Home work #6 –Multivariate

This homework will be similar to the last homework in that you do not need to hand in methods and you do not need to write prose for the results (except where noted – there will be some "interpretation" questions that require a 1-2 sentence answer). Be sure to turn in and spend some time on your CODE section as this is now a more important part of your work.

This homework is going to be very similar to work done in class. As a consequence, I will rarely give you the necessary commands. You will need to look them up in your class notes.

#1 – Multivariate – GLM analogues & hypothesis testing

The first dataset we will use is a dataset found in the MASS library called crabs – i.e. use data(crabs). Look at the crabs dataset. It has 50 individuals of both sexes of two species (O=Orange, B=Blue) of crab. Five morphometric measurements were taken on adults:

FL=frontal lobe size (mm)
RW=rear width (mm)
CL=carapace length (mm)
CW=carapace width (mm)
BD=body depth (mm)

We will be exploring whether the two species differ morphological (and ought to be considered separate species). To simplify things we will examine only females. Type:

```
crabs2=crabs[crabs$sex=="F",-(2:3)]
# only females, drop the sex & index columns
```

Use this as your dataset from here on out when I ask for the crab dataset.

GLM Analogues

The simplest and most direct test is a Hotelling T test for whether the two species differ morphologically on all 4 traits at once (y has four variables). You could look up the Hotelling T test, but instead just use the manova command shown in class (manova on a discrete variable with two levels is approximately equivalent to MANOVA). **1) Calculate the means and standard deviations for all ~~four~~ five traits for each species. Does it look like there are significant differences? 2) Report the approximate F and the p for this test and whether this means the species are morphologically distinct (even if overlapping).**

Now load the two datasets varespec and varechem from the library vegan (i.e. data(varespec) and …). Look at the datasets. The varespec data contains species abundances from a community. varechem contains various environmental factors including. Perform an RDA using only the three non-chemistry columns (pH, humdepth, varesoil) to see how the abundances depend on the soil type. Hint: Use the formula version of the rda command (library vegan). Do a triplot. **3) Interpret the loadings and triplot. Write a short paragraph on conclusions you can draw. Hint: there is a common problem that shows up in this plot. Be sure to mention it.**

#2 – Multivariate – Principle components

PCA on environment

Run a PCA on the environmental (env.csv from my website) data. **4) How many components do you need to cover 75% of the variance and how many components would you include in further analyses (and why?) NB no one right answer so explain your choice.** Look at the loadings. **5) What is the loading on Component 1 for clP and give a verbal interpretation of the first 5 components.**

Now try a factor analysis. Omit the landcover variables (use d[,1:10] to get the first 10 variables). Run a factor analysis for 2 factors with no rotation (rot="none"), promax rotation and varimax rotation. **6) Which of these is easiest to interpret? What interpretation would you give it?**

Now load the dickcissel.csv which has abundance and environment data (note the environment data is at a different set of stations so you will get a slightly different answer than the previous paragraph). Specifically, you will need to run a new PCA on the environmental dataset just on this dataset to get the rows to match to the ros of dickcseel abundance. Create a presence/absence variable (i.e. abundance>0). Run a logistic regression of presence/absence vs. the first principle component of the climate (not the landcover) data (i.e. solve the collinearity of all the climate data by using principle components). **5)What is the regression equation and what is the p-value?**

~~Now try a factor analysis. Omit the landcover variables (use d[,1:10] to get the first 10 variables). Run a factor analysis for 2 factors with no rotation (rot="none"), promax rotation and varimax rotation. **6) Which of these is easiest to interpret? What interpretation would you give it?**~~

PCA on Crabs

Now run a PCA on the crabs. Hint #1: you may want to use crabs2[,2:6] to remove the species column if you get an error. Hint #2: do you need any transformation or scaling? Hint #3: covariance or correlation? Run a summary, a scree plot, and look at the loadings. Perform biplot with crabs2$sp as a label and interpret it. **7) What percentage of variance is covered by the first two principal components, give verbal interpretations of the first four components and describe verbally how (and if) the first two components succeed in dividing the species.**

Let's return to the full crab dataset with males & females (crabs, not crabs2). Run a PCA on it (remember you will need to use crabs[,4:8] to get just the numeric part and check out the hints in the last paragraph). Use the following command to get appropriate labels:

```
lab=factor(c("B","b","O","o")[rep(1:4,each=50)])
```

8) Look at lab and describe in words what these labels mean. Quickly run a PCA and a biplot (using the option xlab=lab on the biplot command) of the full crab data. **9) Describe in words how the first two principal components separate species AND gender.**

#3 – Ordination

Return to the varespec dataset. Many ordination techniques must have more rows than observations. So make a subset `varab<-varespec[,1:20]`. Run a principle component analysis on varab. **10) Write a paragraph interpreting this ordination. Hint: What species structure the first axis? What species structure the second axis? What sites are extreme?**

Now perform two other ordination techniques (NMDS, PCoA with distance based on Morisita distances). **11) Compare and contrast the results form these two techniques with the PCA.**

#4 – Multivariate - Clustering

Clustering

We will continue with the full crab dataset and see if clustering can break these apart. Run a divisive cluster technique (diana) on the full crabs data (with males & females and again using just the numeric parts). Plot the result. **12) Do four distinct clusters appear?** Let's see how well these got broken up. Try the following commands:

```
d=data.frame(assigned=cutree(cl,k=4),class=lab)
table(d)
plot(cl,lab=lab,w=2,cex=0.5)
```

13) Describe in words what these three commands did (i.e. write a 1-2 sentence methods statement) and then describe in words your interpretation of the results.