R Homework 8: Multivariate Sexual Dimorphism Analysis

Marguerite Butler

March 20, 2018

DUE: Tuesday April 3rd

This week the homework will focus on multivariate analyses using data from both sexes of 15 species of *Anolis* lizards. The basic questions are are they differentiated ecologically? Is there evidence that these species really form distinct "ecomorphs"? that is, species that are not necessarily closest relatives but evolving similar morphology because they share the same ecology? For example species which are more sit-and-wait predators need to run fast, and may therefore be selected for long limbs. Another basic question is are sexes different? Do they differ in size? in shape? Are the ecomorphs different in sexual dimorphism?

Read through the assignment, run the example code, then write a script to do the exercises. Produce all of the code needed to do these analyses from reading in the data to the end of analysis. You may do one or several scripts. Please answer questions posed in the exercises in the script using comments as usual.

1 The data

The morphometric data were collected by Marguerite Butler on Caribbean *Anolis* lizard species from the islands of Jamaica and Puerto Rico. On these islands, a very special phenomenon has taken place whereby the same ecological types or "ecomorphs" have evolved repeatedly on multiple islands. This independent evolution of the same ecomorph types is one of the rare examples of replicate adaptive radiation.

Much work has established convergent evolution in morphology to match habitat type amongst males of the species, yet males and females are sexually dimorphic. If males are adapted to their habitats, but females differ from males, then does that mean females are "less adapted"? Or differently adapted than males? These questions were explored in Butler et al. (2007); Butler and Losos (2002).

The variables in this dataset are:

SVL Snout-to-vent length, or body length in mm.

CMASS mass in g.

HINDL Hind limb length in mm.

FOREL Fore limb length in mm.

LAMN Lamellae number, the number of sub digital scales on the fourth digit of the hind foot (the longest toe). Lamellae number correlates with clinging ability in these arboreal lizards.

Ecomorph Ecomorph class, based on their microhabitat type, a factor.

Species There are 15 species in this dataset from the two islands of Puerto Rico and Jamaica. Each species occurs on only one island. Species is a factor.

Sex a factor.

These data are saved in a file named `anolisShapeRaw.csv`:

```
> morph <- read.csv("anolisShapeRaw.csv")
> names(morph) <- c("obs", "ecomorph", "species", "sex", "SVL", "mass", "hindl", "forel", "lamn")
> head(morph)
```

```
  obs ecomorph species sex SVL mass hindl forel lamn
1   1       CG      CU   F 118 37.0  87.0  55.0   30
2   2       CG      CU   F 117 46.2  88.0  53.0   30
3   3       CG      CU   F 119 32.0  91.0  53.0   28
4   4       CG      CU   M 136 51.0 100.0  63.0   31
5   5       CG      CU   M 127 40.5  97.5  59.0   30
6   6       CG      CU   M 133 40.5 102.0  60.5   29
```

How many species and ecomorphs are there?

```
> sp <- unique(morph$species)
> sp
```

```
 [1] CU GA KR PO PU EV GR OP ST CR GU LI SA OC VA
Levels: CR CU EV GA GR GU KR LI OC OP PO PU SA ST VA
```

```
> length(sp)
```

```
[1] 15
```

```
> eco <- unique(morph$ecomorph)
> eco
```

```
[1] CG GB TC TG TW
Levels: CG GB TC TG TW
```

How many observations per species and sex?

```
> aggregate(morph, by=list(morph$species, morph$sex), length)
```

	Group.1	Group.2	obs	ecomorph	species	sex	SVL	mass	hindl	forel	lamn
1	CR	F	19	19	19	19	19	19	19	19	19
2	CU	F	3	3	3	3	3	3	3	3	3
3	EV	F	19	19	19	19	19	19	19	19	19
4	GA	F	12	12	12	12	12	12	12	12	12
5	GR	F	21	21	21	21	21	21	21	21	21
6	GU	F	19	19	19	19	19	19	19	19	19
7	KR	F	19	19	19	19	19	19	19	19	19
8	LI	F	24	24	24	24	24	24	24	24	24
9	OC	F	10	10	10	10	10	10	10	10	10
10	OP	F	21	21	21	21	21	21	21	21	21
11	PO	F	6	6	6	6	6	6	6	6	6
12	PU	F	20	20	20	20	20	20	20	20	20
13	SA	F	25	25	25	25	25	25	25	25	25
14	ST	F	26	26	26	26	26	26	26	26	26
15	VA	F	29	29	29	29	29	29	29	29	29
16	CR	M	20	20	20	20	20	20	20	20	20
17	CU	M	6	6	6	6	6	6	6	6	6
18	EV	M	16	16	16	16	16	16	16	16	16
19	GA	M	10	10	10	10	10	10	10	10	10
20	GR	M	18	18	18	18	18	18	18	18	18
21	GU	M	17	17	17	17	17	17	17	17	17
22	KR	M	18	18	18	18	18	18	18	18	18
23	LI	M	21	21	21	21	21	21	21	21	21
24	OC	M	4	4	4	4	4	4	4	4	4
25	OP	M	18	18	18	18	18	18	18	18	18
26	PO	M	6	6	6	6	6	6	6	6	6
27	PU	M	19	19	19	19	19	19	19	19	19
28	SA	M	21	21	21	21	21	21	21	21	21
29	ST	M	11	11	11	11	11	11	11	11	11
30	VA	M	15	15	15	15	15	15	15	15	15

It's not a balanced dataset, but it's a large dataset with good numbers for most species and sex classes, and at least we don't have any missing values.

1.1 Size-adjusting the data

The data should be log-transformed as well as size-adjusted for morphometric analysis.

```
> logmorph <- log(morph[c("SVL", "mass", "hindl", "forel", "lamn")])
```

SVL is a common measure of size in lizard studies, but we are actually interested in sexual dimorphism in SVL. We will use geometric mean size as a proxy for size. Four of our variables scale strongly with size, lamellae number does not. So we include the remaining four in our size variable:

$$SIZE = \frac{log(SVL) + log(mass) + log(hindl) + log(forel)}{4} \quad (1)$$

We make size-adjusted variables by taking the log-ratio of each variable with size:

$$hindl_{size} = log(hindl) - SIZE \quad (2)$$

```
> species <- morph$species
> sex <- morph$sex
> ecomorph <- morph$ecomorph
> size <- with(logmorph, (SVL+mass+hindl+forel)/4)
> sizedat <- with( logmorph, cbind( cbind(SVL, mass, hindl, forel) - size, lamn ))
> sizedat <- cbind(sizedat, size)
```

And aggregate by species, sex, and ecomorph to make a size by species-sex dataset:

```
> sizesp <- aggregate(sizedat, by=list(species, sex, ecomorph), mean)
> head(sizesp)
```

```
  Group.1 Group.2 Group.3       SVL       mass     hindl       forel     lamn     size
1      CU       F      CG 0.5520243 -0.5820919 0.2660651 -0.23599742 3.378200 4.218636
2      GA       F      CG 0.7875633 -1.1287361 0.4207488 -0.07957591 3.339941 3.571456
3      CU       M      CG 0.5160202 -0.5028535 0.2496686 -0.26283529 3.400827 4.378917
4      GA       M      CG 0.5713207 -0.5102907 0.2024824 -0.26351234 3.399541 4.172789
5      KR       F      GB 1.1078418 -2.2849560 0.9100335  0.26708073 2.953816 2.508219
6      PO       F      GB 1.1602322 -2.2517212 0.8266211  0.26486795 2.802901 2.446125
```

```
> names(sizesp)[1:3] <- c("species", "sex", "ecomorph")
```

2 Do ecomorphs differ in SIZE dimorphism?

Do the ecomorphs differ in sexual SIZE dimorphism? Let's take a look at our size variable. First let's do an ANOVA:

```
> summary(aov(with(sizesp, lm(size~ sex*ecomorph))))
```

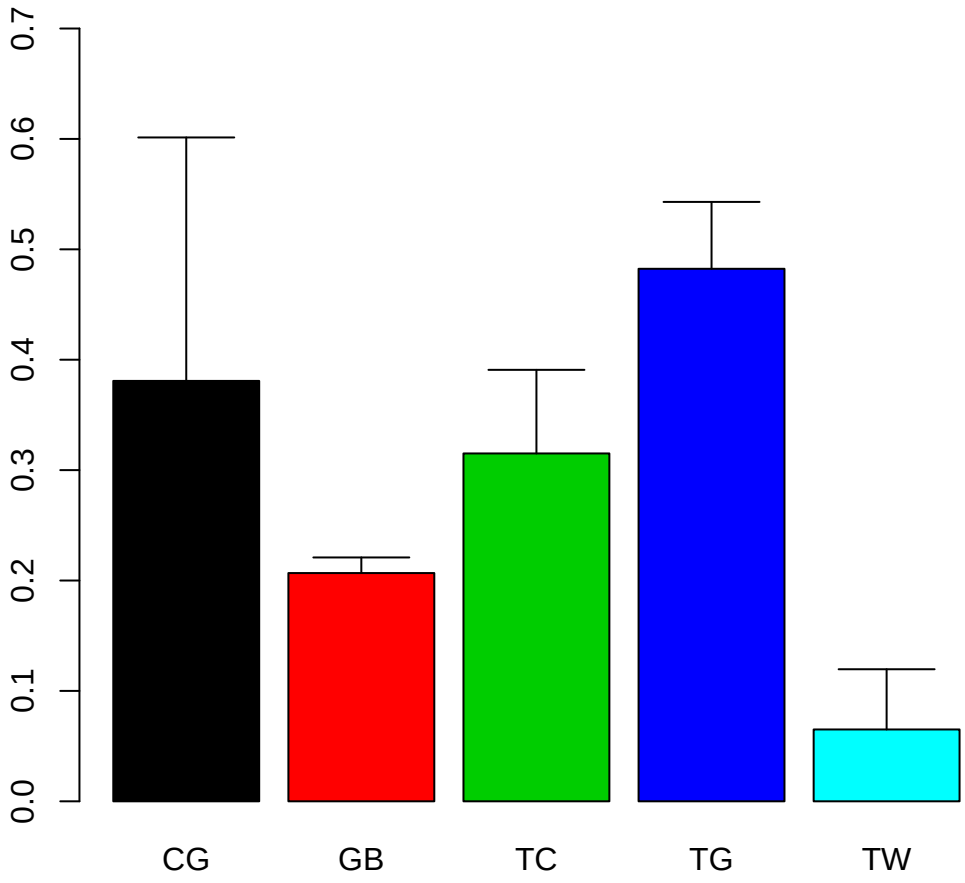
```
             Df Sum Sq Mean Sq F value   Pr(>F)
sex           1  0.737  0.7369   9.715  0.00543 **
ecomorph      4  7.134  1.7835  23.510 2.54e-07 ***
sex:ecomorph  4  0.140  0.0351   0.463  0.76235
Residuals    20  1.517  0.0759
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sexes differ in size, and ecomorphs differ in size after accounting for sex differences, but there is no interaction between the two. How about dimorphism? Let's first make a dimorphism dataset, with one row per species (all variables here are differences between males and females).

```
> sizespm <- sizesp[sizesp$sex=="M",]
> sizespf <- sizesp[sizesp$sex=="F",]
> sizediff <- sizespm[4:9] - sizespf[4:9]
> sizediff <- cbind(sizespf[c(1,3)], sizediff)
> with( sizediff, t.test(size))


        One Sample t-test

data:  size
t = 6.4371, df = 14, p-value = 1.555e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.2090204 0.4179055
sample estimates:
mean of x
 0.313463
```

A t-test shows that ecomorphs do indeed differ in sexual size dimorphism. So there is something about the habitat that is driving the magnitude of size differences between sexes. Let's make a plot to illustrate. Aggregate to make barplots for mean ssd between ecomorphs:

```
> smean <- aggregate(sizediff$size, by=list(sizediff$ecomorph), mean)
> ssd <- aggregate(sizediff$size, by=list(sizediff$ecomorph), sd)
> sn <- aggregate(sizediff$size, by=list(sizediff$ecomorph), length)
> names(smean) <- c("ecomorph", "sexdim")
> names(ssd) <- c("ecomorph", "sd")
> names(sn) <- c("ecomorph", "sn")
> sizediff.eco <- merge(smean, ssd)
> sizediff.eco <- merge(sizediff.eco, sn)
> sizediff.eco$se <- with(sizediff.eco, sd/sqrt(sn))
> bb <- with(sizediff.eco, barplot(sexdim, names.arg=ecomorph,
+   col=ecomorph, ylim=c(0, max(sizediff.eco$sexdim)*1.5)))
> with(sizediff.eco, arrows(bb, sexdim, bb, sexdim+se, angle=90))
```



2.1 Exercises

1. Produce the plot above but make the color code for the ecomorphs match the scheme used below for PCA (read next section).

2. Repeat the analysis using logSVL as the proxy for size.

3 Do ecomorphs differ in shape?

4 Principal Components Analysis

We're going to do a principal components analysis on our size-corrected dataset. In lecture we talked about the fact that with multiple species (groups), we have both within and between species variation. We will first deal with this within vs between group variation by taking species-sex means and using those in the PCA (basically collapsing or ignoring within-group variation).

First obtain species mean data:

```
> sizesp <- aggregate(sizedat, by=list(species, sex, ecomorph), mean)
> names(sizesp)[1:3] <- c("species", "sex", "ecomorph")
```

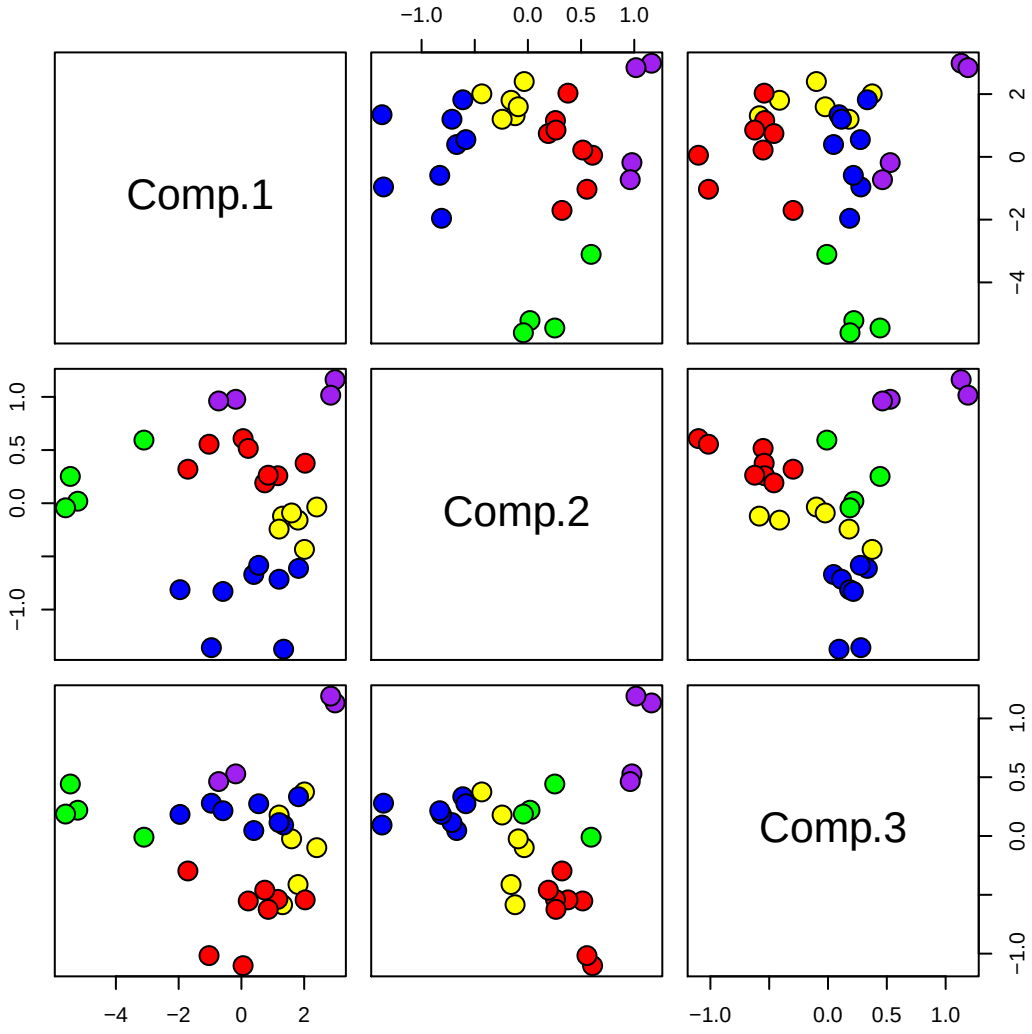
Run a PCA on the species mean data:

```
> vars <- c("SVL", "mass", "hindl", "forel", "lamn", "size")
> cols <- c("green", "yellow", "red", "blue", "purple")
> pca.size <- princomp(sizesp[vars], cor=TRUE, scores=TRUE)
> pca.size

Call:
princomp(x = sizesp[vars], cor = TRUE, scores = TRUE)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
2.28904054 0.66060121 0.52782871 0.19839353 0.07704751 0.00000000

 6  variables and  30 observations.

> summary(pca.size)

Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4       Comp.5 Comp.6
Standard deviation     2.2890405 0.66060121 0.52782871 0.198393528 0.0770475149      0
Proportion of Variance 0.8732844 0.07273233 0.04643386 0.006559999 0.0009893866      0
Cumulative Proportion  0.8732844 0.94601676 0.99245061 0.999010613 1.0000000000      1

> print(loadings(pca.size), cutoff=0)

Loadings:
      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
SVL    0.389  0.564  0.483 -0.156  0.377 -0.362
mass  -0.434 -0.140  0.092  0.062 -0.247 -0.848
hindl  0.410 -0.324 -0.458 -0.652  0.105 -0.284
forel  0.419  0.026 -0.462  0.717  0.163 -0.264
lamn  -0.362  0.710 -0.576 -0.180 -0.026  0.000
size  -0.431 -0.229 -0.058  0.024  0.871  0.000
```

```
                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
SS loadings      1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var   0.167  0.167  0.167  0.167  0.167  0.167
Cumulative Var   0.167  0.333  0.500  0.667  0.833  1.000
```

A quick and dirty pairwise plot function is pairs. A pairwise plot of the data shows that the ecomorphs are well separated by PC1, PC2, and PC3. Crown-giant anoles are low in PC1 and separated from the other ecomorphs, and the remaining separate along PC2 and/or PC3.

```
> scores <- as.data.frame(pca.size$scores)
> pairs(scores[1:3], pch=21, bg=cols[unclass(sizesp$ecomorph)],cex=2)
```



The work of Losos (1990) on the correlated evolution of morphology, ecology, and behavior showed that the species cluster by ecomorph. Let's plot the males and see if we can see the same pattern. As a quick and

dirty solution, let's draw a minimum-convex polygon type graph by "connecting the dots" around ecomorphs. We can use the `polygon` function to do this. In a proper minimum convex polygon you would only connect the outermost points. Let's plot the females using a different plot symbol to see if they fall within the "cloud" of points formed by the males.

Let's organize the data and make some index vectors to make things more convenient. We will want to make several plots with males and females distinguished on the same plot. So let's break up the species scores dataset into males and females:

```
> scoresm <- scores[sizesp$sex=="M",]
> scoresf <- scores[sizesp$sex=="F",]
> head(scoresm)
```

```
      Comp.1      Comp.2      Comp.3      Comp.4       Comp.5        Comp.6
3  -5.603687 -0.04280747  0.18701779 -0.11390793  0.172588470 -1.615760e-15
4  -5.452546  0.25245169  0.44258559  0.00441986 -0.081449631  2.981980e-15
8   1.311250 -0.12077122 -0.58399229 -0.46541175  0.082525414  1.119557e-16
9   1.200376 -0.24258562  0.17876241 -0.16937298 -0.007100267 -1.940072e-16
10  1.597786 -0.09102853 -0.02382757 -0.40074406  0.034351378  8.200098e-16
15 -1.031502  0.55539499 -1.01650042  0.08461626  0.070061371 -4.878992e-18
```

```
> head(scoresf)
```

```
        Comp.1      Comp.2       Comp.3      Comp.4      Comp.5       Comp.6
1  -5.21485194  0.01937789  0.219716856 -0.07811873  0.05579945 9.450159e-16
2  -3.10585465  0.59415591 -0.009791812 -0.18029493 -0.08152823 1.523868e-15
5   1.80423971 -0.15848684 -0.412124743 -0.48271475 -0.03572067 1.990690e-15
6   2.00890368 -0.43380299  0.375256717 -0.08937460 -0.09178047 1.299044e-15
7   2.40144298 -0.03497523 -0.098829399 -0.26141687 -0.01701307 7.862874e-16
11  0.05291868  0.60815956 -1.101667287  0.03066181  0.03839053 4.339721e-16
```

Notice that the species, sex, and ecomorph variables are not in the scores datasets. They are just numerical. But these match exactly the `sizesp` dataset, so we can use the categorical variables from `sizesp`.

Next, to make plots prettier we want to color the points by ecomorph. That's easy, but in order to do that we need index vectors for ecomorphs and for colors that match the scores by sex datasets.

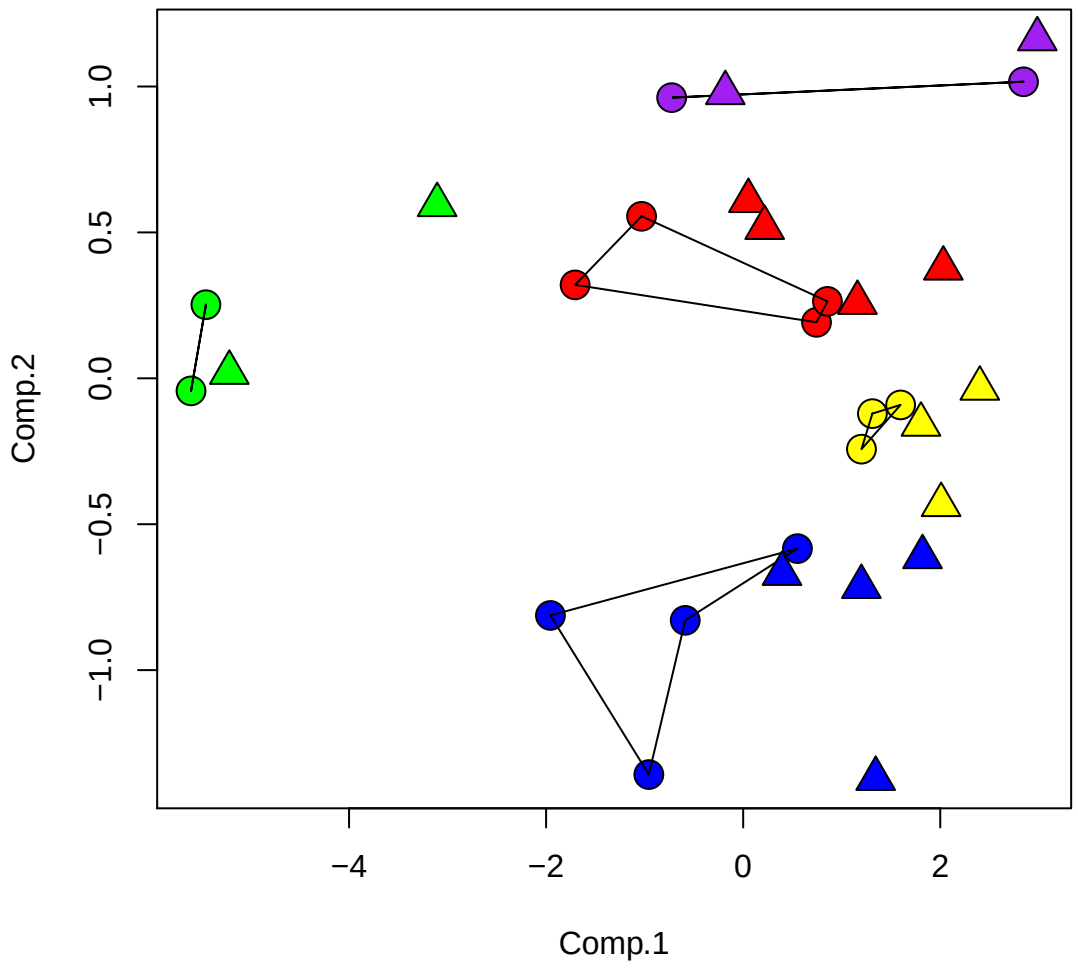
```
> ecos <- sizesp$ecomorph[sizesp$sex=="M"]
> colss <- cols[unclass(sizesp$ecomorph[sizesp$sex=="M"])]
> cbind(ecos, colss)
```

```
     ecos colss
 [1,] "1"  "green"
 [2,] "1"  "green"
 [3,] "2"  "yellow"
 [4,] "2"  "yellow"
 [5,] "2"  "yellow"
 [6,] "3"  "red"
 [7,] "3"  "red"
```

```
 [8,] "3"  "red"
 [9,] "3"  "red"
[10,] "4"  "blue"
[11,] "4"  "blue"
[12,] "4"  "blue"
[13,] "4"  "blue"
[14,] "5"  "purple"
[15,] "5"  "purple"
```

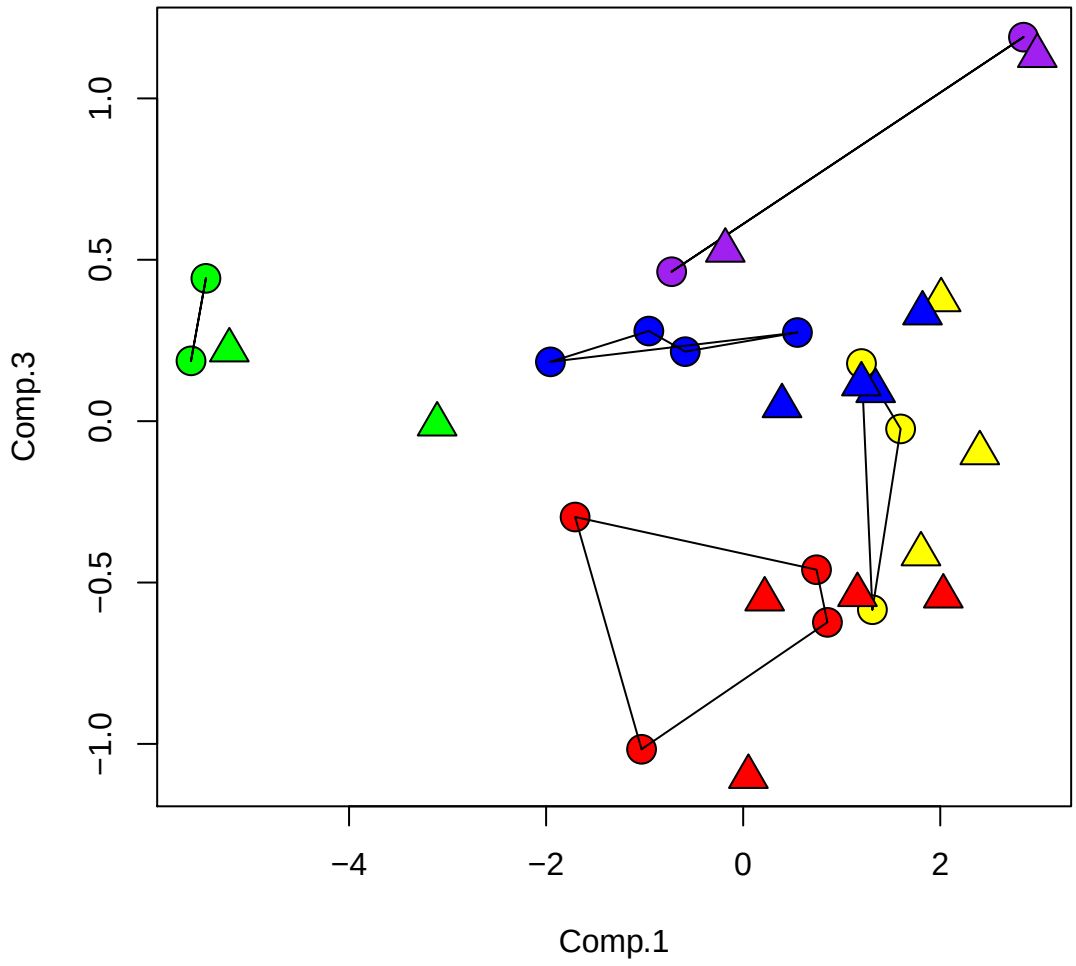
Let's plot PCA1 vs 2. Do the females fall outside of the polygons defined by males? Hmm....

```
> with(scores, plot(Comp.2 ~ Comp.1, type="n")) # make axes big enough to fit all points
> with(scoresm, points(Comp.2 ~ Comp.1, pch=21, bg=colss, cex=2))
> polygon( scoresm[ecos=="CG", ][1:2])
> polygon( scoresm[ecos=="GB", ][1:2])
> polygon( scoresm[ecos=="TC", ][1:2])
> polygon( scoresm[ecos=="TG", ][1:2])
> polygon( scoresm[ecos=="TW", ][1:2])
> with(scoresf, points(Comp.2 ~ Comp.1, pch=24, bg=colss, cex=2))
```



And PCA 1 vs 3:

```
> with(scores, plot(Comp.3 ~ Comp.1, type="n")) # make axes big enough to fit all points
> with(scoresm, points(Comp.3 ~ Comp.1, pch=21, bg=colss, cex=2))
> polygon( scoresm[ecos=="CG", ][c(1,3)])
> polygon( scoresm[ecos=="GB", ][c(1,3)])
> polygon( scoresm[ecos=="TC", ][c(1,3)])
> polygon( scoresm[ecos=="TG", ][c(1,3)])
> polygon( scoresm[ecos=="TW", ][c(1,3)])
> with(scoresf, points(Comp.3 ~ Comp.1, pch=24, bg=colss, cex=2))
```



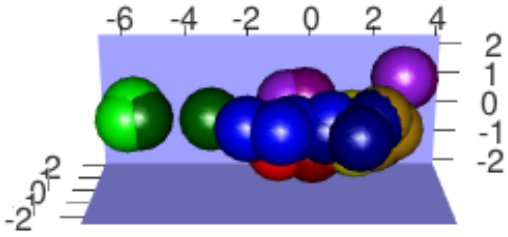
Redo the PCA analysis without size (exclude size from the list of variables):

```
> vars <- c("SVL", "mass", "hindl", "forel", "lamn")
> cols <- c("green", "yellow", "red", "blue", "purple")
> pca.size <- princomp(sizesp[vars], cor=TRUE, scores=TRUE)
> pca.size

Call:
princomp(x = sizesp[vars], cor = TRUE, scores = TRUE)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
2.0675921 0.6391446 0.5265197 0.1983286 0.0000000

 5  variables and  30 observations.
```

```
> summary(pca.size)

Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4 Comp.5
Standard deviation     2.0675921 0.63914463 0.52651973 0.198328553      0
Proportion of Variance 0.8549874 0.08170117 0.05544461 0.007866843      0
Cumulative Proportion  0.8549874 0.93668855 0.99213316 1.000000000      1

> print(loadings(pca.size), cutoff=0)

Loadings:
      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
SVL    0.423  0.618  0.529  0.167 -0.362
mass  -0.479 -0.207  0.067 -0.069 -0.848
hindl  0.458 -0.263 -0.460  0.655 -0.284
forel  0.464  0.101 -0.445 -0.712 -0.264
lamn  -0.408  0.704 -0.553  0.179  0.000

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
SS loadings       1.0    1.0    1.0    1.0    1.0
Proportion Var    0.2    0.2    0.2    0.2    0.2
Cumulative Var    0.2    0.4    0.6    0.8    1.0

> scores <- as.data.frame(pca.size$scores)
```

One very nice feature of R is all of the great packages available. `rgl` is a package that produces excellent 3d graphics. It may be helpful to look at these plots in 3d. After making these plots, you can interactively grab the figure and turn it around to examine the relative positions of males, females, and different ecomorphs. You will need to install the package `rgl`. The function `rgl.snapshot` can save the figure to file, as a png. In order to plot different perspectives, check out the help page for `rgl.viewpoint` http://rgl.neoscientists.org/arc/doc/RGL_INTERFACE03.pdf

```
> require(rgl)
> fcols <- c("darkgreen", "darkgoldenrod", "darkred", "darkblue", "darkmagenta")
> colss2 <- fcols[unclass(sizesp$ecomorph[sizesp$sex=="F"])]
> open3d()

[1] 2

> with(rbind(scoresm, scoresf), spheres3d(Comp.1, Comp.2, Comp.3, col=c(colss, colss2)))
> bbox3d(color=c("#333377","black"), emission="#333377",
+           specular="#3333FF", shininess=5, alpha=0.8)
> rgl.snapshot("anolis_pca3d.png")
```



4.1 Exercises

You may want to look back at the lecture pdf and other examples for how to do PCA.

3. Do a PCA analysis on species-sex mean data (means by species and sex), and compare log-transformed data (i.e. the data contains information on both size and shape) with a PCA on size-adjusted data (shape only). How much variance is explained by size?

4. Are ecomorphs separated in both PCA analyses? Along which axes are they separated? Answer the question and illustrate with a plot.

5. Compare the loadings for the two analyses for PC axes 1, 2, and 3. Can you interpret them as strongly correlated with any of the original variables?

5 Canonical Discriminant Analysis

You will need the package `candisc` to execute this code. First, we set up a multivariate linear model on which to run canonical discriminant analysis. We can test the significance (are they significantly different in multivariate shape morphology?) using MANOVA. We choose to discriminate by species and sex because this is the smallest biological unit in our data. We will take the strategy of finding the best function that discriminates species-sex groups, and then seeing if there are any interesting patterns by ecomorph in the plots.

```
> require(candisc)
> sizedat <- as.data.frame(sizedat)
> shape.manova <- lm(cbind(SVL, hindl, forel, lamn) ~ species + sex, data=sizedat)
> Anova(shape.manova, test="Wilks")

Type II MANOVA Tests: Wilks test statistic
        Df test stat approx F num Df den Df    Pr(>F)
species 14   0.00100   161.57     56 1845.9 < 2.2e-16 ***
```

```
sex      1   0.45924   139.54      4  474.0 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Now it's very simple to run `candisc`

```
> anolis.can <- candisc(shape.manova)
> anolis.can

Canonical Discriminant Analysis for species:

   CanRsq Eigenvalue Difference Percent Cumulative
1 0.92169    11.7702     2.4295 47.7239     47.724
2 0.90329     9.3407     2.4295 37.8733     85.597
3 0.68420     2.1665     2.4295  8.7845     94.382
4 0.58082     1.3856     2.4295  5.6183    100.000

Test of H0: The canonical correlations in the
current row and all that follow are zero

  LR test stat approx F num Df den Df   Pr(> F)
1      0.00100  161.575     56 1845.9 < 2.2e-16 ***
2      0.01280  121.134     39 1407.3 < 2.2e-16 ***
3      0.13238   69.357     24  952.0 < 2.2e-16 ***
4      0.41918   60.086     11  477.0 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The correlations of the canonical axes with the original variables:

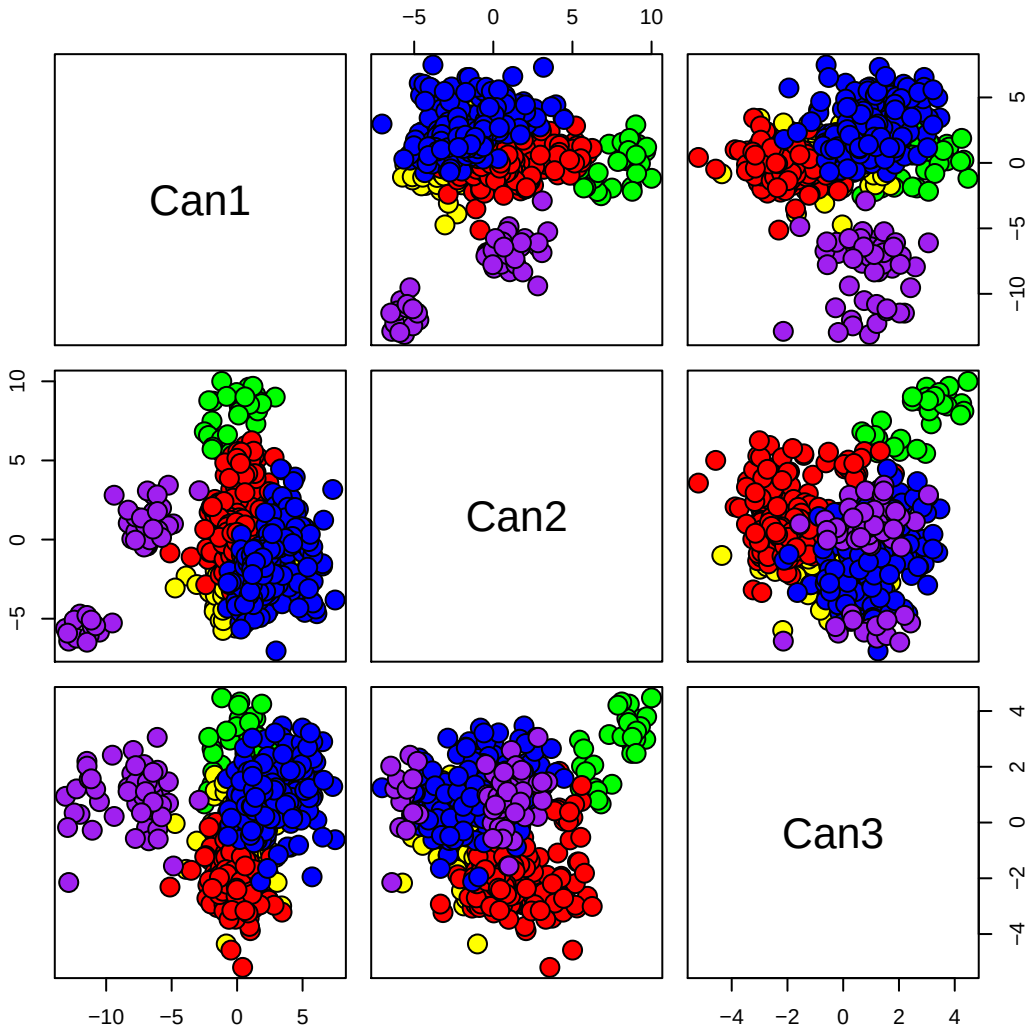
```
> anolis.can$structure

             Can1       Can2       Can3        Can4
SVL   -0.64809876 -0.7219738 -0.3825005  0.08452294
hindl  0.17743263 -0.7857271 -0.5583072  0.24338371
forel -0.05582863 -0.6840990 -0.7701330 -0.11834945
lamn  -0.22545432  0.9231496 -0.1467869  0.18801184
```

Let's plot and see if ecomorphs cluster. First we want to save the scores as a dataframe and add the ecomorph column back:

```
> scores <- as.data.frame(anolis.can$scores)
> scores <- cbind(scores, ecomorph)
```

We want to make a color vector that indicates ecomorph class, and then plot (quick and dirty with pairs function):

```
> ecolsi <- cols[unclass(scores$ecomorph)]
> pairs(scores[3:5], pch=21, bg=ecolsi,cex=2)   # rough look at what's going on
```

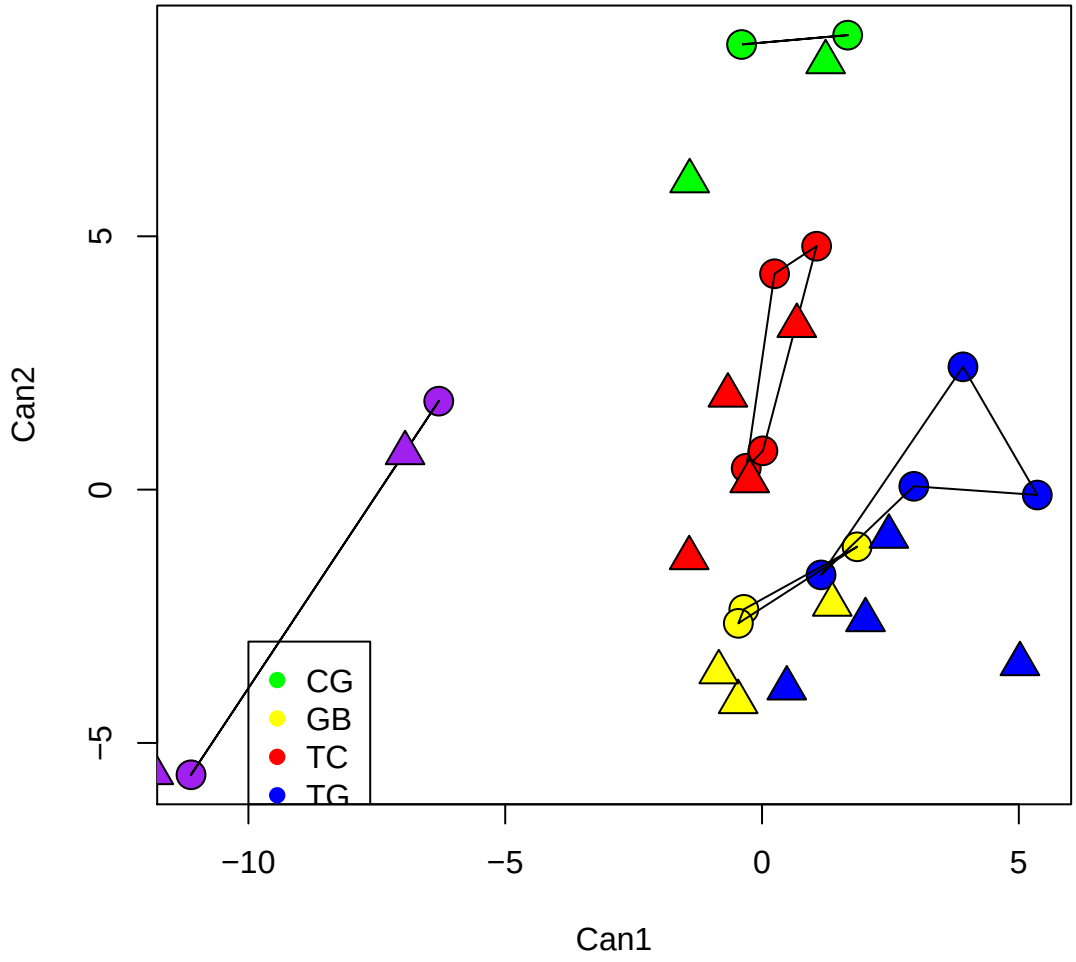


Now let's take species sex means and produce plots similar to the pca exercise:

```
> cansp <- aggregate(scores, by=list(species, sex, ecomorph), mean)
> names(cansp)[1:3] <- c("species", "sex", "ecomorph")
> msp <- cansp$sex=="M"
> fsp <- cansp$sex=="F"
> ecolssp <- cols[unclass(cansp$ecomorph)]
> with(cansp[msp,], plot(Can2 ~ Can1, pch=21, bg=ecolssp[msp], cex=2))
> lapply( as.character(eco), function(e) {
+   polygon( cansp[msp & cansp$ecomorph == e , ][c("Can1", "Can2")]) })

[[1]]
NULL
```

```
[[2]]
NULL

[[3]]
NULL

[[4]]
NULL

[[5]]
NULL
```

```
> with(cansp[fsp,], points(Can2 ~ Can1, pch=24, bg=ecolssp[fsp], cex=2))
> legend(-10, -3, legend=as.character(eco), col=cols[eco], pch=19)
>
```



5.1 Exercises

6. Produce a canonical discriminant analysis for shape and compare to the shape PCA.
7. Are ecomorphs well-separated? Compare Canonical Discriminant Analysis and PCA.
8. What are the differences in the loadings between the two analyses? Are they very similar or not?

6 Do males and females have different niches?

If males and females are morphologically differentiated in shape, then a discriminant function analysis on male shape should misclassify females. Discriminant function analysis tries to find a linear combination that best discriminates between groups. It takes into account both within and between group variation and finds

a transformation that optimally separates groups. It then computes the distance between the groups, and measures how far each individual observation is from each group (this is the Mahalanobis distance). Each test observation is then assigned to the group that it is closest to.

Study the following example. We are going to do a DFA on Fisher's iris dataset. Use half of the iris data to train the DF. Take a random sample. The whole dataset is 150 observations (50 per species).

```
> head(iris)  # iris dataset

  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa

> train <- sample(1:150, 75)
> table(iris$Species[train])  # your answer may differ

    setosa versicolor  virginica
        23         27         25
```

Compute a discriminant function using the training dataset (the sample of individuals chosen to train the DF).

```
> z <- lda(Species ~ ., iris, subset=train)
```

Use the discriminant function to predict the rest of the dataset (non-training individuals)

```
> zz <- predict(z, iris[-train,])$class
```

Look at the results. Each row of the table is the actual species (from species), the column is which species it was classified into. Any off-diagonal counts are misclassifications (from species in the row label into species in the column label)

```
> (ct <- table(iris$Species[-train], zz))

            zz
             setosa versicolor virginica
  setosa         27          0         0
  versicolor      0         22         1
  virginica       0          0        25

> diag(prop.table(ct,1))

    setosa versicolor  virginica
 1.0000000  0.9565217  1.0000000

> sum(diag(prop.table(ct)))
```

```
[1] 0.9866667
```

Now, most people don't want to use half of their data to train and only get classifications on half their data. At the same time, you can't include the observations into the training function that you also want to test, that would bias the classification. One way around this problem is to do something like a "jacknife". Exclude one observation. Construct a classification with the remaining observation, then classify the one excluded one. Then move on to the next observation, exclude, train, classify, etc. Go through your entire dataset and count up how many were classified correctly.

```
> train <- 1:150
> result <- NULL
> z <- lda(Species ~ ., iris, subset=train[-1])
> zz <- predict(z, iris[train[1],])
> (ct <- table(iris$Species[1], zz$class))

            setosa versicolor virginica
  setosa          1          0         0
  versicolor      0          0         0
  virginica       0          0         0

> diag(prop.table(ct,1))

    setosa versicolor  virginica
         1        NaN        NaN

> sum(diag(prop.table(ct)))

[1] 1

> result <- c(result, zz$class)
```

We can automate this a bit by creating a function:

```
> mylda <- function(x) {
+   z <- lda(Species ~ ., iris, prior = c(1,1,1)/3, subset=train[-x])
+   zz <- predict(z, iris[train[x],])
+   return(zz$class)
+ }
```

And use the function in an sapply function, operating over the index for each observation:

```
> predicted <- sapply(1:150, mylda)
> (ct <- table(iris$Species, predicted))

            predicted
             setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         2
  virginica       0          1        49
```

```
> diag(prop.table(ct,1))

    setosa versicolor  virginica
      1.00       0.96       0.98

> sum(diag(prop.table(ct)))

[1] 0.98
```

You can actually do the cross-validation using built-in options to lda. Set `CV=TRUE`. It is always good to check with your own code to validate when you can.

```
> z <- lda(Species ~ ., iris, CV=TRUE)
> (ct <- table(iris$Species, z$class))

             setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         2
  virginica       0          1        49

> diag(prop.table(ct,1))

    setosa versicolor  virginica
      1.00       0.96       0.98

> sum(diag(prop.table(ct)))

[1] 0.98
```

6.1 Exercises

9. Using male shape morphology, are ecomorphs morphologically differentiated? Using male shape morphology, do a DFA and classify each individual to species. How many individuals of each species were misclassified to the wrong species? To the wrong ecomorph? (Members of the same ecomorph class are supposed to be very similar in shape). Write code to do cross-validation using both the `CV=TRUE` option as well as using the jack-knife like code similar to above.

10. Do females share the same niches (as evidenced by convergent morphology)? Construct a discriminant function based on all of the male shape morphology data, and classify each female observation. How many are misclassified to the wrong species? To the wrong ecomorph?

11. What do these results tell you about the similarity of male and female morphology? How does it compare to the plots above?

References

Butler, M. A. and J. B. Losos. 2002. Multivariate sexual dimorphism, sexual selection, and adaptation in greater antillean anolis lizards. ECOLOGICAL MONOGRAPHS, **72**:541–559.

Butler, M. A., S. A. Sawyer, and J. B. Losos. 2007. Sexual dimorphism and adaptive radiation in anolis lizards. NATURE, **447**:202–205.

Losos, J. B. 1990. Ecomorphology, performance capability, and scaling of west-indian anolis lizards - an evolutionary analysis. Ecological Monographs, **60**:369–388.