysis of the sample covariance matrix $\mathbf{S}$ is specified in terms of its eigenvalue-eigenvector pairs $(\hat{\lambda_1}, \hat{\mathbf{e_1}}), (\hat{\lambda_2}, \hat{\mathbf{e_2}}), ..., (\hat{\lambda_p}, \hat{\mathbf{e_p}})$,where $\hat{\lambda_1} \geq \hat{\lambda_2} \geq ... \geq \hat{\lambda_p}$. Let $m < p$ be the number of common factors. Then the matrix of the estimated factor loadings $\{\tilde{\ell_{ij}}\}$ is given by

$$\tilde{\mathbf{L}} = \left(\sqrt{\hat{\lambda_1}} \hat{\mathbf{e_1}} \vdots \sqrt{\hat{\lambda_2}} \hat{\mathbf{e_2}} \vdots \cdots \sqrt{\hat{\lambda_m}} \hat{\mathbf{e_m}} \right) \tag{1}$$

The estimated specific variances are provided by the diagonal elements of the matrix $\mathbf{S}$ - $\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T$, so

$$\tilde{\mathbf{\Psi}} = \begin{pmatrix} \tilde{\psi_1} & 0 & \cdots & 0 \\ 0 & \tilde{\psi_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\psi_p} \end{pmatrix} \text{ with } \tilde{\psi_i} = s_{ii} - \sum_{j=1}^{m} \tilde{\ell_{ij}^2} \tag{2}$$

Communalities are estimated as

$$\tilde{h_i^2} = \tilde{\ell_{i1}^2} + \tilde{\ell_{i2}^2} + ... + \tilde{\ell_{im}^2} \tag{3}$$

[Wichern, page 490]
Note that the given correlation matrix is the observed correlation matrix of the data set. Using R, we can compute the eigenvalues and the eigenvectors of the correlation matrix.

```
>data<-read.table("data.txt", header=F)
>pc<-eigen(data,sym=FALSE)
>eigenvalues<-pc$values
#We only need the e-vectors corresponding to the 1st two e-values
>eigenvectors<-pc$vectors[,1:2]
```

```
>eigenvalues
[1] 4.45644850 0.78240991 0.45842506 0.16883257 0.07908774 0.05479622
>eigenvectors
>eigenvectors
           [,1]       [,2]
[1,] 0.3507933  0.3956532
[2,] 0.2862040  0.8146406
[3,] 0.4399784 -0.2632446
[4,] 0.4468851 -0.1972029
[5,] 0.4488806 -0.1614667
[6,] 0.4474933 -0.2134503
```

Since we are working with 2 factors, we only consider the first two eigenvalue-eigenvector pair of $\mathbf{R}$. Therefore, using equation (1) we can compute the estimated factor loadings and using equations (2) and (3) the communalities and the specific variances.

```
#estimation of the loadings
loadings<-matrix(0,6,2)
for (j in 1:2)
{
 for (k in 1:6)
  {
   loadings[k,j]<- ( sqrt(eigenvalues[j]) ) * ( eigenvectors[k,j] )
  }
}

psi<-matrix(0,6,6) #the psi matrix
h.values<-matrix(0,1,6) #the 1 row matrix with the communalities
for (i in 1:6)
 {
  h <- loadings[i,1]^2 +  loadings[i,2]^2
  psi[i,i]<- (1 - h)
  h.values[i]<-h
 }

> loadings
           [,1]       [,2]
[1,] 0.7405352  0.3499708
[2,] 0.6041852  0.7205817
[3,] 0.9288076 -0.2328502
[4,] 0.9433880 -0.1744338
[5,] 0.9476006 -0.1428236
[6,] 0.9446719 -0.1888052

> h.values
         [,1]      [,2]      [,3]     [,4]      [,5]      [,6]
[1,] 0.670872 0.8842777 0.9169028 0.920408 0.9183454 0.9280524
```

The first two eigenvalues are $\hat{\lambda_1} = 4.456$ and $\hat{\lambda_2} = 0.782$. Thus, 2 factors will account for a cumulative

proportion

$$\frac{\hat{\lambda_1} + \hat{\lambda_2}}{p} = \frac{4.456 + 0.782}{6} = 0.873$$

The above quantities are shown on table 1.

Table 1: Principal component solution for the factor problem with two factors

Variable	Estimated factor loadings		Communalities	Specific variances
X_i	$\tilde{\ell_{ij}} = \sqrt{\hat{\lambda_i}} e_{ij}$		$\tilde{h_i^2}$	$\tilde{\psi_i} = 1 - \tilde{h_i^2}$
	F_1	F_2		
X_1	0.741	0.350	0.671	0.329
X_2	0.604	0.721	0.884	0.116
X_3	0.929	-0.233	0.917	0.083
X_4	0.943	-0.174	0.920	0.080
X_5	0.948	-0.143	0.918	0.082
X_6	0.945	-0.189	0.928	0.072
Eigenvalues	4.456	0.782		
Cumulative proportion of total sample variance	0.743	0.873		

$(\boldsymbol{b})$ Is the choice of only two factors well justified? Or would you choose a different number of factors?

$\boldsymbol{Solution}$ **:**
The choice of only two factors shows that it explains approximately 87% of the total sample variance. There are not enough evidence to say that 2 factors don't explain enough of the sample variance. Moreover if we look at the correlations between the variables we see that the following pairs have high correlations; $Cor(X_3, X_4) = 0.926, Cor(X_5, X_6) = 0.937, Cor(X_3, X_5) = 0.877, Cor(X4, X_5) = 0.874, Cor(X_3, X_6) = 0.878, Cor(X_4, X_6) = 0.894$. Therefore we could have one factor for the set of variables (X_3, X_4, X_5, X_6) and another one for the variables (X_1, X_2). Also we observe that the variable X_1 has the lowest correlations with the rest of the variables. Possibly 3 factors could explain more of the sample variance. However the interpretation of the factors is subjective and different method to extract the factors might have different interpretations.

$(\boldsymbol{c})$ What is the proportion of the total variance due to each one of the two factors?

$\boldsymbol{Solution}$ **:**
The proportion of the total variance due to each of the two factors is given by the eigenvalue that corresponds to the factor divided by the sum of the eigenvalues. Note that

$$\sum_{j=1}^{6} \lambda_j = trace(\boldsymbol{R})$$

$\frac{\hat{\lambda_1}}{6} = \frac{4.456}{6} = 0.743$ the proportion of total variance explained by F_1
$\frac{\hat{\lambda_2}}{6} = \frac{0.782}{6} = 0.130$ the proportion of total variance explained by F_2

(d) Compute the residual matrix and based on it comment on the efficiency of the factor solution you provided.

Solution **:**
To compute the residual matrix we first need to compute the matrix $\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T + \tilde{\mathbf{\Psi}}$, which is an approximation to the given correlation matrix.

$$\begin{aligned}
\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T + \tilde{\mathbf{\Psi}} &= \begin{pmatrix} 0.741 & 0.350 \\ 0.604 & 0.721 \\ 0.929 & -0.233 \\ 0.943 & -0.174 \\ 0.948 & -0.143 \\ 0.945 & -0.189 \end{pmatrix} \begin{pmatrix} 0.741 & 0.604 & 0.929 & 0.943 & 0.948 & 0.945 \\ 0.350 & 0.721 & -0.233 & -0.174 & -0.143 & -0.189 \end{pmatrix} \\
&+ \begin{pmatrix} 0.329 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.116 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.083 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.080 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.082 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.072 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0.700 & 0.606 & 0.638 & 0.652 & 0.633 \\ 0.700 & 1 & 0.393 & 0.444 & 0.470 & 0.435 \\ 0.606 & 0.393 & 1 & 0.917 & 0.913 & 0.921 \\ 0.638 & 0.444 & 0.917 & 1 & 0.919 & 0.924 \\ 0.652 & 0.470 & 0.913 & 0.919 & 1 & 0.922 \\ 0.633 & 0.435 & 0.921 & 0.924 & 0.922 & 1 \end{pmatrix}
\end{aligned}$$

The residual matrix corresponding to the two factor solution is

$$\mathbf{R} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T + \tilde{\mathbf{\Psi}}) = \begin{pmatrix} 0 & -0.195 & -0.037 & -0.036 & -0.031 & -0.030 \\ -0.195 & 0 & 0.029 & 0.023 & 0.012 & 0.015 \\ -0.037 & 0.029 & 0 & 0.009 & -0.036 & -0.043 \\ -0.036 & 0.023 & 0.009 & 0 & -0.045 & -0.030 \\ -0.031 & 0.012 & -0.036 & -0.045 & 0 & 0.015 \\ -0.030 & 0.015 & -0.043 & -0.030 & 0.015 & 0 \end{pmatrix}$$

Note that the entries of the above matrices are given in 3 decimal places. The above matrix manipulations can be easily computed in R, using the following code. (See Appendix A for the produced output in R.)

```
>corr.new<- loadings %*% tran.loadings + psi
>residual<- data - corr.new
```

We can conclude that the efficiency of the provided factor solution is good, since the residual matrix has entries very close to zero. In other words the correlation matrix is well estimated by the principal component method with 2 factors. However there is a high correlation between the first and the second variables, which gives a residual of 0.19.

2. Question 3

We consider the following data table

Table 2: Given data set

Individual	X_1	X_2	X_3
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
4	3.2	53.2	12.0
5	3.1	55.5	9.7
6	4.6	36.1	7.9
7	2.4	24.8	14.0
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5	58.8	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

$(\boldsymbol{a})$ Construct normal probability plots (Q-plots) for each observation. Construct the pairwise scatter plots. Does the assumption of multivariate normality seems to be justified?

$\boldsymbol{Solution}$ **:**
The pairwise scatter plots are shown on figure 1. The correlation coefficient for each pair is also displayed. We observe a positive correlation between all the pairs with the highest the one between X_1 and X_3. The normal probability plots for each variables X_1, X_2 and X_3 are shown on figures 2, 3 and 4 respectively. To check for multivariate normality we can check these 2 types of plots. In other words, we can check (i) the univariate marginals, by looking at the Normal probability plots and (ii) the bivariate marginals using two dimensional scatter plots. [$Zygouras, page2$]

By looking at the Normal probability plots (Q-plots) we observe a positive linear trend for all 3 univariate marginals, with some deviation in some cases. In particular a slightly positive skewness is observed for X_1 and a slightly negative skewness for X_3. The sample size is small anyway and some deviation from the desired plots is expected. Furthermore by looking on the pairwise plots, we know that if we consider any pair, then the points should resemble ellipses. This is because the level curves of the bivariate normal are ellipses. This assumption is not significant violated either. The points for the pair (X_2, X_3) seem to be more spread out than having the shape of an ellipse and so we should be careful with this pair. A bigger sample size might be more accurate. However overall, the multivariate normality seem to be justified.

$(\boldsymbol{b})$ Determine the axes of the 90% confidence ellipsoid for $\boldsymbol{\mu}$. Determine the lengths of these axes.

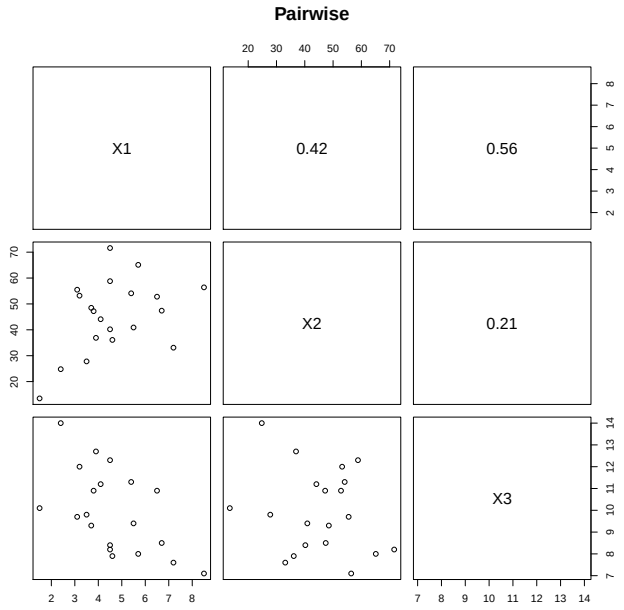

Figure 1: The pairwise scatter plots of the observations

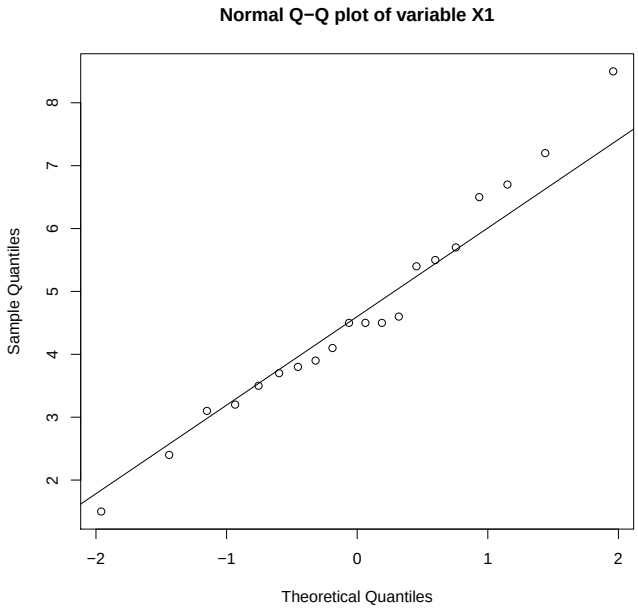

Figure 2: Normal probability plot of X1

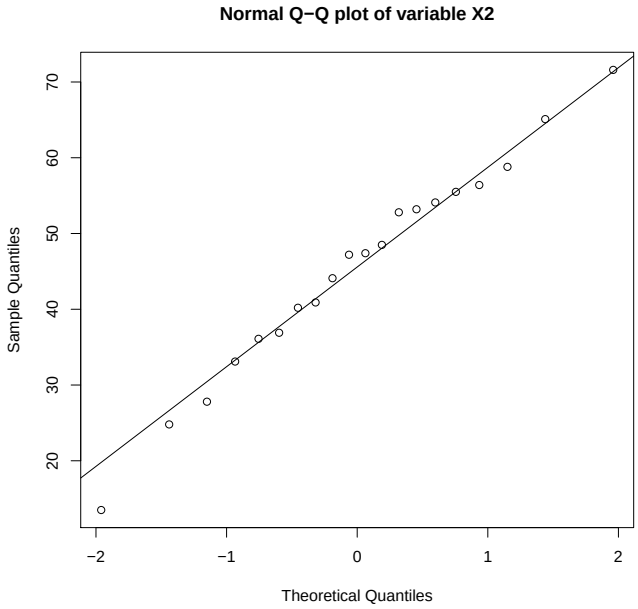

Figure 3: Normal probability plot of X2

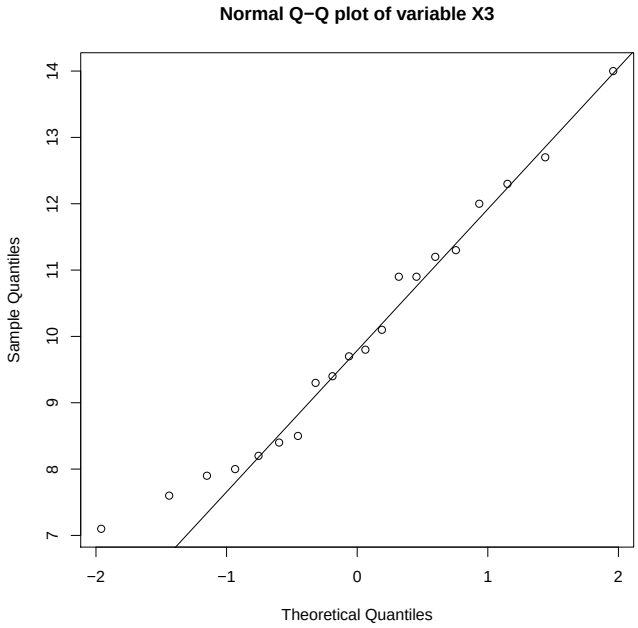

Figure 4: Normal probability plot of X3

Solution **:**
The required confidence interval for the mean of the data, is the case of computation of confidence interval for mean with unknown covariance matrix.

Proposition 1. *Let $\boldsymbol{X_1}, ...\boldsymbol{X_N}$ i.i.d with distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then*

$$\frac{N(N-p)}{p(N-1)}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim F_{p,N-p}$$

where $F_{p,N-p}$ is the F-distribution with p and $N-p$ degrees of freedom. Hence the $100(1-\alpha)\%$ confidence interval for the population mean is

$$\mathcal{R}(\boldsymbol{X_1}, ..., \boldsymbol{X_N}) = \left\{ \boldsymbol{x} \in \mathbb{R}^p : (\bar{\boldsymbol{X}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \leq \frac{p(N-1)}{N(N-p)} F_{p,N-p}(\alpha) \right\} \tag{4}$$

where $F_{p,N-p}(\alpha)$ is the $100(1-\alpha)\%$ percentile of the $F_{p,N-p}$ distribution and $\mathcal{R} = \mathcal{R}(\boldsymbol{X_1}, ..., \boldsymbol{X_N})$, is a $100(1-\alpha)\%$ confidence region for the population mean if

$$\mathbb{P}(\boldsymbol{\mu} \in \mathcal{R}(\boldsymbol{X_1}, ..., \boldsymbol{X_N})) = 1 - \alpha.$$

Hence from the given data we have, (see Appendix B for code and R output)

$$\bar{\boldsymbol{X}} = \frac{1}{N}\sum_{i=1}^{10} \boldsymbol{X_{i.}} = \begin{pmatrix} 4.640 \\ 45.400 \\ 9.965 \end{pmatrix}$$

$$\boldsymbol{S} = \frac{1}{N}(\boldsymbol{X_i} - \bar{\boldsymbol{X}})(\boldsymbol{X_i} - \bar{\boldsymbol{X}})^T = \begin{pmatrix} 2.879 & 10.010 & -1.809 \\ 10.010 & 199.788 & -5.640 \\ -1.809 & -5.640 & 3.628 \end{pmatrix} \text{ and } \boldsymbol{S}^{-1} = \begin{pmatrix} 0.586 & -0.022 & 0.258 \\ -0.022 & 0.006 & -0.002 \\ 0.258 & -0.002 & 0.402 \end{pmatrix}$$

Then the axes of the 90% confidence ellipsoid for $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^T$ can be computed using the region given by (4).

$$\begin{pmatrix} 4.64 - \mu_1 & 45.4 - \mu_2 & 9.965 - \mu_3 \end{pmatrix} \begin{pmatrix} 0.586 & -0.022 & 0.258 \\ -0.022 & 0.006 & -0.002 \\ 0.258 & -0.002 & 0.402 \end{pmatrix}$$

$$\text{x} \begin{pmatrix} 4.64 - \mu_1 \\ 45.4 - \mu_2 \\ 9.965 - \mu_3 \end{pmatrix} \leq \frac{3\text{x}19}{20\text{x}17} F_{3,17}(0.1) = 0.409$$

since $qf(0.90, 3, 17) = 2.437434$
which can be written as

$$\begin{aligned} \{0.586(4.64 - \mu_1)^2 &+ 0.006(45.4 - \mu_2)^2 + 0.402(9.965 - \mu_3)^2 \\ &+ 2(-0.022)(4.64 - \mu_1)(45.4 - \mu_2) + 2(0.258)(4.64 - \mu_1)(9.965 - \mu_3) \\ &+ 2(-0.002)(45.4 - \mu_2)(9.965 - \mu_3)\} \leq 0.409 \end{aligned}$$

Hence the confidence region is an ellipsoid given by the above equation. The centre of the ellipsoid is given by the sample mean vector, i.e if we are on the orthogonal system with axes (x_1, x_2, x_3), $C = (4.640, 45.400, 9.965)$. The major and the minor axes of the ellipsoid are given by the eigenvectors of the sample covariance matrix $\boldsymbol{S}$. To compute the eigenvalues of $\boldsymbol{S}$, we are using the equation $det(\boldsymbol{\lambda I} - \boldsymbol{S}) = 0$, which is equivalent to solve the following system of equations.

$$\begin{aligned}\boldsymbol{Se_1} &= \lambda_1 \boldsymbol{e_1}\\ \boldsymbol{Se_2} &= \lambda_2 \boldsymbol{e_2}\\ \boldsymbol{Se_3} &= \lambda_3 \boldsymbol{e_3}\end{aligned}$$

Thus, using R (see Appendix B), we get the eigenvalues $\lambda_1 = 200.462, \lambda_2 = 4.532$ and $\lambda_3 = 1.301$ and the corresponding eigenvectors $\boldsymbol{e_1} = (-0.051, -0.998, 0.029)^T, \boldsymbol{e_2} = (-0.574, 0.053, 0.817)^T$ and $\boldsymbol{e_3} = 0.817, -0.025, 0.575)^T$.
Then the half lengths of the ellipsoid are given by

$$\begin{aligned}\sqrt{\lambda_1}\sqrt{\frac{p(N-1)}{N(N-p)}F_{p,N-p}(a)} &= \sqrt{200.462}\sqrt{\frac{3\text{x}19}{20\text{x}17}F_{3,17}(0.1)} = 9.051\\ \sqrt{\lambda_2}\sqrt{\frac{p(N-1)}{N(N-p)}F_{p,N-p}(a)} &= \sqrt{4.532}\sqrt{\frac{3\text{x}19}{20\text{x}17}F_{3,17}(0.1)} = 1.361\\ \sqrt{\lambda_3}\sqrt{\frac{p(N-1)}{N(N-p)}F_{p,N-p}(a)} &= \sqrt{1.301}\sqrt{\frac{3\text{x}19}{20\text{x}17}F_{3,17}(0.1)} = 0.729\end{aligned}$$

i.e length=(18.102,2.722,1.458)
Furthermore, the individual simultaneous 90% confidence intervals for μ_1, μ_2 and μ_3 are
(a) The 90% confidence interval for μ_1

$$\left(\bar{X_1} - \sqrt{\frac{p(N-1)}{N(N-p)}F_{p,N-p}(a)S_{11}}\ ,\ \bar{X_1} + \sqrt{\frac{p(N-1)}{N(N-p)}F_{p,N-p}(a)S_{11}}\right) = (3.555, 5.725)$$

(b) The 90% confidence interval for μ_2

$$\left(\bar{X_2} - \sqrt{\frac{p(N-1)}{N(N-p)}F_{p,N-p}(a)S_{22}}\ ,\ \bar{X_2} + \sqrt{\frac{p(N-1)}{N(N-p)}F_{p,N-p}(a)S_{22}}\right) = (36.365, 54.435)$$

(c) The 90% confidence interval for μ_3

$$\left(\bar{X_3} - \sqrt{\frac{p(N-1)}{N(N-p)}F_{p,N-p}(a)S_{33}}\ ,\ \bar{X_1} + \sqrt{\frac{p(N-1)}{N(N-p)}F_{p,N-p}(a)S_{33}}\right) = (8.747, 11.183)$$

3. Question 4

Consider two populations π_1,π_2 with the following bivariate normal distributions

$$\pi_1 \sim \mathcal{N}_2\left(\boldsymbol{\mu_1} = \begin{pmatrix}10\\15\end{pmatrix}, \boldsymbol{\Sigma_1} = \begin{pmatrix}18 & 12\\12 & 32\end{pmatrix}\right)$$
$$\pi_2 \sim \mathcal{N}_2\left(\boldsymbol{\mu_2} = \begin{pmatrix}10\\25\end{pmatrix}, \boldsymbol{\Sigma_2} = \begin{pmatrix}20 & -7\\-7 & 5\end{pmatrix}\right)$$

Assume equal prior probabilities and misclassification costs of $c(2|1) = 10$ and $c(2|1) = 73.89$.

$(\boldsymbol{a})$ Write down explicitly the posterior probabilities $\mathbb{P}(\pi_1|\boldsymbol{x})$ and $\mathbb{P}(\pi_2|\boldsymbol{x})$ for an observation $\boldsymbol{x}$.

$\boldsymbol{Solution}$ **:**

This exercise concerns the multivariate techniques for discrimination and classification. The idea is to consider two populations π_1 and π_2 with corresponding probability distribution densities $f_1(x)$ and $f_2(x)$. Then we classify new data as belonging either to π_1 or to π_2. Therefore, we have the following assumptions

$$\text{Let } P(2|1) = \mathbb{P}(\text{ assign data to } \pi_2|\text{ data comes from } \pi_1) = \int_{R_2} f_1((x))\,dx$$
$$\text{Let } P(1|2) = \mathbb{P}(\text{ assign data to } \pi_1|\text{ data comes from } \pi_2) = \int_{R_1} f_2((x))\,dx$$
$$\text{Let } p_1 = \mathbb{P}(\text{data belongs to } \pi_1)$$
$$\text{Let } p_2 = \mathbb{P}(\text{data belongs to } \pi_2)$$

$$\text{Let } \mathbb{P}(\text{observation is correctly classified as } \pi_1) = P(1|1)p_1 = p_1 \int_{R_1} f_1((x))\,dx$$
$$\text{Let } \mathbb{P}(\text{observation is correctly classified as } \pi_2) = P(2|2)p_2 = p_2 \int_{R_2} f_2((x))\,dx$$
$$\text{Let } \mathbb{P}(\text{observation is incorrectly classified as } \pi_1) = P(1|2)p_2 = p_2 \int_{R_1} f_2((x))\,dx$$
$$\text{Let } \mathbb{P}(\text{observation is incorrectly classified as } \pi_2) = P(2|1)p_1 = p_1 \int_{R_2} f_1((x))\,dx$$

- Criterion 1: Minimize Expected Misclassification Cost (ECM)

Let $c(1|2)$ the cost of misclassifying observation from π_2.
Let $c(2|1)$ the cost of misclassifying observation from π_1.

$$\begin{aligned} ECM &= c(1|2)P(1|2)p_2 + c(2|1)P(2|1)p_1 \\ &= c(1|2)p_2 \int_{R_1} f_2((x))\,dx + c(2|1)p_1 \int_{R_2} f_1((x))\,dx \end{aligned}$$

Proposition 2. *The regions R_1, R_2 that minimise the ECM are defined by the values of $\boldsymbol{x}$, such that*

$$R_1 = \left\{ \boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c(1|2)p_2}{c(2|1)p_1} \right\} \tag{5}$$

and $R_2 = R_1^c$

- Criterion 2: Posterior Probability

Suppose we have an observation $\boldsymbol{x_0}$. The posterior probabilities as computed as follow.

$$\mathbb{P}(\pi_1|\boldsymbol{x_0}) = \frac{\mathbb{P}(\boldsymbol{x_0}|\pi_1)\mathbb{P}(\pi_1)}{\mathbb{P}(\boldsymbol{x_0}|\pi_1)\mathbb{P}(\pi_1) + \mathbb{P}(\boldsymbol{x_0}|\pi_2)\mathbb{P}(\pi_2)} \tag{6}$$

$$= \frac{p_1 f_1(\boldsymbol{x_0})}{p_1 f_1(\boldsymbol{x_0}) + p_2 f_2(\boldsymbol{x_0})} \tag{7}$$

and,

$$\mathbb{P}(\pi_2|\boldsymbol{x_0}) = 1 - \mathbb{P}(\pi_1|\boldsymbol{x_0}) = \frac{p_2 f_2(\boldsymbol{x_0})}{p_1 f_1(\boldsymbol{x_0}) + p_2 f_2(\boldsymbol{x_0})} \tag{8}$$

We would then assign the observation $\boldsymbol{x_0}$ to π_1 if $\mathbb{P}(\pi_1|\boldsymbol{x_0}) \geq \mathbb{P}(\pi_2|\boldsymbol{x_0})$.

Now in this exercise $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ are the $pdfs$ of multivariate normals with the given mean and variance-covariance matrices. Note that the pdf of a random vector $\boldsymbol{X}$ that has a multivariate normal distribution, i.e $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)\sqrt{det(\boldsymbol{\Sigma})}} exp\left\{ -\frac{1}{2}(\mathbf{X} - \mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mu) \right\}$$

Hence, for $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$,

$$\boldsymbol{\Sigma_1}^{-1} = \begin{pmatrix} 0.074 & -0.028 \\ -0.028 & 0.042 \end{pmatrix} \text{ and } det(\boldsymbol{\Sigma_1}) = 18\text{x}32 - 144 = 432.$$

$$\boldsymbol{\Sigma_2}^{-1} = \begin{pmatrix} 0.098 & 0.137 \\ 0.137 & 0.392 \end{pmatrix} \text{ and } det(\boldsymbol{\Sigma_2}) = 20\text{x}5 - 49 = 51.$$

Substituting in the pdf of the multivariate normal distribution we get (note that we assume that the vector $\boldsymbol{x} = (x_1, x_2)^T$)

$$\begin{aligned} f_1(\boldsymbol{x}) &= \frac{1}{2\pi\sqrt{432}} exp\left\{ -\frac{1}{2}(x_1 - 10, x_2 - 15) \begin{pmatrix} 0.074 & -0.028 \\ -0.028 & 0.042 \end{pmatrix} \begin{pmatrix} x_1 - 10 \\ x_2 - 15 \end{pmatrix} \right\} \\ &= \frac{1}{41.57\pi} exp\left\{ -\frac{1}{2}\left[0.074(x_1 - 10)^2 - 2(0.028)(x_1 - 10)(x_2 - 15) + 0.042(x_2 - 15)^2 \right] \right\} \end{aligned}$$

and

$$f_2(\boldsymbol{x}) = \frac{1}{2\pi\sqrt{51}} exp\left\{ -\frac{1}{2}(x_1 - 10, x_2 - 25) \begin{pmatrix} 0.098 & 0.137 \\ 0.137 & 0.392 \end{pmatrix} \begin{pmatrix} x_1 - 10 \\ x_2 - 25 \end{pmatrix} \right\}$$
$$= \frac{1}{14.28\pi} exp\left\{ -\frac{1}{2}\left[0.098(x_1 - 10)^2 + 2(0.137)(x_1 - 10)(x_2 - 25) + 0.392(x_2 - 25)^2 \right] \right\}$$

Therefore, the required posterior probabilities as given by equations (7) and (8) become

$$\mathbb{P}(\pi_1|\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})} \text{ and } \mathbb{P}(\pi_2|\boldsymbol{x}) = \frac{f_2(\boldsymbol{x})}{f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})} \tag{9}$$

since $p_1 = p_2$ with $f_1(\boldsymbol{x})$ and $f_1(\boldsymbol{x})$ as given above.

($\boldsymbol{b}$) Classify the set of observations $(10, 15)^T, (12, 17)^T, (14, 19)^T, (16, 21)^T, (18, 23)^T$

Solution **:**

To classify the given observations to regions R_1 and R_2 we use the posterior probability rule that we have denoted before. This could be done by substituting the coordinates of each vector into the $pdfs$ of $f_1(\boldsymbol{x})$ $f_2(\boldsymbol{x})$. This can be easily done in R using the following code. (Note the notation x_1, x_2 etc in the code denotes the vectors and not the coordinates of the vector)

```
#the given set of observations
x1<-c(10,15); x2<-c(12,17); x3<-c(14,19); x4<-c(16,21); x5<-c(18,23)

#the mean vectors and covariance matrices for the normals
mu1<-c(10,15); mu2<-c(10,25)
sigma1<-matrix(c(18,12,12,32),nrow=2)
sigma2<-matrix(c(20,-7,-7,5),nrow=2)

#Function to check in which region an observation lies
classification <- function(x,mu1,mu2,sigma1,sigma2)
  {#begin of function
     #evaluation of the multivariate normals at the given points
     f1<-dmnorm(x,mu1,sigma1)
     f2<-dmnorm(x,mu2,sigma2)

     posterior1<- f1 / (f1 + f2)
     posterior2<- f2 / (f1 + f2)

     if (posterior1 > posterior2)
      { message="It belongs to R1" }

     if (posterior1 <= posterior2)
      { message="It belongs to R2" }
     list(message=message)
  }#end of function

> classification(x1,mu1,mu2,sigma1,sigma2)
```

```
$message
[1] "It belongs to R1"

> classification(x2,mu1,mu2,sigma1,sigma2)
$message
[1] "It belongs to R1"

> classification(x3,mu1,mu2,sigma1,sigma2)
$message
[1] "It belongs to R1"

> classification(x4,mu1,mu2,sigma1,sigma2)
$message
[1] "It belongs to R2"

> classification(x5,mu1,mu2,sigma1,sigma2)
$message
[1] "It belongs to R2"
```

Summary of the above output is shown on table 3.

Table 3: Classification of observations

Observation	Region
$(10, 15)^T$	R_1
$(12, 17)^T$	R_1
$(14, 19)^T$	R_1
$(16, 21)^T$	R_2
$(18, 23)^T$	R_2

References

[1] Richard A. Johnson & Dean W. Wichern. (2007) *Applied Multivariate Statistical Analysis*. Prentice Hall.

[2] Nikos Zygouras. *Multivariate Statistics*. Lectures notes of ST412, University of Warwick,UK, unpublished.

[3] R Development Core Team. *R: A Language and Environment for Statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2011. URL $http://www.r-project.org/$.

Appendices

A. Question 1 - R output

```
> corr.new
          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 1.0000000 0.6996030 0.6063239 0.6375653 0.6517475 0.6334865
[2,] 0.6996030 1.0000000 0.3933842 0.4442873 0.4696102 0.4347072
[3,] 0.6063239 0.3933842 1.0000000 0.9168429 0.9133951 0.9213818
[4,] 0.6375653 0.4442873 0.9168429 1.0000000 0.9188683 0.9241262
[5,] 0.6517475 0.4696102 0.9133951 0.9188683 1.0000000 0.9221375
[6,] 0.6334865 0.4347072 0.9213818 0.9241262 0.9221375 1.0000000

> residual
           V1             V2           V3           V4          V5           V6
1  0.00000000 -1.946030e-01 -0.037323920 -0.035565293 -0.03074747 -0.03048646
2 -0.19460300  1.110223e-16  0.028615783  0.022712676  0.01238982  0.01529278
3 -0.03732392  2.861578e-02  0.000000000  0.009157101 -0.03639513 -0.04338177
4 -0.03556529  2.271268e-02  0.009157101  0.000000000 -0.04486829 -0.03012616
5 -0.03074747  1.238982e-02 -0.036395127 -0.044868292  0.00000000  0.01486252
6 -0.03048646  1.529278e-02 -0.043381774 -0.030126155  0.01486252  0.00000000
```

B. Question 3 - Code and R output

```
#Input of the data set
> data<-read.table("data1.txt", header=T)
>X1<-data1$X1
>X2<-data1$X2
>X3<-data1$X3

#The normal Q-plots for each observation
>qqnorm(X1,main="Normal Q-Q plot of variable X1")
>qqline(X1)

>qqnorm(X2,main="Normal Q-Q plot of variable X2")
>qqline(X2)

>qqnorm(X3,main="Normal Q-Q plot of variable X3")
>qqline(X3)

#Computation of the sample mean vector
>mean.vector<-matrix(0,1,3)
>mean.vector[1]<-mean(X1)
>mean.vector[2]<-mean(X2)
>mean.vector[3]<-mean(X3)

>mean.vector
     [,1] [,2]  [,3]
```

```
[1,] 4.64 45.4 9.965

#Computation of the variance-covariance matrix
> cov(data)
          X1       X2        X3
X1  2.879368  10.0100 -1.809053
X2 10.010000 199.7884 -5.640000
X3 -1.809053  -5.6400  3.627658

#Computation of the inverse variance-covariance matrix
> solve(cov(data))
            X1           X2           X3
X1  0.58615531 -0.022085719  0.257968742
X2 -0.02208572  0.006067227 -0.001580929
X3  0.25796874 -0.001580929  0.401846765

#Computation of the e-values and e-vectors of the variance-covariance matrix
> sigma<-cov(data)
> eigenvalues<-eigen(sigma)
> eigenvalues
$values
[1] 200.462464   4.531591   1.301392

$vectors
            [,1]        [,2]        [,3]
[1,] -0.05084144 -0.57370364  0.81748351
[2,] -0.99828352  0.05302042 -0.02487655
[3,]  0.02907156  0.81734508  0.57541452
```

C. Question 4 - Alternative method to classify the observations

An alternative approach to classify the observations would be to use the region given by proposition 2 and substitute the $pdfs$ of the multivariate normals. The 2 normal populations have different covariance matrices, i.e $\mathbf{\Sigma_1} \neq \mathbf{\Sigma_2}$. The region R_1 now becomes,

$$R_1 = \left\{ \boldsymbol{x} : -\frac{1}{2}\boldsymbol{x}^T(\mathbf{\Sigma_1}^{-1} - \mathbf{\Sigma_2}^{-1})\boldsymbol{x} + (\boldsymbol{\mu_1}^T\mathbf{\Sigma_1}^{-1} - \boldsymbol{\mu_2}^T\mathbf{\Sigma_2}^{-1})\boldsymbol{x} - k \geq ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right] \right\}$$

$$\text{where } k = \frac{1}{2}ln\left[\frac{det(\mathbf{\Sigma_1})}{det(\mathbf{\Sigma_2})}\right] + \frac{1}{2}(\boldsymbol{\mu_1}^T\mathbf{\Sigma_1}^{-1}\boldsymbol{\mu_1} - \boldsymbol{\mu_2}^T\mathbf{\Sigma_2}^{-1}\boldsymbol{\mu_2}).$$

This classification of the observations can be done in R using the following code. Note that we get the same result as with the previous method.

```
#the given set of observations
x1<-c(10,15); x2<-c(12,17); x3<-c(14,19); x4<-c(16,21); x5<-c(18,23)

mu1<-c(10,15); mu2<-c(10,25)
sigma1<-matrix(c(18,12,12,32),nrow=2)
```

```
sigma2<-matrix(c(20,-7,-7,5),nrow=2)
c21=10; c12=73.89; p1=1; p2=1


#Function to check in which region an observation lies
classification<-function(x, mu1, mu2, sigma1, sigma2, c12, c21, p1, p2)
 {#begin of function

 k<- 0.5*log(det(sigma1)/det(sigma2)) + 0.5*(mu1 %*% solve(sigma1) %*% mu1
     - mu2 %*% solve(sigma2) %*% (mu2) )
 left.ineq<- (-0.5*x%*%(solve(sigma1) - solve(sigma2))%*% (x) +
                (mu1 %*% solve(sigma1) - mu2 %*% solve(sigma2)) %*% (x) - k)
 right.ineq<-log( ((c12)*p1)/ ((c21)*p2) )
 condition<- ( left.ineq >= right.ineq )
 if (condition==TRUE)
  { message="TRUE" }
 else
  { message="FALSE" }
 list(sides.inequality=c(left.ineq,right.ineq), message=message)

 }#end of function

> classification(x1,mu1,mu2,sigma1,sigma2,c12,c21,1,1)
$sides.inequality
[1] 18.539543  1.999992

$message
[1] "TRUE"

> classification(x2,mu1,mu2,sigma1,sigma2,c12,c21,1,1)
$sides.inequality
[1] 9.360349 1.999992

$message
[1] "TRUE"

> classification(x3,mu1,mu2,sigma1,sigma2,c12,c21,1,1)
$sides.inequality
[1] 2.999238 1.999992

$message
[1] "TRUE"

> classification(x4,mu1,mu2,sigma1,sigma2,c12,c21,1,1)
$sides.inequality
[1] -0.5437902  1.9999924

$message
[1] "FALSE"
```

```
> classification(x5,mu1,mu2,sigma1,sigma2,c12,c21,1,1)
$sides.inequality
[1] -1.268736  1.999992

$message
[1] "FALSE"
```