Homework 6- Questions

Anchalya Bala

1. **Work through performing a PCA with the zooplankton data and then answer the following questions:**
2. **Why do we use a Hellinger transformation with the zooplankton data?**

This is because a Hellinger Transformation is best used for species abundance data such as the zooplankton data. This transformation gives low weights to variables with low counts and lots of zeros. The transformation divides each value in a data matrix by its row sum, and taking the square root of the quotient.

1. **How much variation was explained by the first two axes of your PCA?**

We only look at the first two axes because they’re giving us the most important data.

0.4306 0.1615

1. **Interpret your PCA biplot. Does it seem like certain zooplankton species are associated with acidic or neutral conditions? Do the lake types seem to separate out along the first (x) or second (y) axes?**

My interpretation of the PCA Biplot is that certain zooplankton species are associated with acidic or neutral conditions. The lake types are separated along the y axes.

1. **Based on your interpretation from c, does it seem like zooplankton communities differ according to lake type?**

Yes, zooplankton communities differ according to lake types.

1. **Based on the score distance and orthogonal distance calculated with the pcaDiagplot function, are there any lakes that are outliers?**

There are a couple outliers but for the most part everything is fine.

1. **Provide an example from your own work, or work that you are familiar with, that could make use of principle components analysis. Why would the dataset be suitable for PCA?**

PCA can be very handy when we look at cells. Cells can seem to be the same but after genetic analysis, we can conclude that they are different by understanding what each cell type is doing. Using PCA, we can differentiate these cells. This tool will help us bring out strong pattern between the cell types or any complex biological datasets. The dataset that has 3 or higher dimensions or categories are best for PCA such as the cells examples if viewing multiple types of cells.

1. **Next, work through the code for running an RDA, and then answer the following questions:**
2. **Why do we standardize predictors before running an RDA?**

Standardizing predictors is important for an RDA, as different predictors are measured on different scales. I believe we standardize predictors to remove non-significant explanatory variables since some can be large and some can be small. To counter this, we standardize the predictors by making them all have a mean of 0 and a variance of 1.

1. **Which predictor variables were significant based on the step-wise variable selection procedure?**

Fish + pH + DOC,

1. **Based on the permutation test (permutest function), was the final RDA significant? What was the significance value from that test?**

The significance is less than 0.05, we conclude that our RDA does explain variation in our zooplankton communities.

1. **Based on the VIF, were any of the variables collinear?**

I tested for collinearity among predictors in my model using the VIF. There are no collinearity because all of the SIGNIFICANT variables have values that are less than 10.

1. **How much variation in zooplankton communities did the first and second axes of the RDA explain?**

The constrained ( what was included in the model): 0.1473

The unconstrained (the unexplained variation): 0.3737

The proportion explained (by all the axes, RDA1): 0.1528

The proportion explained (RDA2): 0.08661

1. **Now examine the RDA biplot that shows the site scores. Were certain types of lakes associated with the three predictor variables? For example, did lakes with a high pH consistently plot out along the pH arrow? What about DOC or Fish?**

For the most part this RDA biplot shows lakes associated with the right three predictor variables. Of course you can see some lakes with, say, high PH not in the pH arrow and same thing can be said about recovered. Neutral lakes seem to follow the neutral arrow.

1. **Based on your interpretation from f, does it appear that lakes in the three different categories (acidic, neutral, recovered) have different physical properties?**

I think because of the weird discrepancy we see where some lakes don’t follow its appropriate arrows, we can conclude that some lakes have different properties. Maybe there are other factors at play here.

1. **Now examine the RDA biplot with species scores. Does it look like particular species are associated with some of the predictor variables? For example, are some species found along the pH arrow, indicating they like higher pH levels**?

There are a few that run along the lines. The species has a higher pH is it runs close to or overlaps the red line.

1. **Run the cluster analysis in the R Markdown file to see if lakes can be divided into groups with similar chemical/physical characteristics. After you have worked through the analysis, answer the following questions.**
2. **Based on the “elbow method,” how many groups do you think we should use for our cluster analysis?**

The more clusters, the deeper the bend. I would say we should use 3 groups. At three groups is when you really start to see the “elbow’ or bend.

1. **When you look at the cluster means, do you notice differences among the groups from the k-means analysis? For example, do some groups have lakes with larger surface areas, higher maximum depths, elevation, etcetera?**

When I look at the cluster means, I do notice differences among the groups from the k-means analysis. Some groups have lakes with larger surface areas, higher maximum depths, and elevation.

1. **Examine the PCA plot with the clusters circled. What is your interpretation of the plot? Does it look like the clusters are clearly separate? Do the lake status categories (neutral, acidic, recovered) seem to match the clusters that were chosen by our k-means analysis?**

The clusters aren’t clearly separated as there seems to be some discrepancy however the lake status categories do seem to match the clusters that were chosen by our k-means analysis for the most part. Therefore, for the most part they do match but some are out of order and found in the wrong category.

1. **Are there some lakes that seem to fall in the “wrong” cluster according to our lake status categories? Which lakes seem to be interlopers in the wrong cluster?**

The acidic lakes are kilarney, Norway, Davud, Muriel, Turbid, Freeland and Terry, however they were grouped with mostly recovered lakes. Helen, Gem, Howry and Ishn are neutreal lakes. However, they were also grouped with recovered lakes. From looking at the plot, Froos and Evangeline which are recovered lakes were grouped with neutral lakes.