

Multivariate Data Analysis: Tasks for Week 5

Notes & Solutions

1) Continuing Q1 of task sheet week 4, i.e. X' ($n \times p$) is a centred data matrix:

- i) If D is the $n \times n$ distance matrix of the n p -dimensional observations and A is the matrix given by $a_{ij} = -\frac{1}{2}d_{ij}^2$ and $B=HAH$, where H is the centring matrix, shew that $B=kX'X$ for some suitable scalar k .

$$d_{ij}^2 = (x_i - x_j)'(x_i - x_j) = x_i'x_i - 2x_i'x_j + x_j'x_j \quad \text{so} \quad b_{ij} = a_{ij} - \bar{a}_{i+} - \bar{a}_{+j} + \bar{a}_{++} = x_i'x_j$$

noting that $\bar{x}_i = 0$ so $B = X'X$.

- ii) Deduce that deriving a configuration of points from the matrix D by classical scaling is equivalent to referring the original data to principal components

If the data are referred to principal components then the coordinates of the rotated points are $X'A$ where $A=(a_i)$ are the eigenvectors of $S=(n-1)^{-1}XX'$. If we calculate the distance matrix directly from X' , then the principal coordinates from this distance matrix are given by the eigenvectors of $X'X$ which are (from Q1 of week 4) $X'a_i$, i.e. $X'A$, thus showing that deriving a configuration of points from the matrix D by classical scaling is equivalent to referring the original data to principal components.



2) If c_{ij} represents the similarity between cases i and j (c_{ij} is a similarity if $c_{ij}=c_{ji}$ and $c_{ij} \leq c_{ii}$) then the similarity matrix C can be converted to a distance matrix D by defining $d_{ij}=(c_{ii}-2c_{ij}+c_{jj})^{1/2}$. Define $B = HAH$ where $A=(-1/2 d_{ij}^2)$

i) Shew that $B=HCH$.

If $d_{ij} = (c_{ii}-2c_{ij}+c_{jj})^{1/2}$ then $a_{ij} = -1/2 d_{ij}^2 = -1/2(c_{ii}-2c_{ij}+c_{jj})$ so $b_{ij} = a_{ij} - \bar{a}_{i+} - \bar{a}_{+j} + \bar{a}_{++} = c_{ij} - \bar{c}_{i+} - \bar{c}_{+j} + \bar{c}_{++}$ and so $B = HCH$.

ii) Deduce that you can proceed with classical scaling analysis analyzing C directly instead of converting it to a distance matrix and then calculating A .

So, we deduce that you can proceed with classical scaling analysis using C directly in place of the matrix A instead of converting C to a distance matrix and then calculating A from it.



3) The table below gives the road distances between 12 UK towns. The towns are Aberystwyth, Brighton, Carlisle, Dover, Exeter, Glasgow, Hull, Inverness, Leeds, London, Newcastle and Norwich.

i) Is it possible to construct an exact map for these distances?

	A	B	C	D	E	G	H	I	Le	Lo	Ne	No
A	0											
B	244	0										
C	218	350	0									
D	284	77	369	0								
E	197	167	347	242	0							
G	312	444	94	463	441	0						
H	215	221	150	236	279	245	0					
I	469	583	251	598	598	169	380	0				
Le	166	242	116	257	269	210	55	349	0			
Lo	212	53	298	72	170	392	168	531	190	0		
Ne	253	325	57	340	359	143	117	264	91	273	0	
No	270	168	284	164	277	378	143	514	173	111	256	0

These data are contained in data set TOWNS. The Minitab version has the names of the towns in the first column and the data matrix in the next 12 columns. The final 12 columns contain the 12×12 matrix $A = (-\frac{1}{2}d_{ij}^2)$. The R and S-plus versions give a dataframe with just the symmetric matrix of distances.



The key to this is to calculate the eigenanalysis of the centred version of A (i.e. $B=HAH$). This gives the values 394473, 63634, 13544, 10246, –7063, 2465, 1450, –1141, 500, –214, –17 and since some of these are negative it means no it is not possible to construct an **exact** map. A transcript of an **R** session to do this is given below.

In **R** use the function `cmdscale()` (use the help system to find out how).

A transcript is given below. Note that `towns.Rdata` is a dataframe and so the function `as.matrix()` is required to convert this to a matrix.

```
> options(digits=3)
> x<-cmdscale(as.matrix(towns),k=11,eig=TRUE)
Warning messages:
1: In cmdscale(as.matrix(towns), k = 11, eig = TRUE) :
  some of the first 11 eigenvalues are < 0
2: In sqrt(ev) : NaNs produced
> x$eig
 [1] 3.94e+05 6.36e+04 1.35e+04 1.02e+04 2.46e+03 1.45e+03
 [7] 5.01e+02 -9.09e-13 -1.69e+01 -2.14e+02 -1.14e+03
>
```

Note that it warns you of negative eigenvalues.



Multivariate Data Analysis: Tasks for Week 6

Notes & Solutions

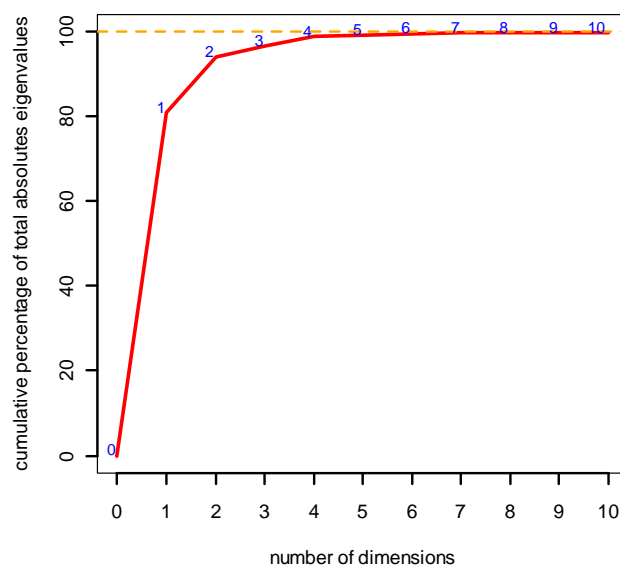
1) Continuing Q3 of the tasks for week 5 (road distances between 12 UK towns)

Determine a configuration of points that will adequately represent the data.

First construct a scree plot (or else eyeball the [positive] eigenvalues and observe the first two dominate, so 2 dimensions is enough).

```
> library(MASS)
> options(digits=3)
> x<-cmdscale(as.matrix(towns),k=11,eig=TRUE)
Warning messages:
1: In cmdscale(as.matrix(towns), k = 11, eig = TRUE) :
  some of the first 11 eigenvalues are < 0
2: In sqrt(ev) : NaNs produced
>
> x$eig
[1] 3.94e+05 6.36e+04 1.35e+04 1.02e+04 2.46e+03 1.45e+03
[7] 5.01e+02 -9.09e-13 -1.69e+01 -2.14e+02 -1.14e+03
>
> CMDscreeplot(towns,raw=FALSE,abs=TRUE,maxcomp=10)
Warning messages:
1: In cmdscale(as.matrix(mydata), k = n, eig = TRUE) :
  some of the first 11 eigenvalues are < 0
2: In sqrt(ev) : NaNs produced
```

Scree plot of absolute values of eigenvalues



- i) Construct a two-dimensional map representing the road distances between these towns.

To do this in Minitab you need to use the code in the course notes (changing 10 to 12 in places) do get the centring matrix and then operate on the matrix obtained by copying the final 12 columns.

To do it in **R** you can use the function `cmdscale()` with

```
x<-cmdscale(as.matrix(towns))
```

and then plot the results by the coordinates of the points are in `x$points` and can be plotted. Note that `towns.Rdata` is already a distance matrix so you should not use the multidimensional scaling menu in the MASS library which presumes that you have a raw data matrix and calls `dist()` internally to create a new distance matrix. The command `as.matrix()` is required for `cmdscale` to recognise the distance matrix. It is also possible in **R** to try two varieties of non-metric scaling (`sammon()` and `isomds()`).

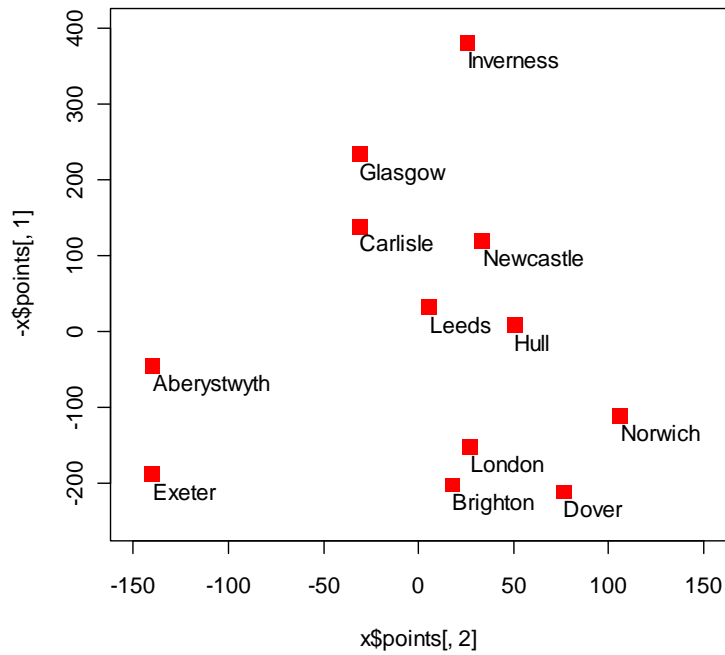
A record of an **R** session to perform these analyses is given below:

```
> plot(x$points[,2],-x$points[,1],pch=15,col="red",
+ xlim=c(-150,150),ylim=c(-250,400),cex=1.5,
+ main="Classical Metric Scaling
+ plot of UK inter-town road distances")
>
> text(x$points[,2],-
x$points[,1],row.names(towns),adj=c(0,1.3),
+ xlim=c(-150,150),ylim=c(-250,400))
>
```

Note the reversal of sign of the vertical axis — an initial plot revealed that Inverness appeared on the lower edge of the plot so re-plotting with the sign changed produces a plot more aligned to the geography of the UK. Fortunately the east-west axis came out correctly but it would be possible to rotate or flip the plot in any way that is required.

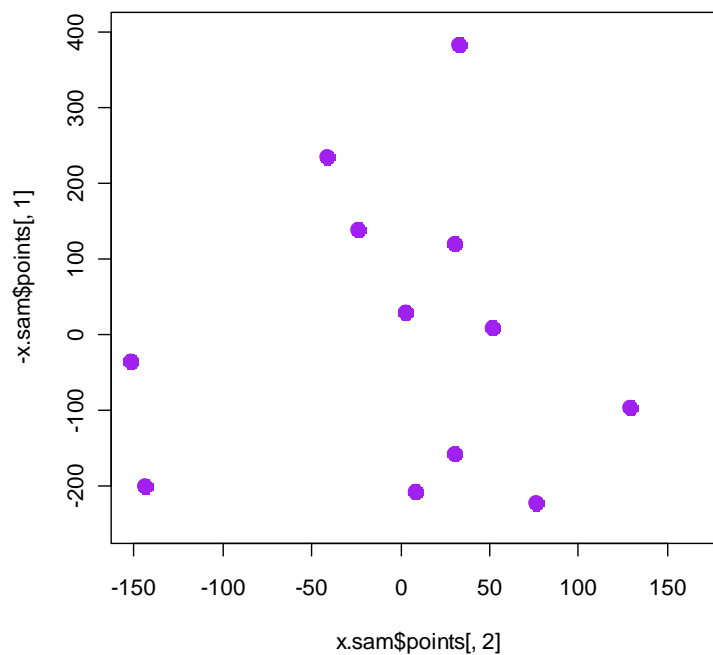


Classical Metric Scaling
plot of UK inter-town road distances



```
> plot(x.sam$points[, 2],
+ -x.sam$points[, 1], pch=19, col="purple",
+ xlim=c(-150, 170), ylim=c(-250, 400), cex=1.5,
+ main="Sammon Mapping
+ plot of UK inter-town road distances")
>
```

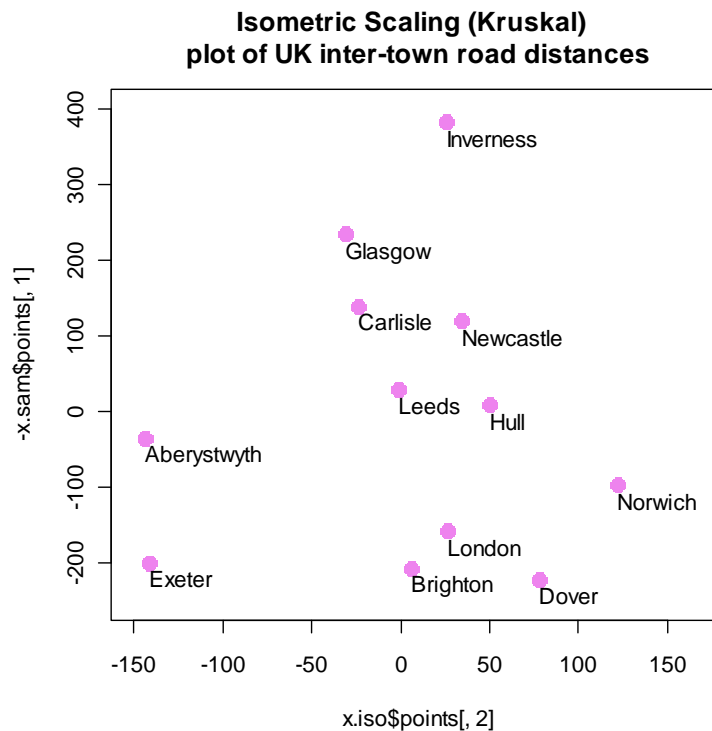
Sammon Mapping
plot of UK inter-town road distances



```

> plot(x.iso$points[,2],-
x.sam$points[,1],pch=16,col="violet",
+ xlim=c(-150,170),ylim=c(-250,400),cex=1.5,
+ main="Isometric Scaling (Kruskal)
+ plot of UK inter-town road distances")
>
>
> text(x.iso$points[,2],-x.sam$points[,1],
+ labels=row.names(towns),adj=c(0,1.3),xlim=c(-
150,170),ylim=c(-250,400))
>

```



- 2) Retrieve the data on beef and pork consumption referenced in §5.2 and verify the calculations given in §5.2 using **R** or **S-PLUS**. Predict the consumption of beef and pork if the prices in cents/lb are 79.3, 41.2 and the disposable income index is 40.4.

First, create a data set with the name `meat` containing the six columns needed (including a column named `constant` containing just 1s), then:

```

> attach(meat)
> y<- cbind(cbe,cpo)
> x<- cbind(constant, pbe,ppo,dinc)
> betahat<- solve(t(x)%*%x)%*%t(x)%*%y

```



```

> betahat
      cbe    cpo
constant 101.448 79.569
pbe      -0.753  0.153
ppo       0.254 -0.687
dinc     -0.241  0.283>
> sigmahat<-t(y-x**betahat)**(y-x**betahat)
> sigmahat<-sigmahat/(17-3-1)
> sigmahat
      cbe    cpo
cbe   4.41 -7.57
cpo  -7.57 16.83>
> x0<- c(1,79.3,41.2,40.4)
> ypred<-x0**betahat
> ypred
      cbe    cpo
[1,] 42.5 74.9
>>

```

So predicted consumption of beef is 42.5 and of pork 74.9 pounds.

- 3) Retrieve the dataset *chap8headsize* referenced in §6.3 and calculate the estimates of the least squares multivariate regression parameters β of length and breadth of heads of first sons upon those of second sons. Is it possible to deduce from these results the estimates for the regression of second sons upon the first? (Note that the individual data files seem no longer to be available but you should be able to download the complete set of files from Brian Everitt's webpage as a zipped archive.)

```

> "headsize" <-
+ matrix(c(191, 195, 181, 183, 176, 208, 189, 197, 188, 192, 179, 183,
174, 190, 188, 163, 195, 186, 181, 175, 192, 174,
+ 176, 197, 190, 155, 149, 148, 153, 144, 157, 150, 159, 152, 150, 158,
147, 150, 159, 151, 137, 155, 153,
+ 145, 140, 154, 143, 139, 167, 163, 179, 201, 185, 188, 171, 192, 190,
189, 197, 187, 186, 174, 185, 195,
+ 187, 161, 183, 173, 182, 165, 185, 178, 176, 200, 187, 145, 152, 149,
149, 142, 152, 149, 152, 159, 151,
+ 148, 147, 152, 157, 158, 130, 158, 148, 146, 137, 152, 147, 143, 158,
150)
+ , nrow = 25, ncol = 4 , dimnames = list(character(0)
+ , c("head1", "breadth1", "head2", "breadth2")))
> attach(data.frame(headsize))

```



The calculations for the regression of first son sizes on second son sizes are:-

```
> y<-cbind(head1,breadth1)
> x<-cbind(rep(1,25),head2,breadth2)
> betahat<- solve(t(x)%*%x)%*%t(x)%*%y
> betahat
```

	head1	breadth1
	34.282	35.802
head2	0.394	0.245
breadth2	0.529	0.471

and this would allow prediction of first son head sizes from second, though it would be more plausible to predict second from first and then the regression analysis would need to be done the other way around. It is not possible to deduce one from the other in the same way as in univariate regression regressing y on x does not give all the information needed for the regression of x on y . Note however that you can get the estimates of the regression coefficients but not the variance matrix) from univariate [multiple] regressions:-

```
> ll<-lm(head1~head2+breadth2)
> ll
Call:
lm(formula = head1 ~ head2 + breadth2)
Coefficients:
(Intercept)          head2          breadth2
      34.282           0.394           0.529
> bb<-lm(breadth1~head2+breadth2)
> bb
Call:
lm(formula = breadth1 ~ head2 + breadth2)
Coefficients:
(Intercept)          head2          breadth2
      35.802           0.245           0.471
```

4) Read the section on Maximum Likelihood Estimation in Background Results.

This material will be required and used extensively in Chapter 8.

Trust you have done this by now.



Multivariate Data Analysis: Tasks for Week 8

Notes & Solutions

(NB no tasks for week 7)

- 1) Read §8.1 – §8.4 paying particular attention to the results highlighted in boxes as well as §8.3.2 and §8.4.

Trust you have done this by now.

- 2) n observations are available on $x \sim N_p(\mu, \Sigma)$ and C is a known $p \times q$ matrix ($p > q$). By finding the distribution of $y = C'x$, show that a test of $H_0: C'\mu = 0$ vs. $H_A: C'\mu \neq 0$ is given by Hotelling's T^2 with $T^2 = n\bar{x}'C(C'SC)^{-1}C'\bar{x}$ (\bar{x} and S are the sample mean and variance). What parameters does the T^2 distribution have?

If $x \sim N_p(\mu, \Sigma)$ then $y = C'x \sim N_q(\mu_y, \Sigma_y)$ where $\mu_y = C'\mu$ and $\Sigma_y = C'\Sigma C$, further $S_y = C'SC$ and $\bar{y} = C'\bar{x}$ and so the T^2 statistic for testing $\mu_y = 0$ which is $n\bar{y}'S_y^{-1}\bar{y} = n\bar{x}'C(C'SC)^{-1}C'\bar{x}$ and the null distribution is $T^2(q, n-1)$.

- 3) Note: parts (i) & (ii) below should give the same p -value.

- i) A sample of 6 observations on sugar content x_1 and stickiness x_2 of a novel toffee give sample statistics of

$$\bar{x} = \begin{pmatrix} 81.17 \\ 60.33 \end{pmatrix} \text{ and } S = \begin{pmatrix} 27.02 & 7.94 \\ * & 4.26 \end{pmatrix}.$$

Test the hypothesis $H_0: 2\mu_1 = 3\mu_2$

[Suggestion: consider using the 2×1 matrix $C = (2, -3)'$]

We have $n=6$ and $H_0: 2\mu_1 - 3\mu_2 = 0$ i.e. $(2, -3) \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = 0$,

i.e. $C = (2, -3)'$. $C'\bar{x} = -18.65$, $C'SC = 51.14$,

so $T^2 = 40.81 : \frac{n-1}{1} \frac{1}{n-1} T^2 \equiv F_{1,5}$ and so we compare 40.81 with $F_{1,5}$

and then conclude that this is highly significant and so reject the null hypothesis.



- ii) By noting that if $x = (x_1, x_2) \sim N_2(\mu, \Sigma)$ where $\mu = (\mu_1, \mu_2)'$ and Σ has element σ_{ij} then $2x_1 - 3x_2 \sim N((2\mu_1 - 3\mu_2), (4\sigma_{11}^2 + 9\sigma_{22}^2 - 12\sigma_{12}))$ test H_0 in i) above using a Student's t -test.

$$2\bar{x}_1 - 3\bar{x}_2 = -18.65; \quad 4s_{11}^2 + 9s_{22}^2 - 12s_{12} = 51.14, \quad n=6$$

so $t = -\frac{18.65}{\sqrt{51.14/6}} = 6.39$ and compare with t_5 giving same p-value as in

part (i) noting that $6.39^2 = 40.8$ and $t_5^2 \equiv F_{1,5}$.



Multivariate Data Analysis: Tasks for Week 9

Notes & Solutions

- 1) Read the solutions to Exercises 2. These contain a detailed guide to the interpretation of principal components and of crimcoords by examining the loadings of the variables in the PCs and Crimcoords and so provide further practice at this important aspect.

Trust you have done this by now.

- 2) Referring to the data set *dogmandibles*. * **excluding the Prehistoric Thai dogs (group 5 on X_{11})** test the hypotheses that Male and Female dogs have
- i) *equally sized mandibles (i.e. variables X_1 & X_2 together)*

This calls for a Hotelling's T^2 -test with (X_1, X_2) . The easiest way of doing this is to use a MANOVA facility in. Values of T^2 can be obtained as $(n-2) \times$ Lawley-Hotelling statistic.

```
> options(digits=7)
>
> mf.manova<-manova(cbind(length, breadth) ~ gender)
>
> summary.manova(mf.manova, test="H")
              Df Hotelling-Lawley approx F num Df den Df Pr(>F)
gender         1           0.08025  2.56806     2    64 0.08457
Residuals    65
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

The T^2 is then $(32 + 35 - 2) \times 0.08025 = 5.21625$, converting this to an F-value gives 2.568 (as in table above) and p-value 0.085 (also as in table above) and so we conclude that there is only weak evidence of a difference in mean sizes of mandibles between male and female.



a) *equally long mandibles (variable X_1)*

b) *equally broad mandibles (variable X_2)*

These can be done using two separate univariate student t-tests:

```
> options(digits=3)
> t.test(length ~ gender)
```

Welch Two Sample t-test

```
data: length by gender
t = 1.74, df = 64.9, p-value = 0.08606
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -1.12 16.42
sample estimates:
mean in group 1 mean in group 2
      133          126
```

```
> t.test(breadth ~gender)
```

Welch Two Sample t-test

```
data: breadth by gender
t = 2.26, df = 64.8, p-value = 0.02699
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 0.092 1.475
```

and we conclude that there is good evidence of a difference in mean breadths between males and females. Note that the apparent contradiction between the multivariate and univariate tests is in part because of the slight loss of power in the multivariate test caused by estimating more parameters and partly because these are different hypotheses.



ii) equal overall mandible characteristics (i.e. variables X_1 – X_9)

We have

```
> xx.manova=manova(cbind(length, breadth, condyle.breadth,
height,
+ molar.length, molar.breadth, first.to.3rd.length,
+ first.to.4th.length, canine.breadth) ~ gender)
>
> summary(xx.manova, "H")
          Df Hotelling-Lawley approx F num Df den Df Pr(>F)
gender      1           0.137      0.871     9    57  0.56
Residuals 65
```

(so $T^2 = 8.9375$, $p=0.556$) and we conclude there is no evidence in a difference in mean characteristics as measured by these variables. Note that there is no need to calculate the p-value again from the T^2 statistics: it is necessarily the same as that given already.

3) Test the hypotheses that *Iris Versicolor* and *Iris Virginica* have

- i) equally sized sepals
- ii) equally sized petals
- iii) equally sized sepals & petals.

This question calls for several two-sample Hotellings T^2 tests; for (i) we need $p=2$ with elements sepal lengths & widths, for (ii) we need $p=2$ with elements petal lengths & widths, for (iii) we need $p=4$ with elements sepal lengths & widths, petal lengths & widths. The easiest way of doing this is to use a MANOVA facility. Values of T^2 can be obtained as $(n-2) \times$ Lawley-Hotelling statistic. Doing this in **R** has no new features and so it is not given here.

To do this 'by hand' (and it is strongly recommended that you do try it by hand for at least one case) you need to calculate the means and covariances of the four measurements separately for the varieties.

