P-set4(s23) (Please do not circulate or post)

1. (a) For the first variable, the decomposition $\boldsymbol{x}_{tj} = \bar{\boldsymbol{x}} + (\bar{\boldsymbol{x}}_t - \bar{\boldsymbol{x}}) + (\boldsymbol{x}_{tj} - \bar{\boldsymbol{x}}_t)$ is

$$\begin{bmatrix} 6 & 5 & 8 & 4 & 7 \\ 3 & 1 & 2 & & \\ 2 & 5 & 3 & 2 & \end{bmatrix} = \begin{bmatrix} 4 & 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & & \\ 4 & 4 & 4 & 4 & \end{bmatrix} + \begin{bmatrix} 2 & 2 & 2 & 2 & 2 \\ -2 & -2 & -2 & & \\ -1 & -1 & -1 & -1 & \end{bmatrix} + \begin{bmatrix} 0 & -1 & 2 & -2 & 1 \\ 1 & -1 & 0 & & \\ -1 & 2 & 0 & -1 & \end{bmatrix}$$

For the second variable, $\begin{bmatrix} 7 & 9 & 6 & 9 & 9 \\ 3 & 6 & 3 & & \\ 3 & 1 & 1 & 3 & \end{bmatrix} = \begin{bmatrix} 5 & 5 & 5 & 5 & 5 \\ 5 & 5 & 5 & & \\ 5 & 5 & 5 & 5 & \end{bmatrix} + \begin{bmatrix} 3 & 3 & 3 & 3 & 3 \\ -1 & -1 & -1 & & \\ -3 & -3 & -3 & -3 & \end{bmatrix} + \begin{bmatrix} -1 & 1 & -2 & 1 & 1 \\ -1 & 2 & -1 & & \\ 1 & -1 & -1 & 1 & \end{bmatrix}$

(b) From the $(\bar{\boldsymbol{x}}_t - \bar{\boldsymbol{x}})$ part in (a), the between group matrix of sum of squares and cross products is

$$\boldsymbol{B} = \sum_{t=1}^{3} n_t(\bar{\boldsymbol{x}}_t - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_t - \bar{\boldsymbol{x}})' = 5\begin{bmatrix} 2 \\ 3 \end{bmatrix}[2\ 3] + 3\begin{bmatrix} -2 \\ -1 \end{bmatrix}[-2\ -1] + 4\begin{bmatrix} -1 \\ -3 \end{bmatrix}[-1\ -3] = \begin{bmatrix} 36 & 48 \\ 48 & 84 \end{bmatrix}$$

From the $(\boldsymbol{x}_{tj} - \bar{\boldsymbol{x}}_t)$ part in (a), the within group matrix of sum of squares and cross products is

$$\boldsymbol{W} = \sum_{t=1}^{3}\sum_{j=1}^{n_t}(\boldsymbol{x}_{tj} - \bar{\boldsymbol{x}}_t)(\boldsymbol{x}_{tj} - \bar{\boldsymbol{x}}_t)' = \begin{bmatrix} 0 \\ -1 \end{bmatrix}[0\ -1] + \begin{bmatrix} -1 \\ 1 \end{bmatrix}[-1\ 1] + \cdots + \begin{bmatrix} -1 \\ 1 \end{bmatrix}[-1\ 1] = \begin{bmatrix} 18 & -13 \\ -13 & 18 \end{bmatrix}$$

MANOVA Table

Source of Variation	Sum of Squares and CrossProduct	d.f.
Treatment	$\boldsymbol{B} = \begin{bmatrix} 36 & 48 \\ 48 & 84 \end{bmatrix}$	$g - 1 = 2$
Residual	$\boldsymbol{W} = \begin{bmatrix} 18 & -13 \\ -13 & 18 \end{bmatrix}$	$n - g = 9$
Total	$\boldsymbol{B} + \boldsymbol{W} = \begin{bmatrix} 54 & 35 \\ 35 & 102 \end{bmatrix}$	$n - 1 = 11$

The covariance matrices can be verified by R.

```
# R command for Q1 (b)
y1=c(6,5,8,4,7, 3,1,2, 2,5,3,2); y2=c(7,9,6,9,9, 3,6,3, 3,1,1,3)
trt=as.factor(c(1,1,1,1,1,2,2,2,3,3,3,3)); y=cbind(y1,y2)
11*cov(manova(y~trt)$fitted)       # B matrix
11*cov(manova(y~trt)$residual)     # W matrix
11*cov(y)                          # B+W
```

(c) $|\boldsymbol{B}| = det(\boldsymbol{B}) = 720, |\boldsymbol{W}| = 155, |\boldsymbol{B} + \boldsymbol{W}| = 4283.$
The Wilks' test concludes that the differences in treatment effects are significant:

$$Wilks'\ \Lambda^* = \frac{|\boldsymbol{W}|}{|\boldsymbol{B} + \boldsymbol{W}|} = 0.0362$$

(d) Bartlett's approximation yields p-value $0.000113 < 0.0001$ (`pchisq(28.209,4)= 0.9999887`).

$$-\left(n - 1 - \frac{p+q}{2}\right)\ln \Lambda^* = 28.209 \sim \chi^2_4 \ > \ \chi^2_4(0.01) = 13.28$$

The conclusion: treatment differences are significant at test level $<< 0.01$.

(e) In (c) and (d), the hypothesis test is:

$$\begin{cases} H_0: & \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 \quad \text{— all three treatment mean vectors are equal} \\ H_a: & \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \text{ for some } i \neq j (i, j = 1, 2, 3) \quad \text{— at least one pair of treatment mean vectors are not equal} \end{cases}$$

2. (a) $p = 4, g = 3, n_i = 30, i = 1, 2, 3$. MANOVA table:

```
          Df  Wilks approx F num Df den Df  Pr(>F)
time       2 0.8301   2.0491      8    168 0.04358 *
Residuals 87
```

Wilks' $\Lambda^* = .8301$, test statistic (J&W p.303) $\frac{\sum n_i - p - 2}{p}\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}} = 2.0491 \sim F_{8,168}$ under the null of equal means.
With p-value = 0.04358, the null of equal means in all periods is barely rejected at test level of 0.05.

```
#=== Q2 Part a ===
> skull=read.table("T6-13.DAT"); colnames(skull)=c("x1","x2","x3","x4","period"); attach(skull)
>  y=cbind(x1,x2,x3,x4); time = as.factor(period)
> summary(manova(y~time),test="Wilks")   # default; p=.04358
> (84/4)*(1-sqrt(.8301))/sqrt(.8301)  #2.0491
> 1-pf(2.0491,df1=8,df2=168)  # 0.04357915
```

(b) Comparing pairs of periods by Hotelling's T^2: T he measures in Period 3 are significantly different from the measures in the other two periods.

$$Under\ H_0: \boldsymbol{\mu}_i = \boldsymbol{\mu}_j, \qquad T^2 \sim \frac{(n_i + n_j - 2)p}{n_i + n_j - p - 1}F_{p,n_1+n_2-p-1} = \frac{232}{55}F_{4,55}$$

Period 1 vs 2: p-value = 0.814; Period 1 vs 3: p-value = 0.020; Period 2 vs 3: p-value = 0.019
Measurements in Period 3 is more different from that of periods 1 and 2. Boxplots also reveal the difference. (omitted)

```
#======Q2 Part b (partial R code) =======#
# Hotelling T2 pairwise between periods
> summary(manova(y[1:60,]~time[1:60]),test="Hotelling")                   # Periods 1 vs 2
> summary(manova(y[c(1:30,61:90),]~time[c(1:30,61:90)]),test="Hotelling")  # Periods 1 vs 3
> summary(manova(y[31:90,]~time[31:90]),test="Hotelling")                  # Periods 2 vs 3
```

(c) i. There are $pg(g-1)/2 = 12$ simultaneous confidence intervals to be constructed. For each variable x_i, there are $g(g-1)/2 = 3 \times (3-1)/2 = 3$ simultaneous C.I.'s (periods 1 vs 2, 2 vs 3, 3 vs 1).

ii. Note that the measurements from different periods are independent. For component i, the confidence interval for the difference of the mean vectors between independent samples of periods 1 and 2 has the form

$$\bar{x}_{1i} - \bar{x}_{2i} \pm c\sqrt{\widehat{var}(\bar{x}_{1i} - \bar{x}_{2i})}, \qquad (i = 1, 2, 3, 4)$$

with

$$\widehat{var}(\bar{x}_{1i} - \bar{x}_{2i}) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)s_{ii} = \left(\frac{1}{30} + \frac{1}{30}\right)s_{ii}$$

where $s_{ii} = w_{ii}/(n-g) = 20.52, 22.12, 24.75, 9.66$ are the diagonal elements in the sample covariance matrix pooled over three independent samples from the three periods. $S_{pool} = \frac{(n_1-1)S_1+(n_2-1)S_2+(n_3-1)S_3}{n-g}$.
As for the multiplier c, at test level $\alpha = 1 - 0.85 = 0.15$, the multiplier for 85% simultaneous Bonferroni confidence intervals is

$$c = t_{n-g}(\alpha/(pg(g-1))) = t_{90-3}(0.15/24) = t_{87}(0.00625) = 2.55$$

```
#====== Q2 (c) =======#
Spool = 29*(cov(y[1:30,])+cov(y[31:60,])+cov(y[61:90,]))/(90-3)
round(diag(Spool),2)    # 20.52   22.12   24.75   9.66   = s_ii
c=qt(1-.15/24, df=90-3)  # 2.550697 = t-quantile
sd = sqrt(diag(Spool)/15)   # 1.17 1.21 1.28 0.80  = s.e.
```

3. The data `basketball.csv` contains measurements on 54 basketball players. Use `(Field, Freethrow, Avgpt)` jointly as the response vector, use regressions to analyze and interpret.

(a) Fit a multivariate linear regression model. Check the residuals. The residuals of which two response variables are mostly correlated?

- Consider the following models with $Y =$ `(Field, Freethrow, Avgpt)` as the response vector, `Height` and `Weight` as explanatory variables:
`Y ~ Height + Weight + Height*Weight; Y ~ Height + Weight; Y ~ Height; Y ~ Weight`
- The explanatory variables do a moderate and limited job in explaining the response variables, as the existence of some large residuals indicates.
The multivariate regression model does retain the correlation structure of the dependence variables. The strengths of the association between the responses with the explanatory variables vary, from strong (Field goal) to weak (Avg pts).
- The interaction model gives an estimate of the covariance matrix Σ. The correlation matrix below shows moderate correlation among the (residuals of the) response variables.
The most correlated pair are Average points (Avgpt) and Field goals (Field).

```
fullfit=lm(cbind(Field,Freethrow,Avgpt)~Height*Weight,data=basket); round(cor(fullfit$residuals),3)
```

The main effects model:

```
> mfit = lm(cbind(Field,Freethrow,Avgpt)~Height+Weight,data=basket)
> summary(mfit)
```

The estimated Σ is

```
> round(cov(mfit$residuals),3)
          Field Freethrow  Avgpt
Field     0.002     0.001  0.119
Freethrow 0.001     0.009  0.141
Avgpt     0.119     0.141 34.376
```

The estimated correlation matrix:

```
> round(cor(mfit$residuals),3)
          Field Freethrow Avgpt
Field     1.000     0.166 0.423
Freethrow 0.166     1.000 0.251
Avgpt     0.423     0.251 1.000
```

(b) Do a (sequential) analysis of variance. Construct two MANOVA tables with different orders of the explanatory variables. Which variable or variables are important? (Hint: Check the correlations of the explanatory variables.)

Multivariate analysis of variance (MANOVA) of the main effect model and interaction model.

It appears either Height or Weight is significant, if the variable is entered first.

The order of input variables matters in MANOVA, which indicates possible strong correlation of the input variables.

Indeed input variables Height and Weight are strongly correlated. Thus only one variable appears significant.

```
> summary(manova(cbind(Field,Freethrow,Avgpt)~Height*Weight,data=basket))
> summary(manova(cbind(Field,Freethrow,Avgpt)~Weight*Height,data=basket))
> summary(manova(cbind(Field,Freethrow,Avgpt)~Height+Weight,data=basket))
          Df  Pillai approx F num Df den Df    Pr(>F)
Height     1 0.35844   9.1256      3     49 6.646e-05 ***
Weight     1 0.07500   1.3244      3     49    0.2771
Residuals 51
> summary(manova(cbind(Field,Freethrow,Avgpt)~Weight+Height,data=basket))
          Df  Pillai approx F num Df den Df    Pr(>F)
Weight     1 0.36912   9.5566      3     49 4.464e-05 ***
Height     1 0.05186   0.8935      3     49    0.4512
Residuals 51
```

(Not required) The data set is small enough to be plotted pairwise, which shows strong correlation between the explanatory variables (Height and Weight), and relatively weak correlation among responses. Only the response variable `Field` seems to have a strong correlation with the explanatory variables `Hight` and `Weight`.

(Not required) Due to the strong correlation between the explanatory variables Height and Weight, the model with Height or Weight as the only input is also plausible.

Try the model with Height (Weight) as input variable only:

```
summary(lm(cbind(Field,Freethrow,Avgpt)~Height,data=basket))
summary(lm(cbind(Field,Freethrow,Avgpt)~Weight,data=basket))
```

4. (a) The coordinates of the sites given $q = 5$ dimensions using MDS. Notice that MDS $q = 3$ gives the first three columns as in $q = 5$, and so on. (Command: `round(cmdscale(as.dist(X),q),2)`)

```
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,]  0.51 -0.28  0.24  0.68  0.12
[2,] -1.32  0.69  0.62  0.05 -0.02
[3,]  0.47 -0.07  0.19 -0.30  0.06
[4,]  0.39  0.09  0.05 -0.34  0.10
[5,]  0.23  0.30 -0.33 -0.05  0.12
[6,]  0.47  0.14 -0.22  0.14 -0.28
[7,]  0.58 -0.35  0.46 -0.18 -0.10
[8,] -1.12 -1.12 -0.32 -0.05 -0.01
[9,] -0.22  0.61 -0.70  0.06  0.01
```

(b) The plots of minimum stress against multidimensional scaling dimension q is given below. (here omitted). The stress decreases with increasing q, and the stress of final configuration for $q = 5$ is 0.000. (both stress and stress-square are acceptable.) The following code is used to calculate Stress and to produce the plots in (b) and (c).

```
StressT12=rep(0,8)
for (i in 1:8)
{
  fit = cmdscale(as.dist(X),i); fitS2 = sum(dist(fit)^2)
  diffS2 = sum((as.dist(X)-dist(fit))^2); StressT12[i]=sqrt(diffS2/fitS2)
}
# round(StressT12,4)
par(mfrow=c(1,2)); plot(1:7,StressT12[1:7],type="b"); title("Archaeologic site MDS Stress dim=1:7",cex.main=.6)
plot(2:5,StressT12[2:5],type="b"); title("Archaeologic site MDS Stress dim=2:5",cex.main=.6)
q=5; par(mfrow=c(1,1))
plot(cmdscale(as.dist(X),q)[,1],cmdscale(as.dist(X),q)[,2],type="n",xlab="mds1",ylab="mds2",cex.axis=.7,cex.lab=.7)
text(cmdscale(as.dist(X),q)[,1],cmdscale(as.dist(X),q)[,2],pch=1:9,cex=.9,lwd=2,col=4)
```

(c) The plot of first two coordinates of $q = 5$ dimensional solution against each other is given below. (here omitted) The results hint a time pattern following the counter-clockwise direction of the circle in the plot, with a couple of exceptions.

(The exact interpretation related to time pattern is not required)

5. (a) The data is a contingency table

```
      Brown Blue Hazel Green
Black    68   20    15     5
Brown   119   84    54    29
Red      26   17    14    14
Blond     7   94    10    16
```

The cell percentage table below shows that the Brown-Brown combination is most common. (Output omitted)
The row percentage table below gives a set of comparable $\mathbb{R}^4$ coordinates of Hair variable. (Output omitted)
The column percentage table below gives a set of comparable $\mathbb{R}^4$ coordinates of Eye variable. (Output omitted)

(b) The expected counts under independence in the table below are quite far from the original data table, indicating dependence between the row (Hair) and column (Eye) variables.

```
      Brown Blue Hazel Green
Black    40   39    17    12
Brown   106  104    45    31
Red      26   26    11     8
Blond    47   46    20    14
```

(c) Below the table of cell *mass* $(x_{ij} - E_{ij})^2/E_{ij}$ (cell contribution to the Chi-square statistic) displays cell level deviations from independence, some are quite large.

```
      Brown  Blue Hazel Green
Black 19.35  9.42  0.23  3.82
Brown  1.52  3.80  1.83  0.12
Red    0.01  2.99  0.73  5.21
Blond 34.23 49.70  4.96  0.38
```

The Chi-squared statistic = 138.29 (df = 9) wih p-value ≈ 0 under the nulll hypothesis that Hair and Eye colors are independent. The data provides evidence that Hair and Eye colors are strongly correlated. The test statistic 138.29 = "total mass" = the sum of all entries in the above table of cell *mass*. "Total inertia" is defined as $\chi^2/n = 138.29/592 = 0.23$.

(d) The plot of correspondence analysis and interpretations (partially shown) are shown below.

Overall: The CA plot explains 99% variation in the data, thus the 2-d plot displays the table reasonably well.

Individual association: 63% Brown eye are of Black hair, the strong pairing (association) is reflected by their closeness in the CA plot.

(Not required) Row-column profiles: In hair variables (blue solid circles), Brown and Red hair have similar eye color profiles (similar row percentages), reflected in their closeness in row direction. Eye variables (red triangles) are apart in column direction, result from different hair profiles (column percentages) for different Eye colors.

```
  #=== R code for Q5 ===
data = read.table("HairEyeAll.txt")
rownames(data) = c("Black","Brown", "Red", "Blond")   # for variable Hair
colnames(data)= c("Brown","Blue", "Hazel", "Green")   # for variable Eye
X = as.matrix(data)
round(X/sum(X),2); round(diag(c(1/X%*%c(1,1,1,1)))%*%X, 2); round(X%*%diag(c(1/(c(1,1,1,1)%*%X))), 2)# Q5 (a)
E=(X%*%c(1,1,1,1))%*%(c(1,1,1,1)%*%X)/sum(X); round(E)       # Q5 (b)
# Q5(c)
round((X-E)^2/E, 2); chisq.test(X)   # Q5(d)
library(ca); ca(X); plot(ca(X),mass=c(TRUE,TRUE))
```

6. (a) i. Derive the conditional expectation $\mathbb{E}(\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \boldsymbol{x}_2, \boldsymbol{X}_3 = \boldsymbol{x}_3)$.

$$
\begin{aligned}
\mathbb{E}(\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \boldsymbol{x}_2, \boldsymbol{X}_3 = \boldsymbol{x}_3) &= \boldsymbol{\mu}_1 + [\Sigma_{12}\ \Sigma_{13}] \begin{bmatrix} \Sigma_{22} & \mathbf{0} \\ \mathbf{0} & \Sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x}_2 - \boldsymbol{\mu}_2 \\ \boldsymbol{x}_3 - \boldsymbol{\mu}_2 \end{bmatrix} \\
&= \boldsymbol{\mu}_1 + [\Sigma_{12}\ \Sigma_{13}] \begin{bmatrix} \Sigma_{22}^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{33}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_2 - \boldsymbol{\mu}_2 \\ \boldsymbol{x}_3 - \boldsymbol{\mu}_2 \end{bmatrix} \\
&= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) + \Sigma_{13}\Sigma_{33}^{-1}(\boldsymbol{x}_3 - \boldsymbol{\mu}_3)
\end{aligned}
$$

ii. Derive the conditional variance $Var(\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \boldsymbol{x}_2, \boldsymbol{X}_3 = \boldsymbol{x}_3)$.

$$
Var(\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \boldsymbol{x}_2, \boldsymbol{X}_3 = \boldsymbol{x}_3) = \Sigma_{11} - [\Sigma_{12}\ \Sigma_{13}] \begin{bmatrix} \Sigma_{22} & \mathbf{0} \\ \mathbf{0} & \Sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{21} \\ \Sigma_{31} \end{bmatrix} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{31}
$$

(b) Since $\boldsymbol{X}_2 \sim N(\boldsymbol{\mu}_2, \Sigma_{22})$, $\boldsymbol{X}_3 \sim N(\boldsymbol{\mu}_3, \Sigma_{33})$, and $\boldsymbol{X}_2 \perp\!\!\!\perp \boldsymbol{X}_3$, assuming $\boldsymbol{X}_2, \boldsymbol{X}_3$ are of the same dimension,

$$
\boldsymbol{X}_2 + \boldsymbol{X}_3 \sim N(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_3, \Sigma_{22} + \Sigma_{33}).
$$

Furthermore,

$$
Cov(\boldsymbol{X}_1, \boldsymbol{X}_2 + \boldsymbol{X}_3) = Cov(\boldsymbol{X}_1, \boldsymbol{X}_2) + Cov(\boldsymbol{X}_1, \boldsymbol{X}_3) = \Sigma_{12} + \Sigma_{13}, \qquad Cov(\boldsymbol{X}_2 + \boldsymbol{X}_3, \boldsymbol{X}_1) = \Sigma_{21} + \Sigma_{31}
$$

So $\boldsymbol{X}_1$ and $\boldsymbol{X}_2 + \boldsymbol{X}_3$ have joint normal distribution with mean and covariance

$$
\boldsymbol{\mu}^* = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 + \boldsymbol{\mu}_3 \end{bmatrix}, \qquad \boldsymbol{\Sigma}^* = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} + \Sigma_{13} \\ \Sigma_{21} + \Sigma_{31} & \Sigma_{22} + \Sigma_{33} \end{bmatrix}
$$

By the multivariate normal property, the conditional distribution of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2 + \boldsymbol{X}_3 = \boldsymbol{x}_0$ is also multivariate normal, with mean

$$
\boldsymbol{\mu}_1 + (\Sigma_{12} + \Sigma_{13})(\Sigma_{22} + \Sigma_{33})^{-1}[\boldsymbol{x}_0 - (\boldsymbol{\mu}_2 + \boldsymbol{\mu}_3)]
$$

and variance-covariance matrix

$$
\Sigma_{11} - (\Sigma_{12} + \Sigma_{13})(\Sigma_{22} + \Sigma_{33})^{-1}(\Sigma_{21} + \Sigma_{31}).
$$