Elaborated Principal Components Analysis

There are many ways to conduct principle components analysis within R. Each ways had its own procedures, advantages, and limitations. Below are 3 different ways of conducting PCA that illustrate many of the points we have discussed in class.

This initial section is the same for all 3 procedures

Load the Doubs dataset

```
doubs.pre <- read.table("C:\\Multi\\doubs.txt", header = TRUE)
```

An assessment of assumptions

PCA is very robust for violations of assumptions (linearity and normality) but is quite sensitive to outliers. Researchers that are confident in their data often do not worry about the assumptions. This is probably okay if you are only use the PCA for description. However, if you are going to use the components in other analyses, then you need to check the assumptions. With PCA it is always wise to check for outliers.

Normality

```
library (car)

scatterplotMatrix(~das + alt + pen + deb + pH + pho + nit + oxy + dbo,
        data = doubs.pre, diag = "boxplot")
```

The boxplots of many of the variables appear as though they might be non-normal. We will try transforming the variables.

```
tDAS <- doubs.pre$das

tALT <- doubs.pre$alt^-0.25

tPEN <- log10(doubs.pre$pen)

tDEB <- sqrt(doubs.pre$deb)

tPH <- log(doubs.pre$pH)

tPHO <- log10(doubs.pre$pho)

tNIT <- sqrt(doubs.pre$nit)
```

```
tOXY <- doubs.pre$oxy

tDBO <- log10(doubs.pre$dbo)

doubs <- data.frame (DAS = tDAS, ALT = tALT, PEN = tPEN, DEB =
        tDEB, PH= tPH, PHO = tPHO, NIT = tNIT, OXY = tOXY, DBO
        = tDBO)

scatterplotMatrix(~DAS + ALT  + PEN + DEB + PH + PHO + NIT +
        OXY + DBO, data = doubs, diag = "boxplot")
```

The transformed values are approximately normally distributed.

Multivariate normality

Univariate normality covered above.

```
 library (mvnormtest)

N <- as.matrix(doubs)

center <- colMeans(N)

n <- nrow(N); p <- ncol(N); cov <- cov(N);

d <- mahalanobis(N,center,cov)

qqplot(qchisq(ppoints(n),df=p),d, main="QQ Plot Assessing Multivariate
        Normality", ylab="Mahalanobis D2")

abline(a=0,b=1)
```

There are few points that are under the line near the far end that are worrisome, but overall the data appears to be multivariate normal.

Linearity

```
scatterplotMatrix(~DAS + ALT  + PEN + DEB + PH + PHO + NIT +
        OXY + DBO, data = doubs, diag = "boxplot")
```

None of the relationships between the variables appear obviously nonlinear. We can proceed.

Outliers

Care has been taken in screening the data. The data is representative of the sites sampled. There are no outliers.

Performing the Principal Components Analysis

Version 1. Vegan

Vegan is an R package that does many of the traditional community ecology analyses. It is important for many of you to become familiar with it. It does have some convenient visualization tools that are built into it. However, many of the features associated with PCA are not included in the vegan version because the author has some strong opinions regarding how it should be done and does not provide alternate capabilities. Vegan also uses an odd vocabulary (e.g. cases are referred to as sites and variables are referred to as species regardless of the dataset being analyzed).

```
library(vegan)

doubs.pca.vg <- rda(doubs, scale = TRUE)

doubs.pca.vg
```

Vegan produces a number of principal components equal to the number of variables in the original dataset. The numbers presented here are the eigenvalues for each of the principal components.

```
summary(doubs.pca.vg, scaling = 1)
```

The scaling variable determines if the comparison is between cases (1) or between variables (2).

These results are easily broken apart. The first section summarizes the model and data. The second part repeats the reporting of the eigenvalues and adds information to assist in its interpretation (proportion of the variation explained). The third part is the species scores (variable coefficients) for the extracted axes. The final part is the site scores (positions of the cases on the extracted axes).

```
 scores(doubs.pca.vg, choices = 1:3, display = "species", scaling = 0)
```

This command provides the loadings for the species (variables) for the first 3 principal components.

Visualizing the solution with vegan

```
biplot(doubs.pca.vg, scaling=1, main="PCA - scaling 1")
```

This command produces a biplot showing the relative position of the sites (cases) as points and species (variables) as vectors. Sites (cases) that are near each other in the ordination space have similar environments (variables or species). Species (variables) with similar vector angles are correlated and more closely a variables vector's direction matches that of an extracted axes the greater its loading on that axis.

```
library(rgl)

ordirgl(doubs.pca.vg, size=2)
```

This snippet of code produces an interactive figure of principal components 1-3. When you are done with the interactive figure, you can exit with:

```
rgl.quit()
```

Version 2. prcomp

```
library(MASS)

predictors <- doubs

fit <- prcomp(predictors, scale=TRUE)

summary(fit)
```

prcomp produces a number of principal components equal to the number of variables in the original dataset. The numbers presented here are modified eigenvalues (in standard deviation units) for each of the principal components and the proportion of the variation explained.

```
fit
```

This command provides the loadings for the species (variables) for the first 3 principal components.

```
biplot(fit)
```

Produces the biplot for the ordination.

```
predict(fit)
```

This command provides the case scores (positions of the cases on the extracted axes).

Version 3. principal

```
library(psych)

pc <- principal(doubs, nfactors = 9, score = TRUE)
```

This command performs the PCA. The nafactors setting specifies the number of axes to extract. The score setting determines if the case scores are produced.

```
pc
```

The first part of the printout repeats the settings that were run. The second part reports the standardized loadings for the requested axes. h2 reports the total amount variation of each variable that is explained across all of the specified axes. u2 reports the amount of variation of each variable that is unexplained.

The final part of this printout lists the SS loadings (the eigenvalues) for each axis and the total proportion of the variation in the dataset explained by that axis.

```
pc$scores
```

This produces the case scores for each of the requested axes.

```
biplot(pc)
```

Produces the biplot for the ordination.

```
pc2 <- principal(doubs, nfactors = 3, score = TRUE, rotate = "varimax")

pc2

pc2$scores

biplot(pc2)
```

This code reproduces the analyses, but using a varimax rotation. Note the difference in the loadings.