

## Multivariate Data Analysis: Tasks for Week 10

### Notes & Solutions

1) Suppose we have samples of sizes  $n_1$  and  $n_2$  with means  $\bar{X}_1$  and  $\bar{X}_2$  and variances  $S_1$  and  $S_2$  from populations  $N_p(\mu_1, \sigma_1^2)$  and  $N_p(\mu_2, \sigma_2^2)$ , let  $S = [(n_1 - 1)S_1 + (n_2 - 1)S_2] / (n - 2)$  where  $n = n_1 + n_2$ .

i) Shew that the UIT of  $H_0: \mu_1 = \mu_2$  vs  $H_A: \mu_1 \neq \mu_2$  is given by Hotelling's

$$T^2 = \frac{n_1 n_2}{n} (\bar{X}_1 - \bar{X}_2)' S^{-1} (\bar{X}_1 - \bar{X}_2)$$

ii) Deduce that the greatest difference between the two populations is exhibited in the direction  $S^{-1}(\bar{X}_1 - \bar{X}_2)$ .

[Suggestion: adapt the argument of §5.6.4]

When the data are projected into one dimension we require a two-sample t-test for the equality of two Normal means and we use the

statistic  $t_{\beta}^2 = \frac{n_1 n_2 \beta' (\bar{X}_1 - \bar{X}_2) (\bar{X}_1 - \bar{X}_2)' \beta}{n \beta' S \beta}$  and maximizing this wrt  $\beta$  means

we require that  $\beta$  is the eigenvector of the [rank 1  $p \times p$  matrix]  $n_1 n_2 S^{-1} (\bar{X}_1 - \bar{X}_2) (\bar{X}_1 - \bar{X}_2)' / n$  corresponding to the only non-zero eigenvalue. Easily seen by the usual procedure that this maximum value is  $T^2 = \frac{n_1 n_2}{n} (\bar{X}_1 - \bar{X}_2)' S^{-1} (\bar{X}_1 - \bar{X}_2)$  and that the eigenvector is proportional to  $S^{-1}(\bar{X}_1 - \bar{X}_2)$  which therefore exhibits the maximum deviation between the two populations.



2) Referring to the data set dogmandibles.\* **excluding the Prehistoric Thai dogs (group 5 on  $X_{11}$ )**

- i) What combination of length and breadth of mandible exhibits the greatest difference between Males and Females?

First, the Minitab analysis:

```
MTB > ANOVA 'length' 'breadth' = gender;
SUBC> MANOVA;
SUBC> Eigen;
SUBC> NoUnivariate.
```

**ANOVA: length, breadth versus gender**

```
MANOVA for gender          s = 1      m = 0.0      n = 31.0

Criterion      Test Statistic          F          DF          P
Wilk's         0.92571          2.568      ( 2, 64) 0.085
Lawley-Hotelling 0.08025          2.568      ( 2, 64) 0.085
Pillai's       0.07429          2.568      ( 2, 64) 0.085
Roy's         0.08025
```

**EIGEN Analysis for gender**

```
Eigenvalue  0.08025  0.00000
Proportion  1.00000  0.00000
Cumulative  1.00000  1.00000

Eigenvector      1      2
length    -2.4E-03  -0.01334
breadth    0.111885  0.13027
```

This shows that the combination exhibiting the greatest difference between males and females is proportional to  $0.112 \times \text{breadth} - 0.0024 \times \text{length}$ , or rescaling breadth - 0.0214  $\times$  length.



**Next, in R:**

```

> library(MASS)
> attach(dogmandibles)
> moddogs=dogmandibles[1:67,]
> detach(dogmandibles)
> attach(moddogs)
> lda(gender~length+breadth)
Call:
lda(gender ~ length + breadth)

Prior probabilities of groups:
      1      2
0.5223881 0.4776119

Group means:
  length  breadth
1 133.40 10.274286
2 125.75  9.490625

Coefficients of linear discriminants:
              LD1
length  0.01938347
breadth -0.90204609
Warning message:
In lda.default(x, grouping, ...) : group 3 is empty
>

```

So the linear combination is  $0.01938 \times \text{length} - 0.90205 \times \text{breadth}$ ,  
or rescaling  $\text{breadth} - 0.0214 \times \text{length}$  as before



- ii) What combination of length and breadth of mandible exhibits the greatest difference between the four species?

In R:

```
> lda(species~length+breadth)
Call:
lda(species ~ length + breadth)

Prior probabilities of groups:
      1      2      3      4
0.2388060 0.2985075 0.2537313 0.2089552

Group means:
  length breadth
1 125.3125  9.70625
2 111.0000  8.18000
3 133.2353 10.72353
4 157.3571 11.57857

Coefficients of linear discriminants:
              LD1      LD2
length 0.08941756 -0.1189762
breadth 0.60275578  1.5483264

Proportion of trace:
  LD1  LD2
0.9333 0.0667
Warning message:
In lda.default(x, grouping, ...) : group 5 is empty
```

So the linear combination is  $0.08942 \times \text{length} + 0.6028 \times \text{breadth}$ , or rescaling  $\text{breadth} + 0.0000674 \times \text{length}$  (i.e. primarily breadth).



## Multivariate Data Analysis: Tasks for Week11

### Notes & Solutions

[Note that these questions are more substantial than on previous task sheets. Question 1 is a past examination question. Question 3 is only of benefit to those wanting more practice on PCA interpretation and practical data analysis]

- 1) An archaeologist wishes to distinguish pottery from two different sources on the basis of its chemical composition. Measurements by Neutron Activation Analysis of the concentrations in parts per million of trace elements Cr and V in 19 samples of pottery from Tell el-Amarna gave mean results of 2.3 and 6.7, respectively, with sample variances 0.62 and 1.41 and covariance 0.09. Similar measurements on 23 samples from Memphis gave mean results of 2.9 and 5.9 with sample variances 0.7 and 1.36 and sample covariance 0.08.
- i) Assuming that these measurements are adequately modelled by bivariate Normal distributions with a common variance, calculate the linear discriminant rule for distinguishing Amarna from Memphis pottery on the basis of the concentrations of Cr and V.

First, calculate pooled variance matrix as  $W = [18S_{TA} + 22S_M]/40$

$$= \begin{pmatrix} 0.664 & 0.085 \\ * & 1.383 \end{pmatrix} \text{ and so } W^{-1} = \begin{pmatrix} 1.518 & -0.093 \\ * & 0.729 \end{pmatrix}$$

If  $x = (x_{Cr}, x_V)'$  then the linear discriminant rule is to classify  $x$  as from Amarna if  $((2.3, 6.7) - (2.9, 5.9))W^{-1}(x - \frac{1}{2}((2.3, 6.7) + (2.9, 5.9)))' > 0$ ,  
 i.e. if  $(-0.6, 0.8)W^{-1}(x - (2.6, 6.3)') > 0$ ,  
 i.e. if  $(-0.985, 0.639)(x - (2.6, 6.3)') > 0$   
 i.e. if  $-0.985x_{Cr} + 0.639x_V - 1.465 > 0$



- ii) Prove that the estimated probabilities of misclassifying Memphis pottery as Amarna and vice versa are the same using this rule.

Classification rule is to allocate to Amarna if

$$h(x) = -0.985x_{Cr} + 0.639x_V - 1.465 > 0$$

Now if  $x$  is from Memphis then

$$\begin{aligned} E[h(x)] &= -0.985 \times 2.9 + 0.639 \times 5.9 - 1.465 = -0.551 \text{ and } \text{var}(h(x)) = \\ &= (-0.985)^2(0.664) + (0.639)^2(1.383) + 2(-0.985)(0.639)(0.085) = \\ &1.102, \end{aligned}$$

so  $P[\text{classify as Amarna} \mid \text{from Memphis}]$

$$\begin{aligned} &= P[h(x) > 0 \mid h(x) \sim N(-0.551, 1.102)] = 1 - \Phi(0.551/\sqrt{1.102}) \\ &= 1 - \Phi(0.525) = 1 - 0.700 = 0.300. \end{aligned}$$

If  $x$  is from Amarna then

$$\begin{aligned} E[h(x)] &= -0.985 \times 2.3 + 0.639 \times 6.7 - 1.465 = 0.551 \text{ and } \text{var}(h(x)) = \\ &= (-0.985)^2(0.664) + (0.639)^2(1.261) + 2(-0.985)(0.639)(0.085) = \\ &1.102. \end{aligned}$$

So  $P[\text{classify as Memphis} \mid \text{from Amarna}]$

$$\begin{aligned} &= P[h(x) < 0 \mid h(x) \sim N(0.551, 1.102)] \\ &= \Phi(-0.551/\sqrt{1.102}) = \Phi(-0.525) = 0.30 \end{aligned}$$

Thus the two misclassification probabilities are equal.



- iii) *By how much is this misclassification probability an improvement over those using each of the elements separately?*

If only Cr is used then rule is to classify as Amarna if

$$h_{Cr}(x_{Cr}) = (2.3 - 2.9)(0.664^{-1}(x_{Cr} - 2.6)) > 0, \text{ i.e. if } x_{Cr} < 2.6.$$

If  $x$  is from Memphis, then  $x_{Cr} \sim N(2.9, 0.664)$  and

$$\begin{aligned} \text{so } P[\text{classify as Amarna} \mid \text{from Memphis}] &= \Phi(-0.3/0.664^{1/2}) \\ &= \Phi(-0.368) = 0.356 \end{aligned}$$

Similarly if only V is used then rule is to classify as Amarna if  $x_V > 6.3$  and so  $P[\text{classify as Amarna} \mid \text{from Memphis}]$

$$= 1 - \Phi(0.4/1.383^{1/2}) = 1 - \Phi(0.340) = 1 - 0.633 = 0.367.$$

Thus the improvement in misclassification probability over using just Cr is 5.6% and over using just V it is 6.7%

- iv) *What advice would you give to the archaeologist in the light of these results?*

The archaeologist needs to be warned that the error rate will be at least 30% if (s)he uses either or both of the elements. This is substantial and may give cause for not proceeding with the study. Measuring more trace elements will certainly not worsen the situation.



2) Referring to the data set *dogmandibles.\** (including the Prehistoric Thai dogs (group 5 on  $X_{11}$ ))

- i) Using `STAT>MULTIVARIATE>DISCRIMINANT` in Minitab or `lda()` in S-PLUS look at the discrimination between the 5 species (using the nine measurements) and estimate the classification rate. [In S-PLUS it is easy to find the cross-validation (or jackknife) estimate of classification rate].

NB The computer analysis in the solutions given here and in Q3 have been produced using Minitab. The S-PLUS analysis is considerably easier and has not been given but if there are any difficulties then this can be provided

```
MTB > Discriminant 'species' 'length'-'canine breadth';
SUBC> Predict C51-C59.
```

### Discriminant Analysis: species versus length, breadth, ...

```
Linear Method for Response: species
Predictors: length breadth condyle height molar le molar br first to
            first to canine b
```

| Group | 1  | 2  | 3  | 4  | 5  |
|-------|----|----|----|----|----|
| Count | 16 | 20 | 17 | 14 | 10 |

#### Summary of Classification

| Put into   | ...True Group... |       |       |       |       |
|------------|------------------|-------|-------|-------|-------|
| Group      | 1                | 2     | 3     | 4     | 5     |
| 1          | 15               | 0     | 0     | 0     | 2     |
| 2          | 0                | 20    | 0     | 0     | 0     |
| 3          | 0                | 0     | 17    | 0     | 0     |
| 4          | 0                | 0     | 0     | 14    | 0     |
| 5          | 1                | 0     | 0     | 0     | 8     |
| Total N    | 16               | 20    | 17    | 14    | 10    |
| N Correct  | 15               | 20    | 17    | 14    | 8     |
| Proportion | 0.938            | 1.000 | 1.000 | 1.000 | 0.800 |

N = 77    N Correct = 74    Proportion Correct = 0.961

So estimated classification rate without using cross-validation is 96%:

With cross-validation gives

```
MTB > Discriminant 'species' 'length'-'canine breadth';
SUBC> XVal;
SUBC> Predict C51-C59.
```

### Discriminant Analysis: species versus length, breadth, ...

```
Linear Method for Response: species
Predictors: length breadth condyle height molar le molar br first to
            first to canine b
```

| Group | 1  | 2  | 3  | 4  | 5  |
|-------|----|----|----|----|----|
| Count | 16 | 20 | 17 | 14 | 10 |



Summary of Classification

| Put into   | ...True Group... |       |       |       |       |
|------------|------------------|-------|-------|-------|-------|
| Group      | 1                | 2     | 3     | 4     | 5     |
| 1          | 15               | 0     | 0     | 0     | 2     |
| 2          | 0                | 20    | 0     | 0     | 0     |
| 3          | 0                | 0     | 17    | 0     | 0     |
| 4          | 0                | 0     | 0     | 14    | 0     |
| 5          | 1                | 0     | 0     | 0     | 8     |
| Total N    | 16               | 20    | 17    | 14    | 10    |
| N Correct  | 15               | 20    | 17    | 14    | 8     |
| Proportion | 0.938            | 1.000 | 1.000 | 1.000 | 0.800 |

N = 77      N Correct = 74      Proportion Correct = 0.961

Summary of Classification with Cross-validation

| Put into   | ...True Group... |       |       |       |       |
|------------|------------------|-------|-------|-------|-------|
| Group      | 1                | 2     | 3     | 4     | 5     |
| 1          | 14               | 1     | 0     | 0     | 3     |
| 2          | 0                | 19    | 0     | 0     | 0     |
| 3          | 0                | 0     | 17    | 0     | 0     |
| 4          | 0                | 0     | 0     | 13    | 0     |
| 5          | 2                | 0     | 0     | 1     | 7     |
| Total N    | 16               | 20    | 17    | 14    | 10    |
| N Correct  | 14               | 19    | 17    | 13    | 7     |
| Proportion | 0.875            | 0.950 | 1.000 | 0.929 | 0.700 |

N = 77      N Correct = 70      Proportion Correct = 0.909

so an estimate of 91%.

- ii) *Perform the discriminant analysis just on the first four [modern] species and then use this to classify the prehistoric Thai dogs.*

```
MTB > COPY C1-C11 C101-C111;
SUBC> USE C11 = 5.
MTB > copy c1-c11 c1-c11 ;
SUBC> omit c11 = 5.
MTB > Discriminant 'species' 'length'-'canine breadth';
SUBC> Predict c101-c109.
```

**Discriminant Analysis: species versus length, breadth, ...**

Linear Method for Response: species  
 Predictors: length breadth condyle height molar le molar br first to  
 first to canine b

| Group | 1  | 2  | 3  | 4  |
|-------|----|----|----|----|
| Count | 16 | 20 | 17 | 14 |

Summary of Classification

| Put into   | ...True Group... |       |       |       |
|------------|------------------|-------|-------|-------|
| Group      | 1                | 2     | 3     | 4     |
| 1          | 16               | 0     | 0     | 0     |
| 2          | 0                | 20    | 0     | 0     |
| 3          | 0                | 0     | 17    | 0     |
| 4          | 0                | 0     | 0     | 14    |
| Total N    | 16               | 20    | 17    | 14    |
| N Correct  | 16               | 20    | 17    | 14    |
| Proportion | 1.000            | 1.000 | 1.000 | 1.000 |

N = 67      N Correct = 67      Proportion Correct = 1.000



## Prediction for Test Observations

| Observation | Pred Group | From Group | Sqrd Distnc | Probability |
|-------------|------------|------------|-------------|-------------|
| 1           | 1          | 1          | 14.722      | 0.987       |
|             |            | 2          | 23.353      | 0.013       |
|             |            | 3          | 71.906      | 0.000       |
|             |            | 4          | 63.387      | 0.000       |
| 2           | 1          | 1          | 14.226      | 1.000       |
|             |            | 2          | 36.185      | 0.000       |
|             |            | 3          | 88.289      | 0.000       |
|             |            | 4          | 49.060      | 0.000       |
| 3           | 1          | 1          | 17.875      | 1.000       |
|             |            | 2          | 52.703      | 0.000       |
|             |            | 3          | 93.366      | 0.000       |
|             |            | 4          | 37.821      | 0.000       |
| 4           | 1          | 1          | 8.635       | 0.998       |
|             |            | 2          | 21.142      | 0.002       |
|             |            | 3          | 66.208      | 0.000       |
|             |            | 4          | 69.793      | 0.000       |
| 5           | 1          | 1          | 39.810      | 1.000       |
|             |            | 2          | 83.256      | 0.000       |
|             |            | 3          | 113.618     | 0.000       |
|             |            | 4          | 75.934      | 0.000       |
| 6           | 1          | 1          | 27.584      | 1.000       |
|             |            | 2          | 69.580      | 0.000       |
|             |            | 3          | 65.908      | 0.000       |
|             |            | 4          | 66.938      | 0.000       |
| 7           | 1          | 1          | 39.170      | 1.000       |
|             |            | 2          | 59.289      | 0.000       |
|             |            | 3          | 126.981     | 0.000       |
|             |            | 4          | 74.862      | 0.000       |
| 8           | 1          | 1          | 8.226       | 1.000       |
|             |            | 2          | 33.646      | 0.000       |
|             |            | 3          | 78.406      | 0.000       |
|             |            | 4          | 55.935      | 0.000       |
| 9           | 1          | 1          | 16.727      | 1.000       |
|             |            | 2          | 42.650      | 0.000       |
|             |            | 3          | 108.792     | 0.000       |
|             |            | 4          | 73.268      | 0.000       |
| 10          | 1          | 1          | 12.329      | 1.000       |
|             |            | 2          | 40.900      | 0.000       |
|             |            | 3          | 75.376      | 0.000       |
|             |            | 4          | 57.443      | 0.000       |

Note the apparent 100% correct classification on just the modern species and that with 'near certainty' all the prehistoric are classified as 'modern', even though

iii) Compare the results of these analyses with the results of the more informal exploratory analyses with Crimcoords in Exercises 2.

The plots on crimcoords revealed a consistent slight difference from modern dogs.



- 3) The datafile CLAYPOTS has 272 observations on the trace element content of clay samples from pots found at various archaeological sites around the Aegean. Column 1 gives the group number (i.e. archaeological site for most of the pots) and columns 2—9 give the amounts of 9 trace elements (which have been labelled A to I) found in samples of clay from the pots. It is suggested that before investigating the specific questions below it is advisable to do some exploratory analysis with PCA etc. Groups 1, 3 and 4 are from known sources; groups 2 and 5 are from unknown sources but are believed to come from one or other of 1,3 or 4.
- i) Construct a display on crimcoords of groups 1,3 and 4 and add in the points from groups 2 and 5.
  - ii) Which are the best classifications of these pots?

Note that in Minitab you can copy all the variables to new columns, omitting those rows with  $C1 > 5$ , this gives you the data for just the first five groups. Then you may(?) need to separate out the data for groups 2 and 5 into entirely separate columns again. To plot the data onto crimcoords you need to use the matrix multiplication facility, having first found the matrix of eigenvectors, either by decomposing the  $W^{-1}B$  matrix into symmetric factors or by doing a multivariate analysis of variance (MANOVA) between the groups 1,3 and 4 (i.e. not including 2 and 5) with the eigenanalysis option active and then either cut&paste or by copying down the eigenvectors by hand and re-entering them into columns. In S-PLUS you can use the function `lda()` and the generic function `predict()` — use the Help system to find out the details.



```
Welcome to Minitab, press F1 for help.
MTB > RETR "C:\TEACHING\MVA\CLAYPOTS.MTW"
Retrieving worksheet from file: C:\TEACHING\MVA\CLAYPOTS.MTW
# Worksheet was saved on 22/11/01 11:10:19
```

## Results for: CLAYPOTS.MTW

```
# First remove all data from groups 6 and above, done using
the Copy columns in Manip with the mit rows option
MTB > Copy 'Group'-'I' 'Group'-'I';
SUBC> Omit 'Group' = 6:99.
# Next, copy out the data for groups 2 and 5, with the Use
rows option, not forgetting to cancel the omit rows option
from last time
MTB > Copy 'Group'-'I' c101-c110;
SUBC> Use 'Group' = 2 5.
# Next, copy remove the data from groups 2 and 5 from the main
body by copying the columns into themselves, with the omit
rows option, not forgetting to cancel the use rows option from
last time
MTB > Copy 'Group'-'I' 'Group'-'I';
SUBC> Omit 'Group' = 2 5.
# Now do the MANOVA to get the W**(-1)B matrix. This is in
Balanced Manova within ANOVA in version 13 but in version 12
it is in Multivariate, Balanced manova and some details of
output may be different.
MTB > ANOVA 'A'-'I' = Group;
SUBC> MANOVA;
SUBC> Eigen;
SUBC> NoUnivariate.
```

## ANOVA: A, B, C, D, E, F, G, H, I versus Group

```
MANOVA for Group          s = 2      m = 3.0      n =
24.5
```

| Criterion        | Test Statistic | F      | DF         | P     |
|------------------|----------------|--------|------------|-------|
| Wilk's           | 0.06928        | 15.862 | ( 18, 102) | 0.000 |
| Lawley-Hotelling | 6.35730        | 17.659 | ( 18, 100) | 0.000 |
| Pillai's         | 1.42097        | 14.179 | ( 18, 104) | 0.000 |
| Roy's            | 4.91873        |        |            |       |



## EIGEN Analysis for Group

|            |        |        |         |         |         |         |
|------------|--------|--------|---------|---------|---------|---------|
| Eigenvalue | 4.9187 | 1.4386 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Proportion | 0.7737 | 0.2263 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Cumulative | 0.7737 | 1.0000 | 1.00000 | 1.00000 | 1.00000 | 1.00000 |

|             |        |        |        |        |        |        |
|-------------|--------|--------|--------|--------|--------|--------|
| Eigenvector | 1      | 2      | 3      | 4      | 5      | 6      |
| A           | -0.010 | -0.001 | -0.007 | -0.003 | 0.026  | -0.031 |
| B           | -0.027 | 0.197  | -0.099 | -0.103 | -0.086 | 0.036  |
| C           | -0.012 | 0.030  | 0.154  | 0.051  | -0.006 | 0.085  |
| D           | 0.433  | -0.332 | 0.260  | -0.284 | 0.043  | -0.018 |
| E           | -3.026 | -3.674 | -4.324 | -1.873 | -1.920 | 0.182  |
| F           | 12.912 | -8.621 | 0.784  | -3.279 | -2.330 | -2.055 |
| G           | -0.006 | -0.017 | 0.010  | 0.005  | -0.002 | -0.001 |
| H           | -0.060 | 0.054  | -0.037 | 0.151  | -0.218 | -0.168 |
| I           | -7.888 | 0.678  | -5.970 | 12.367 | 9.882  | -1.723 |

|            |         |         |         |
|------------|---------|---------|---------|
| Eigenvalue | 0.00000 | 0.00000 | 0.00000 |
| Proportion | 0.00000 | 0.00000 | 0.00000 |
| Cumulative | 1.00000 | 1.00000 | 1.00000 |

|             |        |        |        |
|-------------|--------|--------|--------|
| Eigenvector | 7      | 8      | 9      |
| A           | -0.033 | -0.044 | -0.045 |
| B           | -0.005 | -0.001 | 0.107  |
| C           | -0.016 | 0.012  | 0.095  |
| D           | 1.537  | -0.207 | -0.025 |
| E           | -3.408 | 2.042  | -7.957 |
| F           | -1.129 | -0.909 | -4.936 |
| G           | 0.021  | 0.023  | -0.002 |
| H           | 0.017  | 0.037  | 0.024  |
| I           | -0.001 | -0.041 | -3.756 |

# Now highlight the first two eigenvectors by positioning the cursor just to the left of the -0.010 of eigenvector 1, holding down the ALT key and then with the left button depressed moving the cursor just to the right of 0.678, (this selects just the columns rather than all the rows). Now click the copy icon to copy into the clipboard, then move to the data window and click the cell in row 1 of C12, then click the paste icon to paste the two eigenvectors into C12 and C13.

#

# Now move back the data for groups 2 and 5 to be in the same columns as the 'training data' for groups 1,3 and 4. Use Manip>Stack>Stack Blocks of columns to do this.

```
MTB > Stack ('Group'-'I') (C101-C110) ('Group'-'I').
```

# Next, copy training data (groups 1,3, 4) and 'new cases' (groups 2 and 5) into a matrix m1.

```
MTB > Copy 'A'-'I' m1.
```

# copy the two eigenvectors (which have just been pasted into the data sheet) into matrix m2

```
MTB > Copy C12 C13 m2.
```

# rotate all the data (training and new) onto the crimcoords.

```
MTB > Multiply m1 m2 m3.
```



# and copy the resulting matrix back into columns for plotting, naming them sensibly.

```
MTB > Copy m3 c15 c16.
```

```
MTB > name c15 'Crimcoord 1' c16 'Crimcoord 2'
```

# Now produce the plots with **Graph>Plot** and under **Data display** change the default in 'For each' from 'Graph' to 'Group' and give **Group** as the name of the variable, then go into **Edit Attributes** to choose pretty symbols for the plot (alternatively use the lines of code given below).

```
MTB > Plot 'Crimcoord 1'*'Crimcoord 2';
```

```
SUBC> Symbol 'Group';
```

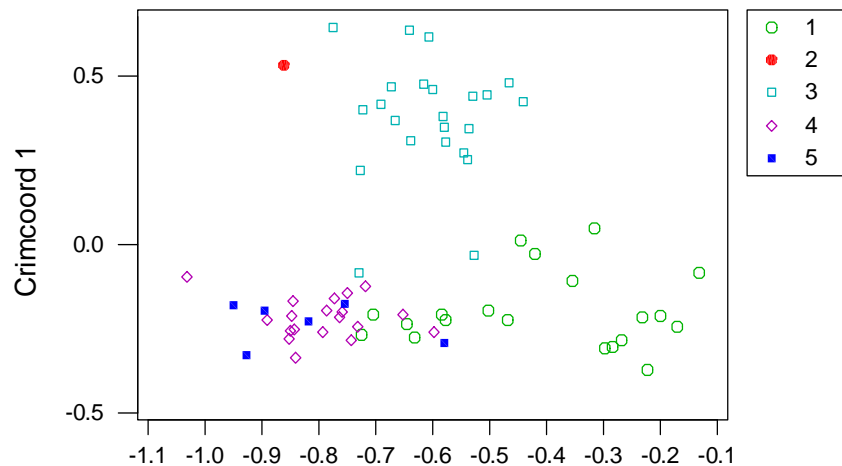
```
SUBC> Type 1 6 11 15 12;
```

```
SUBC> Color 9 2 11 12 4;
```

```
SUBC> ScFrame;
```

```
SUBC> ScAnnotation.
```

## Plot Crimcoord 1 \* Crimcoord 2



It is clear from the plot that most of group 5 are like group 4 with one in the overlap region with group 1, the group 2 pot is most like group 3 out of the three possibilities on offer but is not a 'typical' group 3 pot.



- iii) Compare your opinions with the results from the ready-made analysis in *STATS>MULTIVARIATE>DISCRIMINANT* using the options to predict the membership of groups 2 and 5.

**First without cross-validation**

```
MTB > Discriminant 'Group' 'A'-'I';
SUBC> Predict C102-C110.
```

**Discriminant Analysis: Group versus A, B, C, D, E, F, G, H, I**

```
Linear Method for Response: Group
Predictors: A B C D E F G H I
```

```
Group      1      3      4
Count      20     23     19
```

Summary of Classification

```
Put into      ....True Group....
Group          1      3      4
1              16     1      1
3              0     21     0
4              4      1     18
Total N        20     23     19
N Correct      16     21     18
Proportion    0.800  0.913  0.947
```

N = 62      N Correct = 55      Proportion Correct = 0.887

Squared Distance Between Groups

```
          1      3      4
1      0.0000  21.6392  8.7280
3      21.6392  0.0000  22.8474
4      8.7280  22.8474  0.0000
```

Linear Discriminant Function for Group

```
          1      3      4
Constant -64.24 -72.82 -78.14
A         1.90  1.55  1.86
B        -8.32 -13.37 -12.23
C         5.51  5.88  8.46
D        56.90 80.71 65.72
E        77.11 10.79 33.65
F       196.22 750.93 248.60
```



|   |       |         |        |
|---|-------|---------|--------|
| G | 0.59  | 0.83    | 0.99   |
| H | 6.44  | 2.74    | 4.81   |
| I | 37.33 | -290.41 | -66.66 |

Summary of Misclassified Observations

| Observation | True Group | Pred Group | Group | Squared Distance | Probability |
|-------------|------------|------------|-------|------------------|-------------|
| 5 **        | 1          | 4          | 1     | 18.52            | 0.090       |
|             |            |            | 3     | 34.33            | 0.000       |
|             |            |            | 4     | 13.89            | 0.910       |
| 9 **        | 1          | 4          | 1     | 7.631            | 0.265       |
|             |            |            | 3     | 26.576           | 0.000       |
|             |            |            | 4     | 5.591            | 0.735       |
| 12 **       | 1          | 4          | 1     | 10.001           | 0.313       |
|             |            |            | 3     | 31.864           | 0.000       |
|             |            |            | 4     | 8.431            | 0.687       |
| 14 **       | 1          | 4          | 1     | 11.013           | 0.054       |
|             |            |            | 3     | 30.185           | 0.000       |
|             |            |            | 4     | 5.277            | 0.946       |
| 24 **       | 3          | 4          | 1     | 8.920            | 0.061       |
|             |            |            | 3     | 15.680           | 0.002       |
|             |            |            | 4     | 3.444            | 0.937       |
| 34 **       | 3          | 1          | 1     | 12.22            | 0.854       |
|             |            |            | 3     | 19.98            | 0.018       |
|             |            |            | 4     | 16.01            | 0.129       |
| 56 **       | 4          | 1          | 1     | 6.681            | 0.512       |
|             |            |            | 3     | 28.222           | 0.000       |
|             |            |            | 4     | 6.776            | 0.488       |

Prediction for Test Observations

| Observation | Pred Group | From Group | Sqrd Distnc | Probability |
|-------------|------------|------------|-------------|-------------|
| 1           | 3          | 1          | 68.554      | 0.000       |
|             |            | 3          | 29.937      | 1.000       |
|             |            | 4          | 58.479      | 0.000       |
| 2           | 4          | 1          | 25.307      | 0.032       |
|             |            | 3          | 37.330      | 0.000       |
|             |            | 4          | 18.517      | 0.967       |
| 3           | 4          | 1          | 78.880      | 0.000       |
|             |            | 3          | 86.709      | 0.000       |
|             |            | 4          | 63.197      | 1.000       |
| 4           | 4          | 1          | 92.554      | 0.001       |
|             |            | 3          | 102.880     | 0.000       |
|             |            | 4          | 79.229      | 0.999       |
| 5           | 4          | 1          | 80.461      | 0.001       |
|             |            | 3          | 98.965      | 0.000       |
|             |            | 4          | 65.380      | 0.999       |
| 6           | 4          | 1          | 138.275     | 0.007       |
|             |            | 3          | 152.710     | 0.000       |
|             |            | 4          | 128.415     | 0.993       |
| 7           | 1          | 1          | 202.608     | 0.600       |



|   |         |       |
|---|---------|-------|
| 3 | 226.511 | 0.000 |
| 4 | 203.419 | 0.400 |

Note that these classifications agree with those obtained informally with the crimcoord plot but that the classification of the group 2 plot as type 3 has been made with 'near certainty' making no allowance for the fact that it does not look like a typical group 3 pot.

```
MTB > Discriminant 'Group' 'A'-'I';
SUBC> XVal;
SUBC> Predict C102-C110.
```

**Discriminant Analysis: Group versus A, B, C, D, E, F, G, H, I**

Linear Method for Response: Group  
 Predictors: A B C D E F G H I

|       |    |    |    |
|-------|----|----|----|
| Group | 1  | 3  | 4  |
| Count | 20 | 23 | 19 |

Summary of Classification

|            |                    |       |       |
|------------|--------------------|-------|-------|
| Put into   | ....True Group.... |       |       |
| Group      | 1                  | 3     | 4     |
| 1          | 16                 | 1     | 1     |
| 3          | 0                  | 21    | 0     |
| 4          | 4                  | 1     | 18    |
| Total N    | 20                 | 23    | 19    |
| N Correct  | 16                 | 21    | 18    |
| Proportion | 0.800              | 0.913 | 0.947 |

N = 62      N Correct = 55      Proportion Correct = 0.887

Summary of Classification with Cross-validation

|            |                    |       |       |
|------------|--------------------|-------|-------|
| Put into   | ....True Group.... |       |       |
| Group      | 1                  | 3     | 4     |
| 1          | 12                 | 1     | 1     |
| 3          | 1                  | 21    | 0     |
| 4          | 7                  | 1     | 18    |
| Total N    | 20                 | 23    | 19    |
| N Correct  | 12                 | 21    | 18    |
| Proportion | 0.600              | 0.913 | 0.947 |

N = 62      N Correct = 51      Proportion Correct = 0.823

Squared Distance Between Groups

|   |   |   |
|---|---|---|
| 1 | 3 | 4 |
|---|---|---|



1            0.0000   21.6392   8.7280  
 3            21.6392   0.0000   22.8474  
 4            8.7280   22.8474   0.0000

Summary of Misclassified Observations

| Observation | True  | Pred  | X-val | Group | Squared | Distance |      |
|-------------|-------|-------|-------|-------|---------|----------|------|
| Probability | Group | Group | Group |       | Pred    | X-val    | Pred |
| X-val       |       |       |       |       |         |          |      |
| 5 **        | 1     | 4     | 4     | 1     | 18.52   | 30.13    | 0.09 |
| 0.00        |       |       |       | 3     | 34.33   | 40.13    | 0.00 |
| 0.00        |       |       |       | 4     | 13.89   | 17.32    | 0.91 |
| 1.00        |       |       |       |       |         |          |      |
| 8 **        | 1     | 1     | 4     | 1     | 23.84   | 45.18    | 0.95 |
| 0.44        |       |       |       | 3     | 46.00   | 62.95    | 0.00 |
| 0.00        |       |       |       | 4     | 29.83   | 44.73    | 0.05 |
| 0.56        |       |       |       |       |         |          |      |
| 9 **        | 1     | 4     | 4     | 1     | 7.631   | 9.622    | 0.27 |
| 0.12        |       |       |       | 3     | 26.576  | 26.928   | 0.00 |
| 0.00        |       |       |       | 4     | 5.591   | 5.599    | 0.73 |
| 0.88        |       |       |       |       |         |          |      |
| 10 **       | 1     | 1     | 4     | 1     | 5.174   | 6.208    | 0.63 |
| 0.49        |       |       |       | 3     | 24.754  | 24.666   | 0.00 |
| 0.00        |       |       |       | 4     | 6.196   | 6.125    | 0.37 |
| 0.51        |       |       |       |       |         |          |      |
| 12 **       | 1     | 4     | 4     | 1     | 10.001  | 13.259   | 0.31 |
| 0.10        |       |       |       | 3     | 31.864  | 33.507   | 0.00 |
| 0.00        |       |       |       | 4     | 8.431   | 8.790    | 0.69 |
| 0.90        |       |       |       |       |         |          |      |
| 14 **       | 1     | 4     | 4     | 1     | 11.013  | 14.930   | 0.05 |
| 0.01        |       |       |       | 3     | 30.185  | 31.761   | 0.00 |
| 0.00        |       |       |       | 4     | 5.277   | 5.500    | 0.95 |
| 0.99        |       |       |       |       |         |          |      |
| 16 **       | 1     | 1     | 4     | 1     | 10.90   | 14.74    | 0.59 |
| 0.24        |       |       |       | 3     | 29.16   | 30.51    | 0.00 |
| 0.00        |       |       |       | 4     | 11.61   | 12.45    | 0.41 |
| 0.76        |       |       |       |       |         |          |      |
| 19 **       | 1     | 1     | 3     | 1     | 19.87   | 33.52    | 0.94 |
| 0.14        |       |       |       | 3     | 26.35   | 30.01    | 0.04 |
| 0.83        |       |       |       | 4     | 27.46   | 37.11    | 0.02 |
| 0.02        |       |       |       |       |         |          |      |
| 24 **       | 3     | 4     | 4     | 1     | 8.920   | 8.822    | 0.06 |
| 0.06        |       |       |       |       |         |          |      |



|      |       |   |   |   |        |        |      |
|------|-------|---|---|---|--------|--------|------|
| 0.00 |       |   |   | 3 | 15.680 | 23.329 | 0.00 |
| 0.94 |       |   |   | 4 | 3.444  | 3.469  | 0.94 |
| 0.89 | 34 ** | 3 | 1 | 1 | 12.22  | 12.77  | 0.85 |
| 0.00 |       |   |   | 3 | 19.98  | 33.24  | 0.02 |
| 0.11 |       |   |   | 4 | 16.01  | 16.91  | 0.13 |
| 0.71 | 56 ** | 4 | 1 | 1 | 6.681  | 6.679  | 0.51 |
| 0.00 |       |   |   | 3 | 28.222 | 28.482 | 0.00 |
| 0.29 |       |   |   | 4 | 6.776  | 8.446  | 0.49 |

**Prediction for Test Observations**

| Observation | Pred Group | From Group | Sqrd Distnc | Probability |
|-------------|------------|------------|-------------|-------------|
| 1           | 3          | 1          | 68.554      | 0.000       |
|             |            | 3          | 29.937      | 1.000       |
|             |            | 4          | 58.479      | 0.000       |
| 2           | 4          | 1          | 25.307      | 0.032       |
|             |            | 3          | 37.330      | 0.000       |
|             |            | 4          | 18.517      | 0.967       |
| 3           | 4          | 1          | 78.880      | 0.000       |
|             |            | 3          | 86.709      | 0.000       |
|             |            | 4          | 63.197      | 1.000       |
| 4           | 4          | 1          | 92.554      | 0.001       |
|             |            | 3          | 102.880     | 0.000       |
|             |            | 4          | 79.229      | 0.999       |
| 5           | 4          | 1          | 80.461      | 0.001       |
|             |            | 3          | 98.965      | 0.000       |
|             |            | 4          | 65.380      | 0.999       |
| 6           | 4          | 1          | 138.275     | 0.007       |
|             |            | 3          | 152.710     | 0.000       |
|             |            | 4          | 128.415     | 0.993       |
| 7           | 1          | 1          | 202.608     | 0.600       |
|             |            | 3          | 226.511     | 0.000       |
|             |            | 4          | 203.419     | 0.400       |

Note that the cross-validated estimate of the classification rate is slightly lower and [of course] the classification of the new pots is unaltered by the cross-validation option.



1)  $x_1, \dots, x_n$  are independent measurements of  $N_p(\mu, \sigma^2 I_p)$

- i) Shew that the maximum likelihood estimate of  $\mu$ , subject to  $\mu' \mu = r_0^2$  (a known constant) is the same whether  $\sigma$  is known or unknown.

This example is very like example 5.5.3 in the lecture notes:

We have  $\ell(\mu; X) = -\frac{1}{2}(n-1)\text{trace}(S\sigma^{-2}) - \frac{1}{2}n(\bar{x} - \mu)'(\bar{x} - \mu)\sigma^{-2} - \frac{1}{2}n\text{plog}(2\pi) - \frac{1}{2}n\text{plog}(\sigma^2)$

Let  $\Omega = \ell(\mu) - \lambda(\mu' \mu - r_0^2)$  then  $\frac{\partial \Omega}{\partial \mu} = n(\bar{x} - \mu)\sigma^{-2} - 2\lambda\mu$ .

So we require  $\hat{\mu} = \frac{n\bar{x}}{n+2\lambda\sigma^2}$  then  $\mu' \mu = r_0^2$  implies  $(n+2\lambda\sigma^2)^2 r_0^2 = n^2 \bar{x}' \bar{x}$

and so  $\hat{\mu} = \frac{\bar{x} r_0}{\sqrt{\bar{x}' \bar{x}}}$  which does not depend on  $\sigma^2$ .

- ii) Find the maximum likelihood estimate of  $\sigma$  when neither  $\mu$  nor  $\sigma$  are known.

$$\frac{\partial \Omega}{\partial \sigma} = (n-1)\text{tr}(S)\sigma^{-3} + n(\bar{x} - \mu)'(\bar{x} - \mu)\sigma^{-3} - n\text{p}\sigma^{-1}$$

$$\text{so } \hat{\sigma} = \sqrt{\frac{1}{n\text{p}} \left[ (n-1)\text{tr}(S) + n(\bar{x} - \hat{\mu})'(\bar{x} - \hat{\mu}) \right]}$$

$$= \sqrt{\frac{1}{n\text{p}} \left[ (n-1)\text{tr}(S) + n(\sqrt{\bar{x}' \bar{x}} - r_0)^2 \right]}$$



- iii) Hence, in the case when  $\sigma = \sigma_0$  (a known constant) construct the likelihood ratio test of  $H_0 : \mu' \mu = r_0^2$  vs  $H_A : \mu' \mu \neq r_0^2$  based on  $n$  independent observations of  $N_p(\mu, \sigma_0^2 I_p)$ .

Under  $H_0$

$$\ell_{\max} = K - \frac{1}{2}n(\sqrt{\bar{x}'\bar{x}} - r_0)^2 \sigma_0^{-2}$$

Under  $H_A$  we have

$$\hat{\mu} = \bar{x} \quad \text{so} \quad \ell_{\max} = K$$

so LRT statistic is  $n(\sqrt{\bar{x}'\bar{x}} - r_0)^2 \sigma_0^{-2}$  and under  $H_0$  this  $\sim \chi_1^2$

[1 d.f. since  $p$  parameters in  $\mu$  estimated under  $H_A$  and  $p$  with 1 constraint under  $H_0$ ]

- iv) In an experiment to test the range of a new ground-to-air missile thirty-nine test firings at a tethered balloon were performed and the three dimensional coordinates of the point of ignition of the missile's warhead measured. These gave a mean result of  $(0.76, 0.69, 0.66)'$  relative to the site expressed in terms of the target distance. Presuming that individual measurements are independently normally distributed with unit variance, are the data consistent with the theory that the range of the missile was set correctly?

We have  $\sigma_0=1=r_0$  and so

$$n(\sqrt{\bar{x}'\bar{x}} - r_0)^2 \sigma_0^{-2} = 39(\sqrt{(0.76, 0.69, 0.66)'(0.76, 0.69, 0.66)} - 1)^2$$

= 1.894 ( $\ll 3.84 = \chi_{1,0.95}^2$ ) and so yes, the data are consistent with the theory that the range was set correctly



