

Multivariate Data Analysis

Notes & Solutions to Exercises 2

1) The data given in file *dogmandibles.** (in various formats) are extracted, via Manly (1994), from Higham *etc* (1980), *J.Arch.Sci*, 149–165. The file contains 9 measurements of various dimensions of the mandibles of 5 canine species as well as records of the sex and the species, eleven variables in total. These are X_1 : length of mandible

X_2 : breadth of mandible

X_3 : breadth of articular condyle

X_4 : height of mandible below first molar

X_5 : length of 1st molar

X_6 : breadth of 1st molar

X_7 : length between 1st to 3rd molar inclusive (1st to 2nd for Cuons)

X_8 : length between 1st to 4th premolar inclusive

X_9 : breadth of lower canine

X_{10} : gender (1 = male, 2 = female, 3 = unknown)

X_{11} : species (1 = modern dog from Thailand, 2 = Golden Jackal,

3 = Cuon, 4 = Indian Wolf, 5 = Prehistoric Thai dog)

All measurements are in mm; molars, premolars and canines are types of teeth; an articular condyle is the round knobby bit in a joint; a Cuon, or Red Dog, is a wild dog indigenous to south east Asia and notable for lacking one pair of molars.

i) Ignoring the group structure, what interpretations can be given to the first two principal components?

Step 1 is to perform a PCA on the linear measurements for the complete data set (i.e. all 5 groups). Initial inspection shews (but not given here) that the standard deviations of the measurements vary widely — this is inevitable given that X_1 has values in the 100s and X_9 below 10 — so basing the PCA on the correlation matrix is preferable. (PCA on the covariance matrix gives the first eigenvalue as 0.956, with subsequent ones 0.027 and below, and first PC heavily dominated by X_1 , however the overall conclusions on the PCs are much the same but less clear-cut.)



R Analysis:

```

> attach(dogmandibles)
> dogmandibles[1:5,]
  length breadth condyle.breadth height molar.length molar.breadth
1    123    10.1             23    23             19             7.8
2    127     9.6             19    22             19             7.8
3    121    10.2             18    21             21             7.9
4    130    10.7             24    22             20             7.9
5    149    12.0             25    25             21             8.4
  first.to.3rd.length first.to.4th.length canine.breadth gender species
1                   32                   33             5.6      1      1
2                   32                   40             5.8      1      1
3                   35                   38             6.2      1      1
4                   32                   37             5.9      1      1
5                   35                   43             6.6      1      1
> dog.pc<-princomp(dogmandibles[, -c(10,11)], cor=T)
>
> summary(dog.pc)
Importance of components:
              Comp.1      Comp.2      Comp.3      Comp.4      Comp.5
Standard deviation  2.6993793 0.85254056 0.58404915 0.43677899 0.38952230
Proportion of Variance 0.8096276 0.08075838 0.03790149 0.02119732 0.01685862
Cumulative Proportion 0.8096276 0.89038602 0.92828751 0.94948483 0.96634345
              Comp.6      Comp.7      Comp.8      Comp.9
Standard deviation  0.35707481 0.296851411 0.262761145 0.135064109
Proportion of Variance 0.01416694 0.009791196 0.007671491 0.002026924
Cumulative Proportion 0.98051039 0.990301585 0.997973076 1.000000000
> print(dog.pc$loadings,digits=1)

Loadings:
              Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
length        -0.4  -0.1  -0.3   0.2   0.2   0.1   0.1   0.1
breadth       -0.3   0.3   0.2   0.2  -0.3   0.4   0.2  -0.6
condyle.breadth -0.3   0.3  -0.7  -0.3  -0.3  -0.2   0.2   0.2
height        -0.3   0.4   0.2   0.6   0.3  -0.2  -0.3   0.3
molar.length  -0.3  -0.1   0.1  -0.4  -0.2   0.3  -0.7   0.1
molar.breadth -0.3   0.1   0.4  -0.4  -0.2  -0.7   0.1  -0.2
first.to.3rd.length -0.3  -0.7   0.1   0.4  -0.4  -0.1   0.1   0.1
first.to.4th.length -0.3  -0.3  -0.2   0.1   0.7   0.1   0.1  -0.4
canine.breadth -0.3   0.1   0.3  -0.3   0.1   0.3   0.5   0.6

```

Note that in **R** the standard deviations on each component are the square roots of the eigenvalues. The rest of these solutions will concentrate on the interpretation of plots. These have been produced in a different package but equivalent one can of course be produced in **R**.

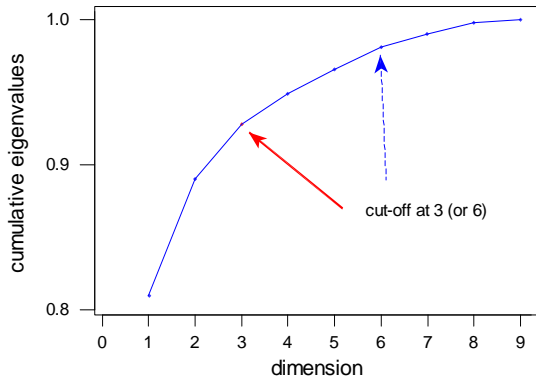


PC1 has coefficients all of the same sign and roughly the same magnitude. Thus low scores will be obtained (in this case, since the signs are all negative) by mandibles with all values of the variables which are large and there will be low values on PC1 when all variable are small. Thus PC1 reflects size and large mandibles will appear at the extreme negative end of the axis, small ones at the positive end.

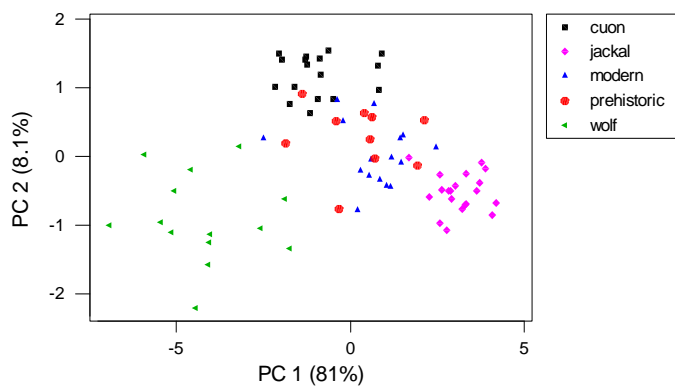
PC2 has negative signs for X_1 , X_5 , X_7 & X_8 and positive signs for the other variables with X_6 & X_9 much smaller. These last two refer to the breadths of the 1st molar and the canine, so these are not important to PC2. Inspection of the variables reveals that those with negative signs are all lengths and those with positive signs are breadths and height so PC2 contrasts lengths with breadths and so can be interpreted as reflecting the **shape** of the mandible. [Aside: these interpretations of size and shape for linear measurements on physical objects are very common and are likely to be appropriate for high order PCs, though not necessarily the second one in the case of shape. This is not entirely for mathematical reasons, just the way the world is. One mathematical reason for size to be predominant is that linear measurements on objects are likely to be positively correlated – end of aside].

Although not asked for, the next steps given here for illustration were to produce a scree plot and plots on the PCs. This is provided here for comparison with the plots on crimcoords and is always a virtually vital step in any analysis for any purpose of multivariate data. The labels included the percentage of variation accounted for by that PC — a useful aid to the interpretation of the scatter plots which might alternatively be obtained by using equal scaling on the axes using `eqscalplot()` in the **MASS** library. The PC plots have the different groups distinguished, even though this information was ignored in the construction of the PCs.



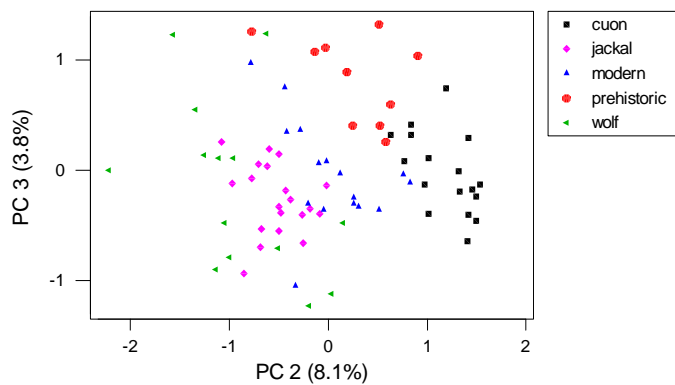


This plot shows that 3 PCs are adequate to capture most of the variation in the data.



The plot on the first two PCs displays 89% of the variation. It separates out the wolves to the left of the plot (i.e. they are **bigger**) and the jackals to the right (i.e. they are **smaller**) than the rest. Note that the

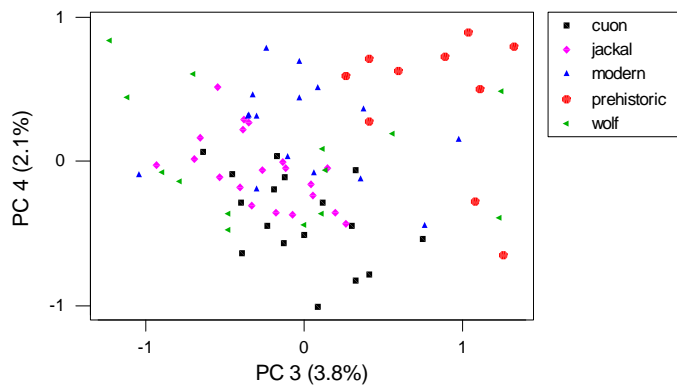
prehistoric and modern dogs are near inseparable on this plot and that this plot displays most of the variation.



The plot on PCs 2 & 3 displays about 12% of the information. It separates the prehistoric (& the cuons) in the top right hand corner. To appear in the top r.h. corner cases have to have large values for those variables

with positive coefficients on **both** PCs 2 & 3 and small values for those with negative coefficients on PCs 2 & 3, i.e. large values for X_2 and X_4 (ignoring any variable with a very small coefficient, even if positive) and small values for X_1 and X_8 , i.e. prehistoric dogs and cuons have short ‘chunky’ mandibles.





The plot on PCs 3 & 4 shows that the prehistoric separates from most groups other than the modern dogs on the 4th PC, though this separation is very slight noting that PC4% contains only 2.1% of the variation. However, the fact

that each of the groups is separated from the others on at least one of these plots suggests that it a discriminant analysis will be able to distinguish them.

ii) Construct a display of the measurements on the first two crimcoords, using different symbols for the five different groups.

```
> library(MASS)
>
> dog.lda<-lda(species~length+breadth+condyle.breadth+height+
+ molar.length+molar.breadth+first.to.3rd.length+
+ first.to.4th.length+canine.breadth)
>
> print(dog.lda,digits=2)
Call:
lda(species ~ length + breadth + condyle.breadth + height +
molar.length +
  molar.breadth + first.to.3rd.length + first.to.4th.length +
  canine.breadth)
```

Prior probabilities of groups:

1	2	3	4	5
0.21	0.26	0.22	0.18	0.13

Group means:

	length	breadth	condyle.breadth	height	molar.length	molar.breadth
1	125	9.7		21	21	19
2	111	8.2		19	17	18
3	133	10.7		24	24	21
4	157	11.6		26	25	25
5	123	10.3		20	23	19
	first.to.3rd.length	first.to.4th.length	canine.breadth			
1		32		37		5.9
2		30		33		4.8
3		29		38		6.6
4		40		45		7.4
5		33		36		6.2



Coefficients of linear discriminants:

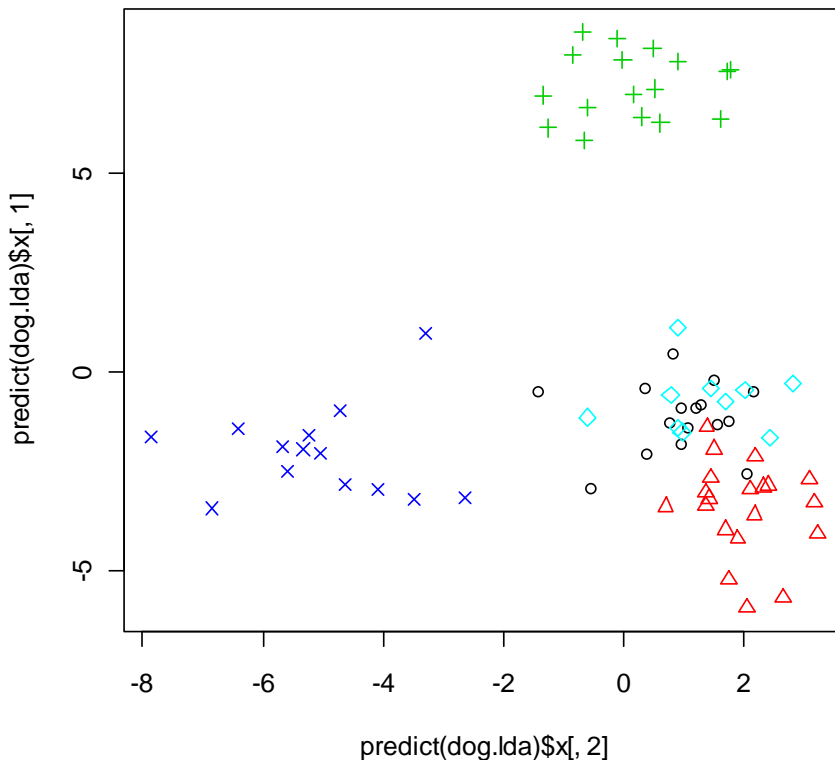
	LD1	LD2	LD3	LD4
length	0.150	-0.027	-0.079	-0.015
breadth	-0.042	0.024	0.552	0.093
condyle.breadth	-0.347	-0.024	-0.087	-0.282
height	0.226	0.051	0.432	0.058
molar.length	0.885	-0.746	-1.131	0.680
molar.breadth	0.818	0.118	0.415	1.057
first.to.3rd.length	-1.375	-0.181	0.338	0.018
first.to.4th.length	-0.239	-0.090	0.014	-0.232
canine.breadth	1.512	0.487	1.279	-1.028

Proportion of trace:

LD1	LD2	LD3	LD4
0.6539	0.2563	0.0859	0.0039

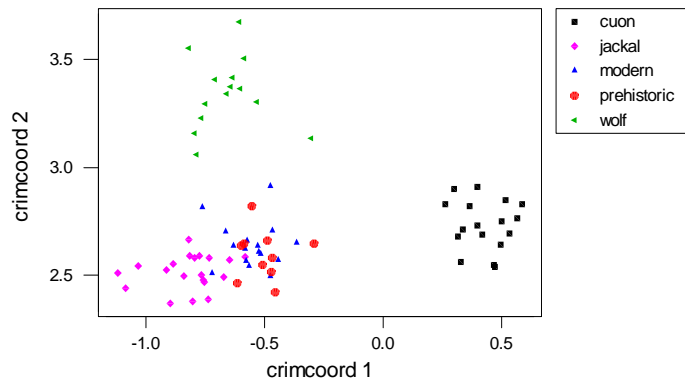
```
type<-unclass(species)
plot(predict(dog.lda)$x[,2],predict(dog.lda)$x[,1], pch=type,col=type)
```

```
type<-unclass(species)
plot(predict(dog.lda)$x[,2],predict(dog.lda)$x[,1], pch=type,col=type)
```



This basic plot needs to be enhanced with a legend and proper labelling of axes and this is done below (though produced in a different plotting package)





This plot shows clear separation of all the groups from each other with the exception of the modern and prehistoric dogs which are intermingled on this display.

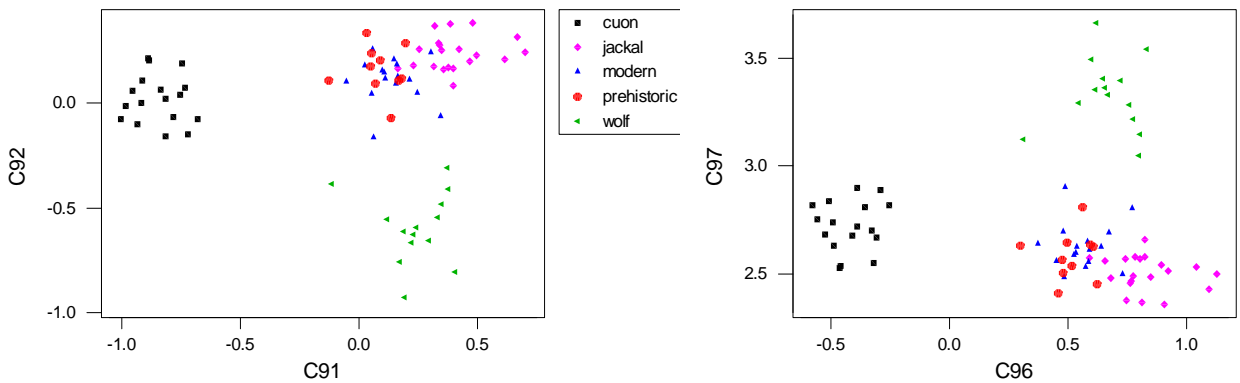
iii) If the linear discriminant analysis were performed on the data after transformation to the full set of nine principal components what differences (if any) would there be in the plot on crimcoords and the eigenvalues and eigenvectors of the matrix $W^{-1}B$?

There are two ways of looking at this. One is to do it and see what happens, the other is to look mathematically, at least initially. Both are useful. However, on general principles, there should be no fundamental difference in the displays since a preliminary transformation to principal components is merely a rotation &/or a reflection of the data and no information is lost or gained. So plots on crimcoords after a PCA transformation should be expected to be essentially identical, up to perhaps a reflection. To get some idea mathematically, suppose the original data matrix is denoted by X' and the matrix of eigenvectors (of either the covariance or the correlation matrix, whichever is used) is denoted by $A=(a_i)$. Then we know that since $a_i'a_i=1$ and $a_i'a_j=0$ for $i \neq j$ that $A'A = I_p$. The data referred to PCs is Y' where $Y'=X'A$. If W and B are the within and between groups variances of the original data X' then those of Y' are $A'WA$ and $A'BA$ respectively. So the crimcoords of the data referred to PCs are the eigenvalues of $(A'WA)^{-1} A'BA$, i.e. of $A^{-1}W^{-1}A'^{-1}A'BA = A'W^{-1}BA$. It is easy to see that the eigenvalues of this are identical to those of $W^{-1}B$. The original data referred to crimcoords are $X'U$ where U is the matrix of eigenvectors of $W^{-1}B$ and the



[PCA transformed-]data referred to crimcoords after the PCA transformation are $Y'V=X'AV$ where V is the matrix of eigenvectors of $A'W^{-1}BA$. It can be shown (but not here) that these differ only in scale and an arbitrary sign difference.

The try it and see approach is straightforward and is not given in detail here. The plots below are again in a different package and axes are not labelled (since this is just for a quick verification that nothing is essentially changed). It can be seen that the three plots are essentially identical except for [arbitrary] reflections.



iv) Which group is separated from the other four by the first crimcoord?

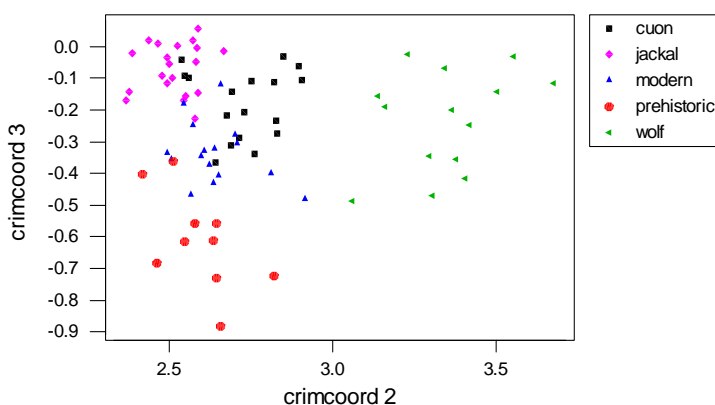
Cuons

v) Which group is separated from the other four by the second crimcoord?

Wolves

vi) Which group is separated from the other four by the third crimcoord?

Need a plot of the third crimcoord:



This shows that the third crimcoord separates the prehistoric dogs from the rest. Further, the third in conjunction



with the second separates the jackals from the rest.

- vii) *What features of the mandibles provide discrimination between the various species?*

The first three crimcoords (from the original data) are

	LD1	LD2	LD3
length	0.150	-0.027	-0.079
breadth	-0.042	0.024	0.552
condyle.breadth	-0.347	-0.024	-0.087
height	0.226	0.051	0.432
molar.length	0.885	-0.746	-1.131
molar.breadth	0.818	0.118	0.415
first.to.3rd.length	-1.375	-0.181	0.338
first.to.4th.length	-0.239	-0.090	0.014
canine.breadth	1.512	0.487	1.279

High scores on crimcoord 1 are obtained by those cases which have big teeth, long mandibles and short distances between molars and premolars. These characteristics distinguish cuons from the others.

High scores on crimcoord 2 are obtained by long narrow mandibles with long narrow teeth, these characteristics distinguish wolves from the other species [rather more specifically than just overall size as was deduced from the PCA above].

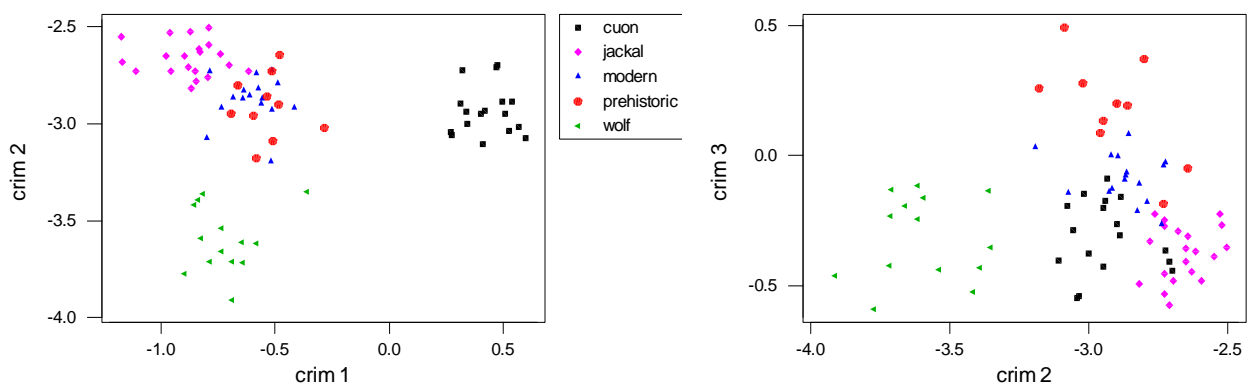
Low scores on crimcoord 3 are obtained by short broad (i.e. 'chunky') molars, short broad ('chunky') mandibles, and longer distances between molars. These features distinguish the prehistoric dogs from the others. [note that this is again a little more specific than obtained from just the PCA].



2) * The question of prime interest in the study of canines was related to an investigation of the origin of the prehistoric dogs. Try calculating the discriminant analysis based on the four groups of modern canines and then plot the prehistoric cases on the same coordinate system a (c.f. informal data classification method (iii) on p140 of course notes) and seeing to which of the modern groups the majority of the prehistoric are closest.

(The interpretation of the results of this exercise are **within** the scope of MAS465; the required computer skills to produce it are useful but a little beyond the scope of PA4370, i.e. if you do not attempt it ensure that you look carefully at the printed solution in due course.)

Below are plots produced in a different package but code to do the equivalent in R is given later



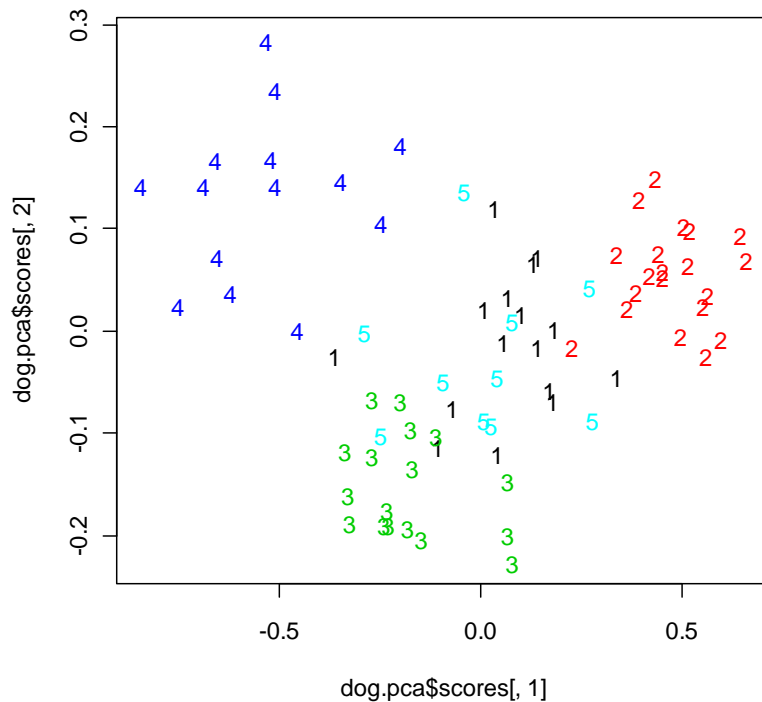
This shows that the prehistoric are superimposed on the modern on crimcoords 1 & 2 but there is a distinction on crimcoord 3.

Below is guidance on producing the plots in **R**. One difficulty is that to add in points for the prehistoric samples onto existing plots on crimcoords for the modern dogs you may need to extend the plotting range or avoid trying to plot points outside the plotting area (hence use of the parameter `ylim=c(.,.)` below). Note also the removal from both the PCA and the LDA the columns 10 and 11 which are factors indicating gender and species. The code and examples below do not provide complete solutions to Q1 and Q2 but are



intended as sufficient for you to adapt to your particular needs. Just for illustration and for comparison, the analyses below have been done after taking [natural] logs of all measurements.

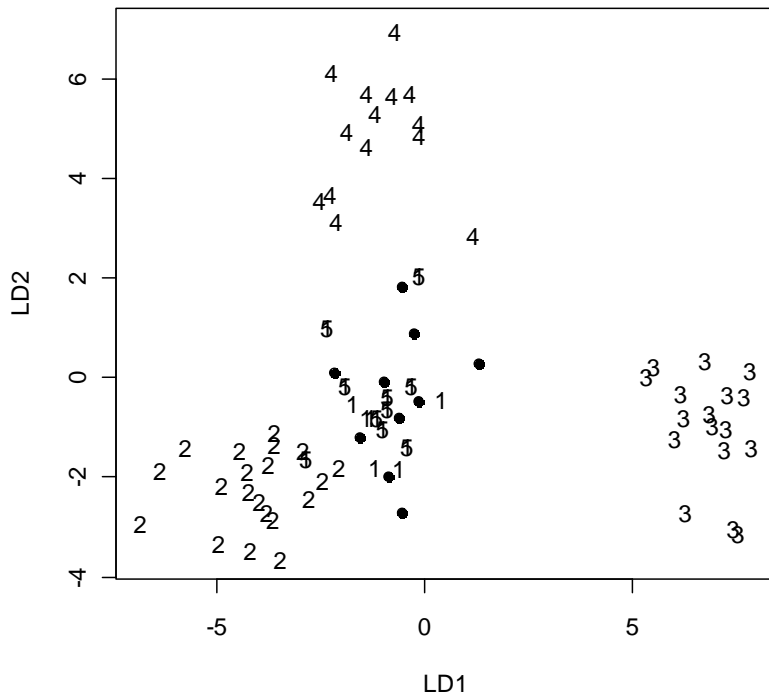
```
> attach(dogmandibles)
> library(MASS)
> □
> dog.pca<-princomp(log(as.matrix(dogmandibles[-c(10,11)])))
>
> plot(dog.pca$scores[,1],dog.pca$scores[,2],type="n")
> text(dog.pca$scores[,1],dog.pca$scores[,2],labels=species,col=type)
>
```



```
>
> mod <-
log(as.matrix(dogmandibl
es[1:67, -c(10,11)]))
> pre <- log(as.matrix(dogmandibles[68:77, -c(10,11)]))
> spec<-species[1:67]
> mod.lda <- lda(mod, spec)
Warning message:
In lda.default(x, grouping, ...) : group 5 is empty
```



```
> plot(predict(mod.lda, dimen = 2)$x, type="n")
> text(predict(mod.lda)$x[,1],predict(mod.lda)$x[,2], labels=species)
> points(predict(mod.lda,pre, dimen= 2)$x, pch=19)
>
```



```
> plot(predict(mod.lda)$x[,2], predict(mod.lda)$x[,3], type="n", ylim=c(-7, 4))
> text(predict(mod.lda)$x[,2], predict(mod.lda)$x[,3], labels=tp)
> points(predict(mod.lda,pre)$x[,2], predict(mod.lda,pre)$x[,3], pch=19)
>
```

