S&DS 363 Pset 6

Yavuz Ramiz Colak, Liana Wang, Ryo Tamaki, David Lieberman

```
library(vegan)
library(vegan3d)
library(mgcv)
library(MASS)
library(rgl)
```

More about our data: Relative abundance of 14 species was measured on 10 plots. Plots were ordered from pioneer (early stage) to climax (late stage). The final column contains that stage of the forest on a scale from 1 to 10.

```
forest <-read.csv("http://reuningscherer.net/stat660/data/Wisconsin.Forest.csv")
rownames(forest)=forest[,1]
forestenv=matrix(forest[,17],ncol=1)
rownames(forestenv)=forest[,1]
colnames(forestenv)=c("Stage")
forest=forest[,-c(1,17)]
forestenv=data.frame(forestenv)
colnames(forest)
```

```
##  [1] "Common.name"       "Bur.oak"           "Black.oak"
##  [4] "Shagbark.hickory"  "Black.cherry"      "White.oak"
##  [7] "Black.walnut"      "Red.oak"           "Butternut"
## [10] "American.elm"      "Basswood"          "Slippery.elm"
## [13] "Yellowbud.hickory" "Ironwood"          "Sugar.maple"
```

Question 1

Fit Correspondence Analysis to your data.

```
forest_cca <- cca(forest[,2:15])
forest_cca
```

```
## Call: cca(X = forest[, 2:15])
##
##               Inertia Rank
## Total          0.7748
## Unconstrained  0.7748    9
## Inertia is scaled Chi-square
##
## Eigenvalues for unconstrained axes:
##    CA1    CA2    CA3    CA4    CA5    CA6    CA7    CA8    CA9
## 0.4790 0.0961 0.0758 0.0529 0.0271 0.0198 0.0140 0.0060 0.0039
```

```
#summary(forest_cca) <- this gives us a clearer view of the inertia and importance of components
```

We fit our model here, and discuss the model in Question 2. We also plot CCA and DCA in Question 2 and discuss the results. In our model, we see that there are 9 unconstrained axes, and eigenvalues range from 0.4790 to 0.0039. The total inertia of our model is 0.7748. Before performing correspondence analysis, we would hope that total inertia explained by first two or three directions would be relatively high–in our data set, we see that this is indeed the case. The inertia explained by the first two/three directions is realtively high: First two eigenvalues get more than 55% of the overall inertia.

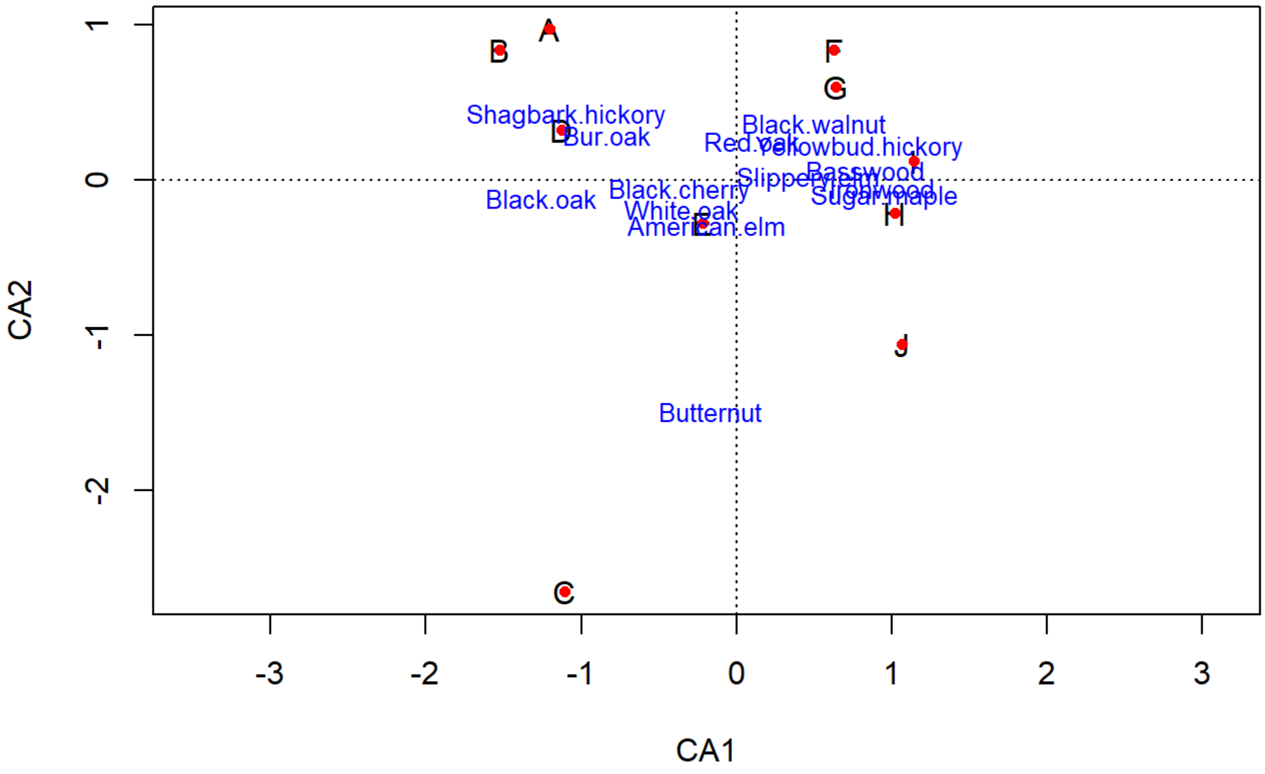
Question 2

Discuss the inertia, make a two dimensional plot of the first two CA directions.

We can see from the summary of eigenvalues that the inertia associated with the first direction CA1 is approximately .479^2 =*0.2291*, and the inertia associated with the second direction CA2 is .09615^2 = *0.0092*. From this, we might think that much more 'information' is contained in CA1 than in CA2 (or any of the subsequent directions), or that the first direction might tell us a better story about our data.This makes sense given that the gradient or direction our data is mainly changing upon is the stage of the plot.

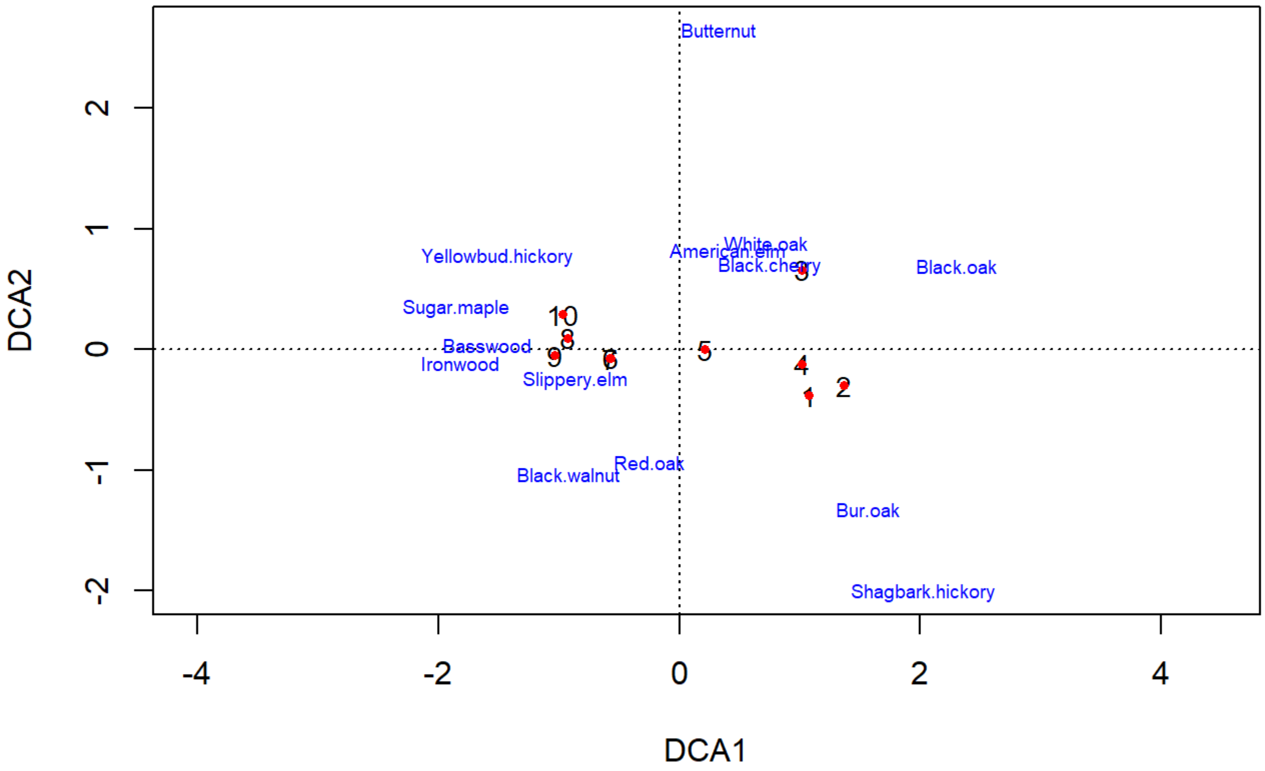
Here we have a plot of CA1 by CA2:

```
plot(forest_cca, type="n")
text(forest_cca, dis="wa") #labeling the points
points(forest_cca, pch=16, col="red", cex=0.76)
text(forest_cca, "species", col="blue", cex=0.8) #labeling the species
```



After seeing the plot of CA1 and CA2, it seems like there might be a little bit of an arch effect. We decide to make a detrended plot since we have ecological data and it is possible that the axes are being interpreted in terms of some environmental gradient.

```
#Plotting detrended correspondence analysis
forest_dca<-decorana(forest[,2:15])
plot(forest_dca, type="n")
text(forest_dca, display=c("sites"), labels=forest[,1], cex=0.86)
points(forest_dca, pch=16, col="red", cex=0.6)
text(forest_dca, "species", col="blue", cex=0.6)
```

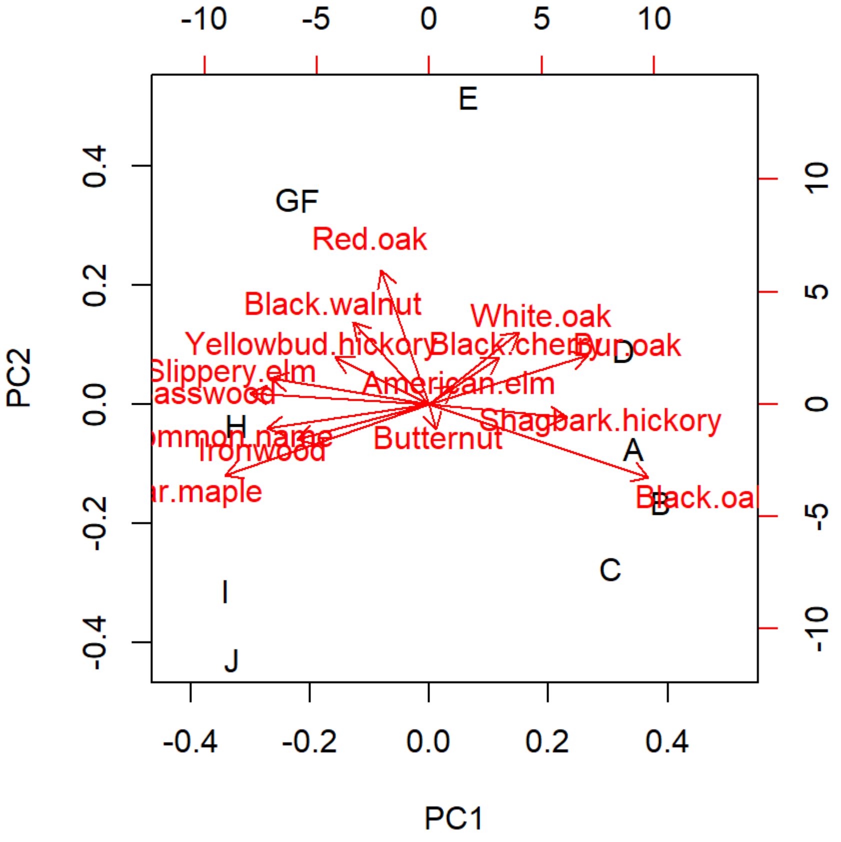


The detrended analysis gives us different directions and thus different eigenvalues / levels of inertia (.4785 for DCA1 and .07698 for DCA2, the directions plotted above). From the DCA plot, we can see that most of the variation amongst the points is happening across the horizontal axis (DCA1) and there is little variation around DCA2.

Question 3

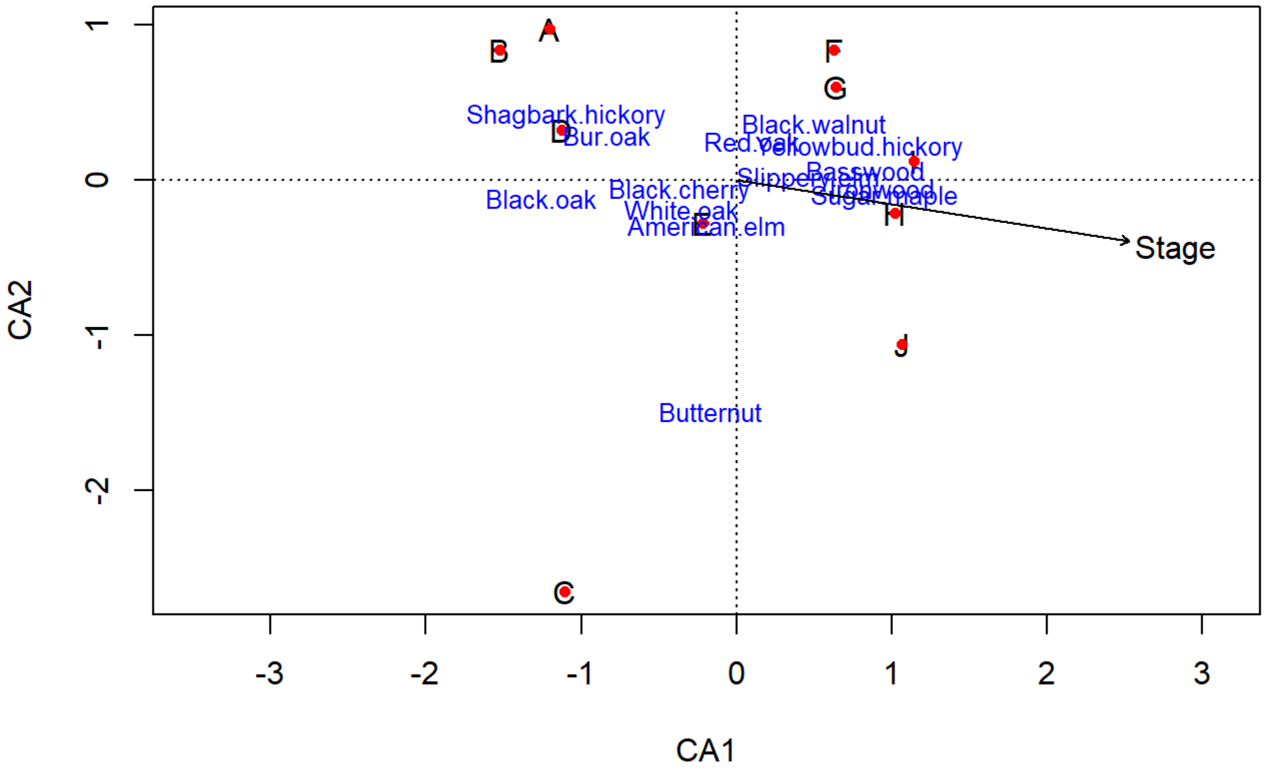
Comment on whether or not there is any evidence of ‘data snaking’ in higher dimensional space.

```
forest_pca <- prcomp(forest)
biplot(forest_pca)
```



From the PCA scoreplot, we can see a “horseshoe” shape amongst the plot points which also does not correspond to the developmental gradient (stage of plot) which our data represents. From this, we might think that there is data snaking in higher dimensional space such that we need some different kind of ordination in order to truly represent the changes over our gradient. (This approach was used in the lecture notes! But as Professor also used CA plot to observe snaking, we included that method as well.)

```
#Fitting Environmental Variable into CA Plot
plot(forest_cca, type="n")
text(forest_cca, dis="wa") #labeling the points
points(forest_cca, pch=16, col="red", cex=0.76)
text(forest_cca, "species", col="blue", cex=0.8) #labeling the species
fit <- envfit(forest_cca, forestenv)
plot(fit, col="black",lwd=3)
```



We also see snaking in our Correspondence Analysis Plot. That is, our data develops in a snake-like pattern with parameter Stage. The path starts from early stages A and B, and goes down to C, and curves back to D and E, finally tilting towards right to old stages F, G, H, I and J.

Question 4

In a few sentences, describe what you conclude from your plot.

Especially from the detrended plot, it would appear that as we move from more late-stage/“climax” plots toward early-stage or “pioneer” plots, the relative abundance of different species increases or decreases and corresponds to the plot(s) in which they are found most. First, for example, we see butternut in the top of the detrended graph (DCA1 x DCA2), and the points with the least distance away from butternut are are 3, 8, and 10. If we go back to the data, we can see that these are the only plots where butternut was counted. Secondly, we can also see (in the detrended graph) that species which are common in the same stages appear close together on the graph: black cherry and white oak are both highly populated in stages 3, 4, and 5 and less populated in stage 10, and they are almost overlapping on the graph.

Question 5

Perform Multidimensional Scaling (metric or non-metric) for 1, 2, and 3 dimensions.

Here, we are performing MDS for 1 to 3 dimensions using Euclidean Distance. We are also performing randomization to check structure. Here, we are making use of the metaMDS function. Given a set of observed similarities/distances between all pairs of 14 items measured in many dimensions, our goal is finding a low dimensional (maybe 2 or 3) representation of the objects that almost match the original similarities/distances.

```
results <- matrix(NA,21,3)
#j is number of dimensions to try
for (j in 1:3){
  for (i in 1:20){
    temp <- forest[shuffle(nrow(forest)),2]
    for (k in 3:15) { temp <- cbind(temp,forest[shuffle(nrow(forest)),k]) }
    #store stress
    results[i,j] <- metaMDS(temp, k=j, distance="euclidean")$stress
  }
  results[21,j] <- metaMDS(forest[,2:15],k=j, distance="euclidean")$stress
}
```

We study our results in 2 or 3 dimensional ordination plots. (Please note that we hid the results in our pdf file to save space, but the process of performing Multidimensional Scaling can be seen in our R file.)

Question 6

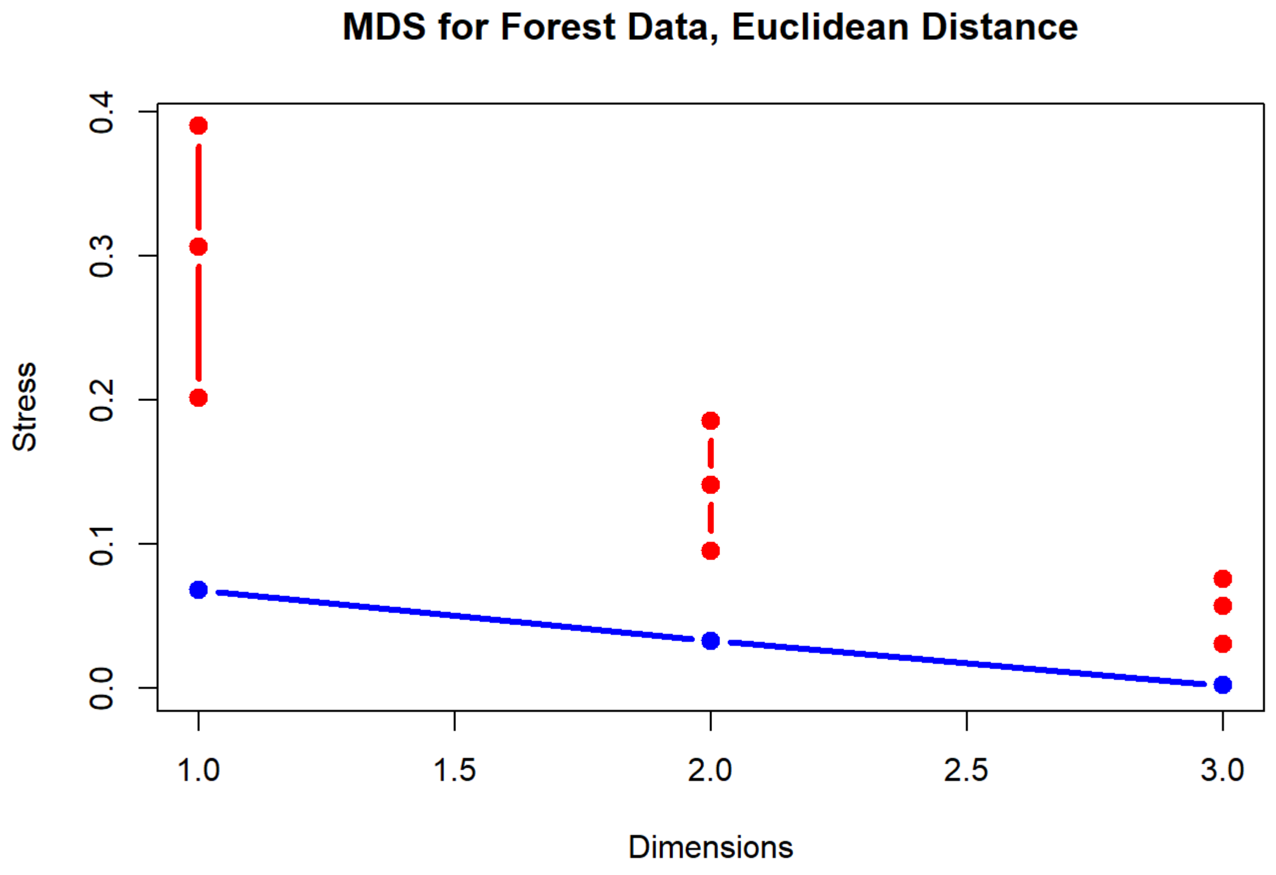
Discuss the stress (or SStress) of each dimensional solution. Make a scree plot if you're able.

3 Dimensions

```
#plot stress results

plot(c(1:3),results[21,],type="b", col="blue", lwd=3,
     ylim=c(0, max(results)), xlab="Dimensions", ylab="Stress", pch=19,
     main="MDS for Forest Data, Euclidean Distance")
mins <- apply(results[1:20,],2,min)
maxs <- apply(results[1:20,],2,max)
meds <- apply(results[1:20,],2,median)

for (i in 1:3){
  points(rep(i,3),c(mins[i], meds[i], maxs[i]),type="b", col="red", lwd=3, pch=19)
}
legend(3.5,(.9*max(results)),c("MDS Solution", "20 Permutations"), lwd=3, col=c("blue","red"))
```



Here we make a scree plot on 1, 2 and 3 dimensions. Here we observe that in 1 and 2 dimensions, we get much lower stress values than what we would achieve in random results. However, in 3 dimensions, our results are not too different from random results. It is also important to note that the stress for 1 dimension is in the typical range of 10% for ordination data. Our stress level for 2 diensions is better around 5%. Our goodness of fit is good per the chart provided in our lecture notes. (While our stress level for 3 dimensions is lowest among the three, it is not different from random results as noted earlier.)

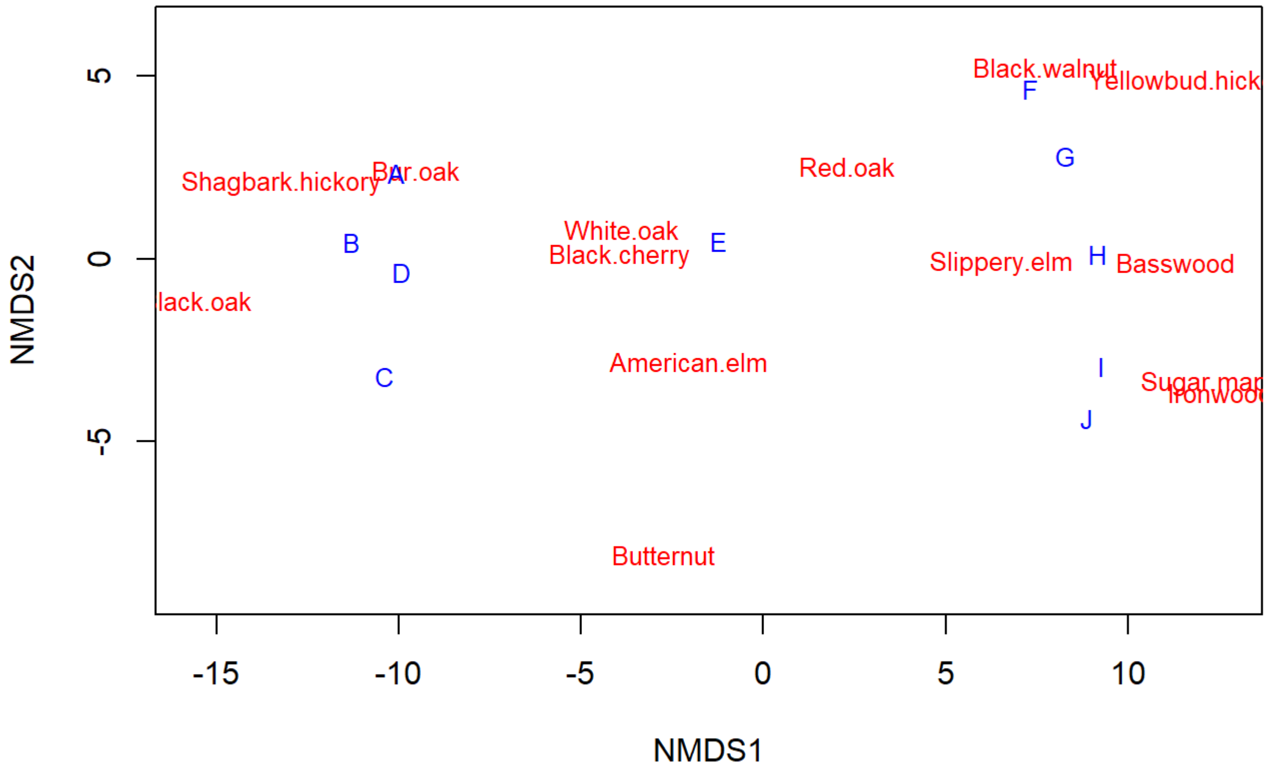
From this, we conclude that two dimensions is the best because we get something that is still significantly different from random results (the permutations indicated by the red dots) – note that three dimensions is not too far off from random results – but the stress level is much lower than if we were to use one dimension and is lower than shuffled stress. While we see only a very slightly tilted elbow around 2 dimensions level (not a sharp elbow as in lecture notes), we note that it is the minimum acceptable stress level for the reasons listed above.

With these, we can note that there is a structure to our data.

Question 7

Make a two dimensional plot of your results.

```
fig<-ordiplot(forest.mds2,type="none",cex=1.1)
text(fig,"species",col="red",cex=0.8)
text(fig,"sites",col="blue",cex=0.8)
```

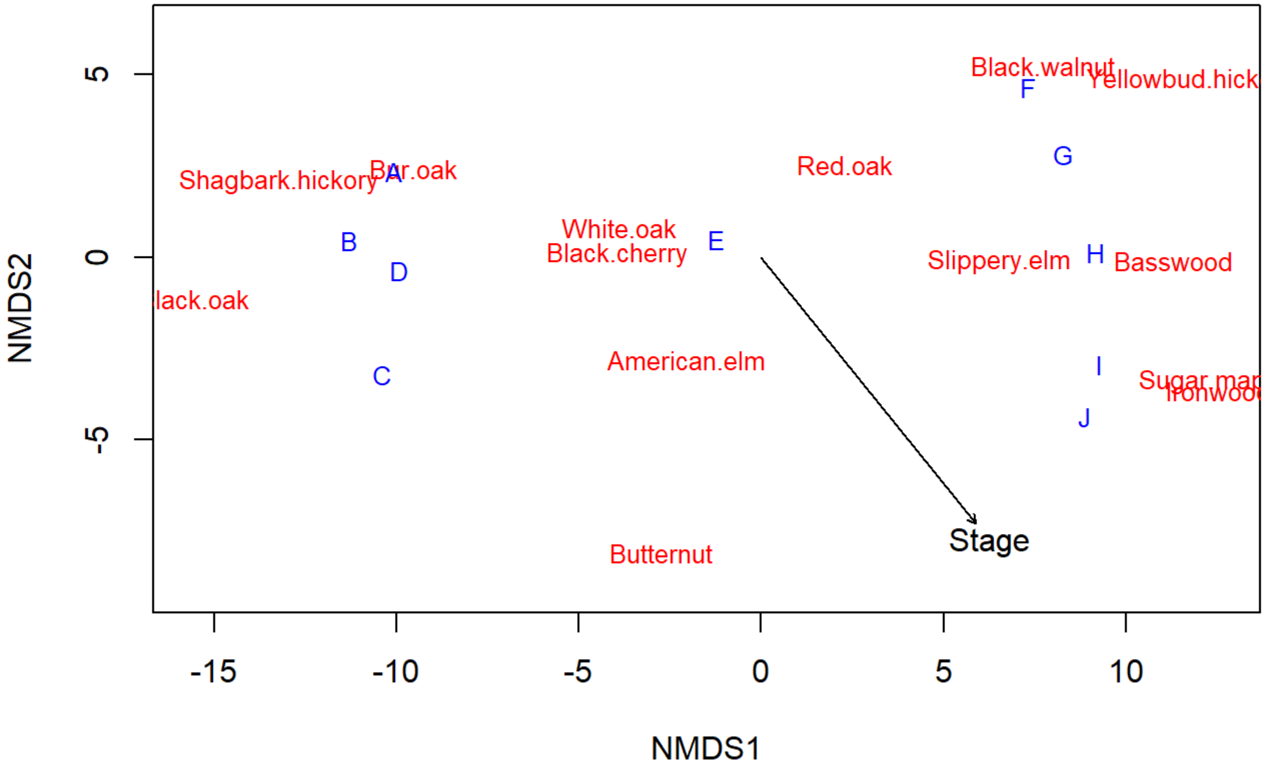


In this plot, it appears that we can group tree species into three broad categories, with some species in between: plots A, B C, and D, are a cluster primarily associated with shagbark hickory, bur oak, and some black oak. Meanwhile, on the other end we see plots F, G, H, I, and J associated with higher counts of black walnut, slippery elm, yellowbud hickory, basswood, sugar maple, and ironwood. Since plot E contains some of most species (unsurprising, given that it is a mid-stage plot), it is located between these two larger clusters. We also see species like white oak, red oak, black cherry, American elm, and butternut, which seem to have relatively uniform distribution across all plots and thus are located in the middle of the graph.

Question 8

If possible, overlay some other variables to interpret your ordination axes.

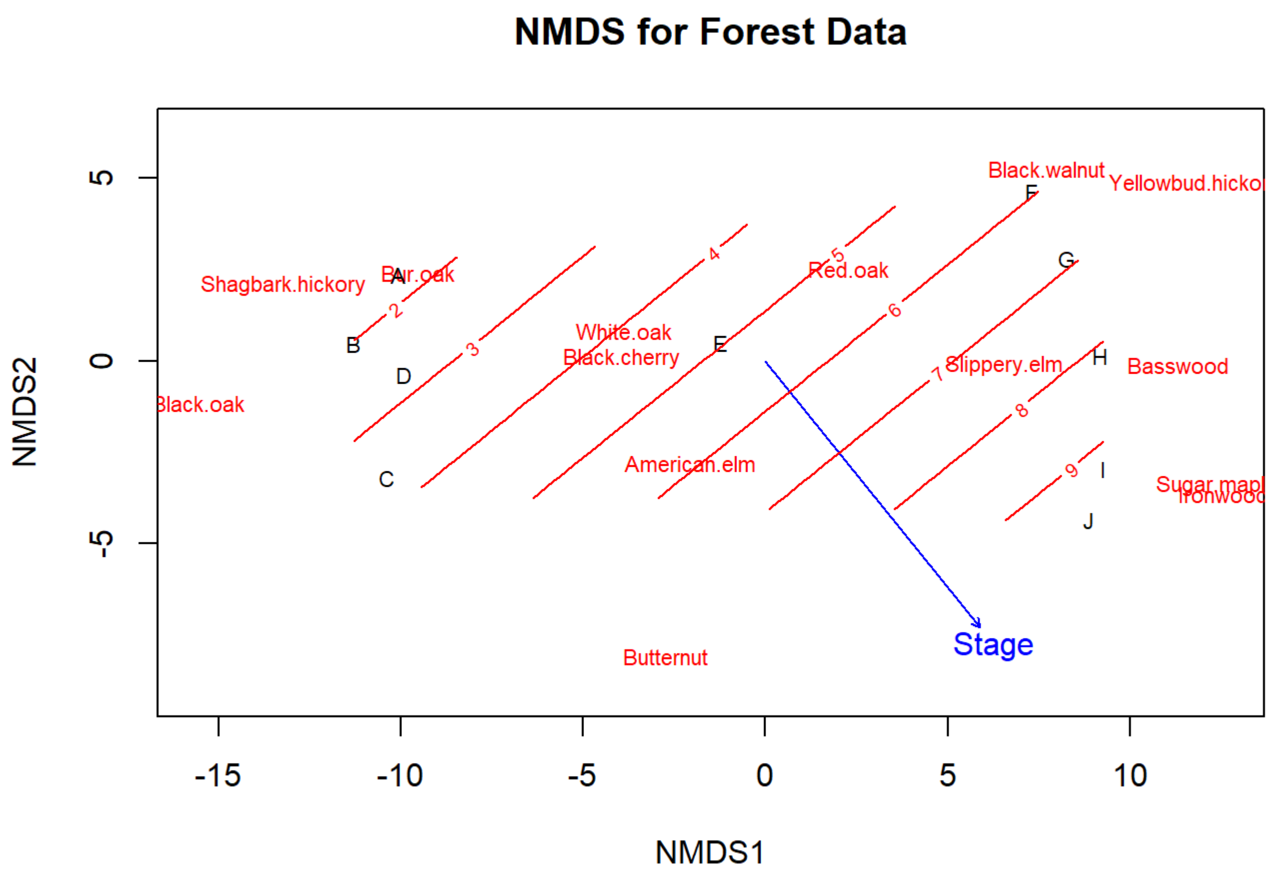
```
fig<-ordiplot(forest.mds2,type="none",cex=0.8)
text(fig,"species",col="red",cex=0.8)
text(fig,"sites",col="blue",cex=0.8)
fit <- envfit(forest.mds2, forestenv,permutations=1000)
plot(fit, col="black",lwd=3)
```



We see the development stage overlaid on this plot, which helps us explain the ordination axes, especially since we know that the plots are ordered from A-J from early to late stage. Knowing the general direction of stage as going down and to the right shows us that species like black oak, shagbark hickory, and bar oak peak in early-stage plots. White oak, red oak, and black cherry peak in early-middle stage plots, while American elm and butternut especially fall a little off the stage direction because their relatively uniform distribution across all plots means they that the ordination in the stage direction is not highly related to the level of counts we get for them. Then, we see species like sugar maple and ironwood which occur exclusively in middle-late and late-stage plots oriented in the rightmost region of the graph. Overlaying stage allows us to see that the axes explain how common species are according to whether they are more or less likely to be found in a certain stage environment.

To further develop this relationship, we can also use regression splines for Stage to see if it actually seems to be ‘linearly’ related across the MDS surface. We can also create a 3-D wire plot of Spline across the MDS surface. We start with using the Regression Spline:

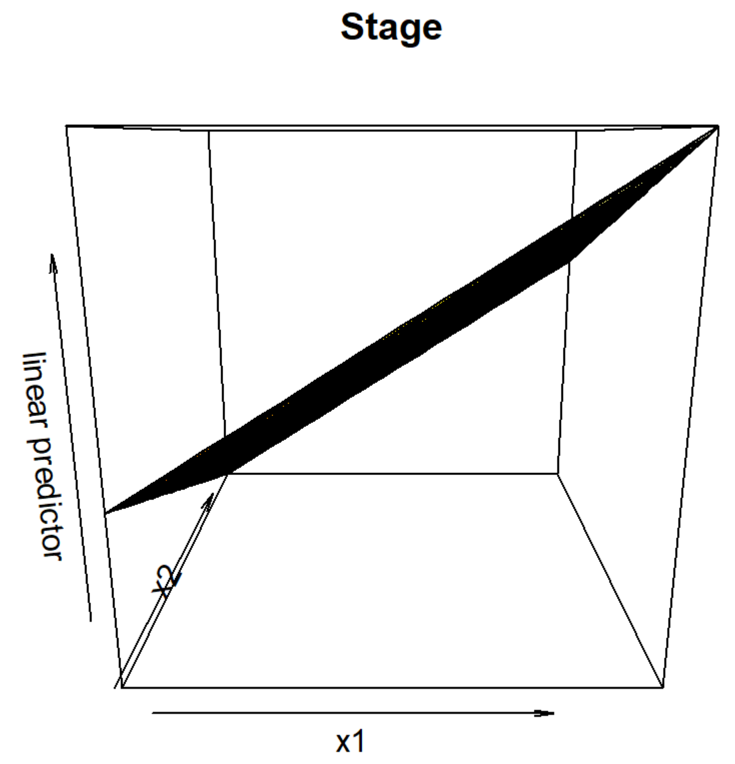
```
fig<-ordiplot(forest.mds2,type="none",cex=1.1,main="NMDS for Forest Data")
text(fig,"species",col="red",cex=0.7)
text(fig,"sites",col="black",cex=0.7)
plot(fit)
tmp1 <- with(forestenv, ordisurf(forest.mds2, Stage, add = TRUE))
```



Overlaying the splines makes it a little clearer which species are most likely to be found in which stage by segmenting out the graph. Here, butternut falls below the splines as mentioned earlier because its distribution seems relatively orthogonal to stage. Since there are so few counts of yellowbud hickory overall, and the variation could be random, it also falls slightly off the splines. However, we can see that shagbark hickory and black oak are definitely more common in stage 1 and 2 plots, while sugar maple and ironwood are most common in stage 9 and 10 plots.

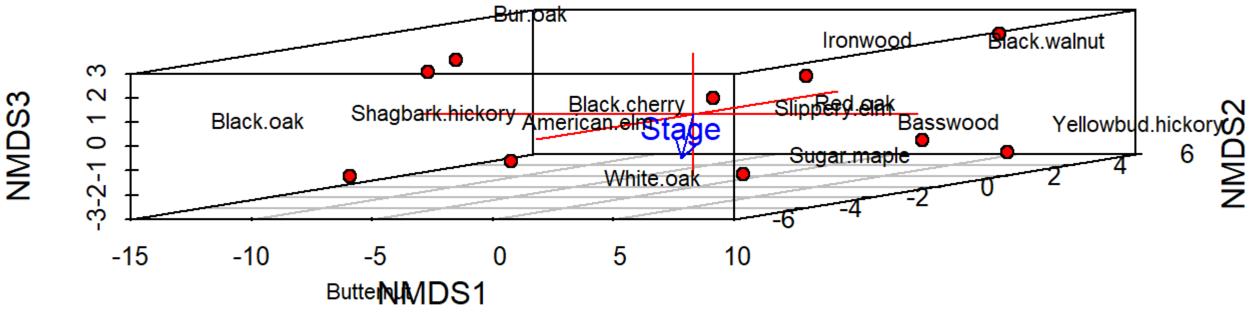
We also create the 3-D Wire Plot to better see the linear relation:

```
vis.gam(tmp1,main="Stage")
```



Our earlier stress test results suggested that 2-D results were better. 3-D stress results were no better than the random stress results. But, for a sanity check, we are also showing the 3-D result here below:

```
fit3 <- envfit(forest.mds3, forestenv, choices = c(1:3),permutations=1000)
pl <- ordiplot3d(forest.mds3, envfit = fit3, pch=19)
points(pl, "points", pch=16, col="red", cex = 0.7)
text(pl, "arrows", col="blue", pos=3)
sp <- scores(forest.mds3, choices=1:3, display="species", scaling="symmetric")
text(pl$xyz.convert(sp), rownames(sp), cex=0.7, xpd=TRUE)
```

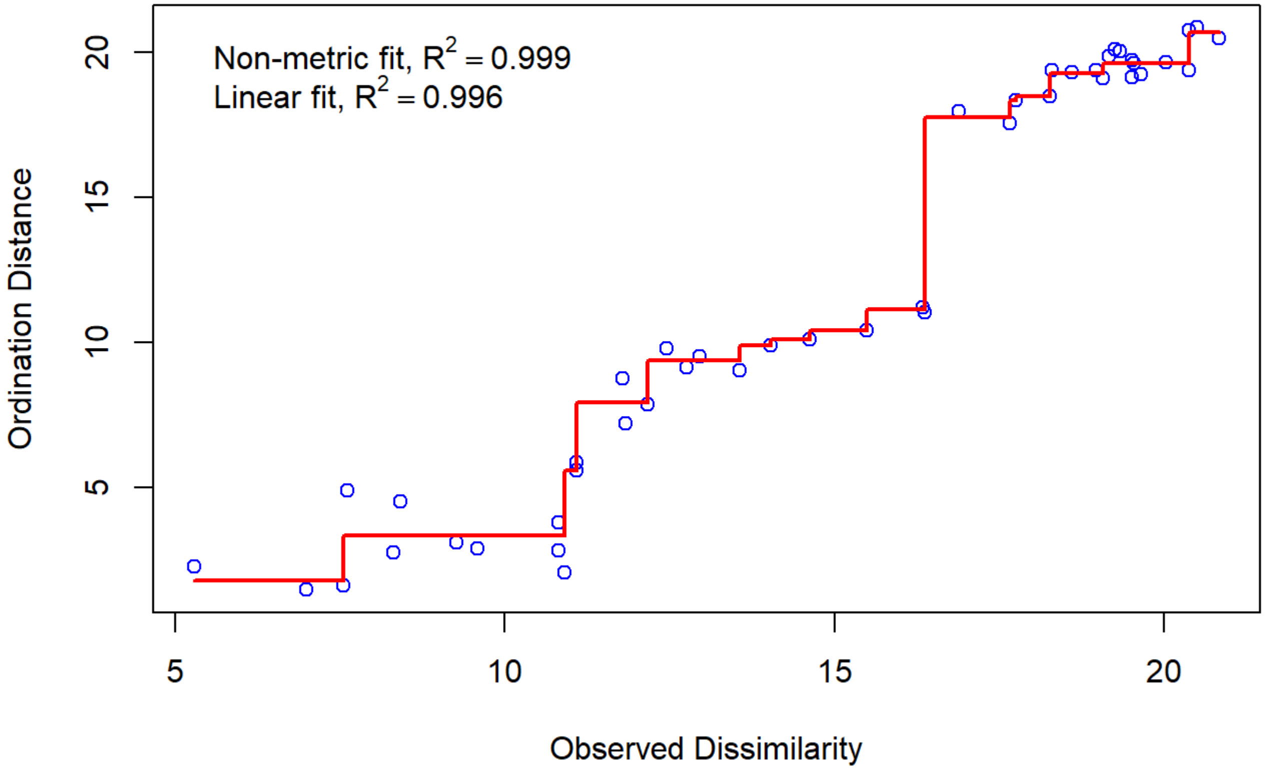


We

then show the stress plots for 2-D and 3-D separately below:

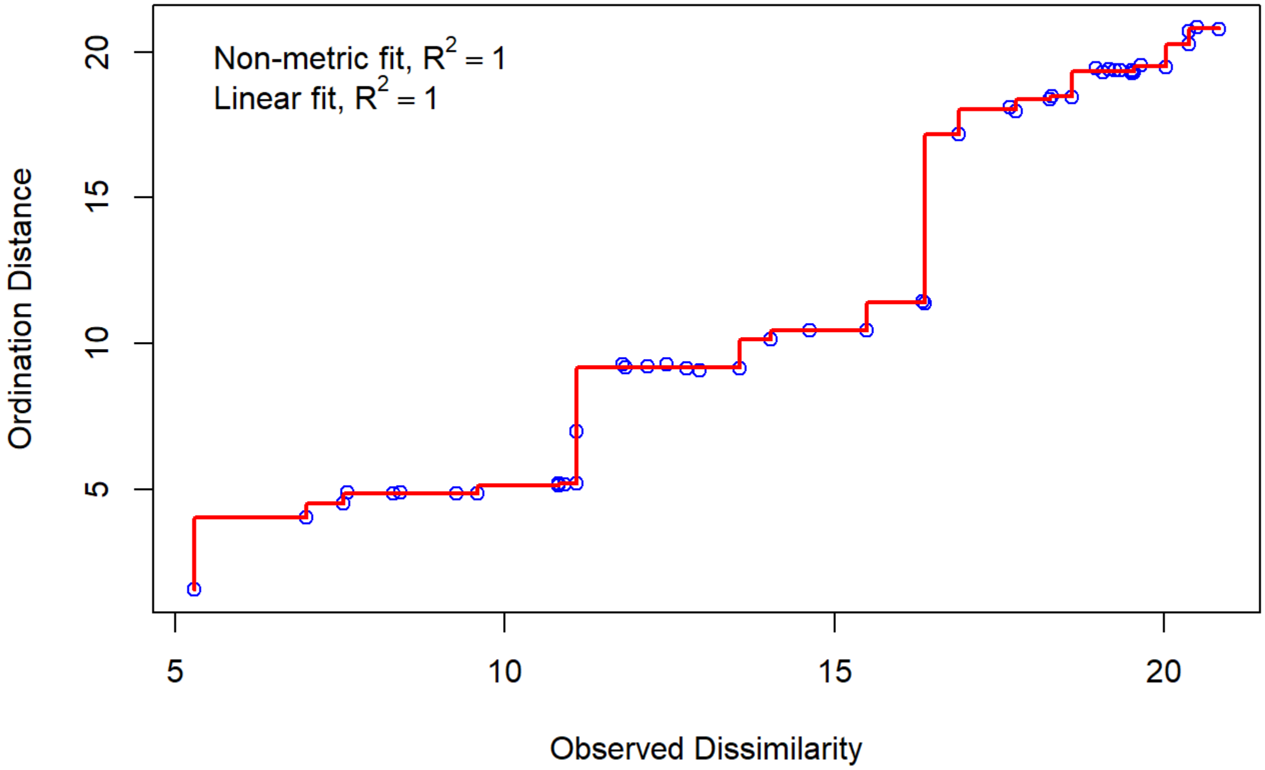
2-D

```
stressplot(forest.mds2)
```



3-D

```
stressplot(forest.mds3)
```



Based on the stress plots, we see that the 2-D result is more adequate. Overall, MDS is most successful in this case.

Bonus

Try canonical correspondence analysis, or calculate p-values for the overlaid additional variables.

```
fit <- envfit(forest.mds2, forestenv,permutations=1000)
fit
```

```
##
## ***VECTORS
##
##          NMDS1    NMDS2     r2   Pr(>r)
## Stage  0.62718 -0.77888 0.9671 0.000999 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 1000
```

The overlaid additional variable is Stage. In our 2-D results, which was chosen base don our previous stress analysis, we see that the p-value of Stage is 0.000999. The Stage variable is significant at a significance level of 0.001.