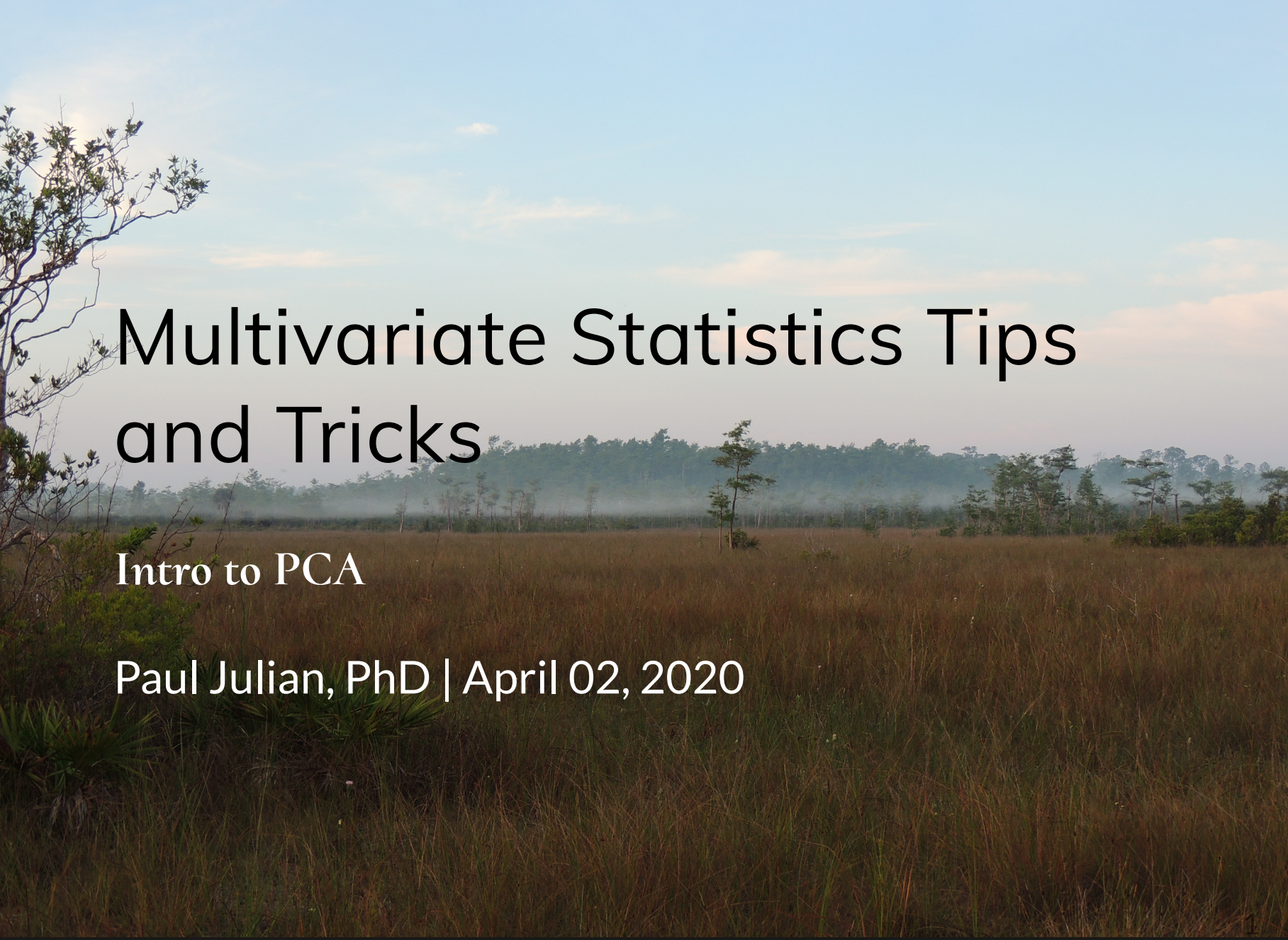


Multivariate Statistics Tips and Tricks

Intro to PCA

Paul Julian, PhD | April 02, 2020

- Thanks to Dr Zarah Pattison and the Modelling, Evidence and Policy Research group for the invitation.
- Today's discussion will be loosely based on a recent blog post on Principal Component Analysis.

[1] https://swampthingecology.org/blog/pca-basics-in-rstats/

About Me

- Not a statistician (just play one on TV)
- Wetland Biogeochemist
- PhD in Soil and Water Science (University of Florida)
- Likes long walks on the beach...or a wetland...

Find me at...

@swampthingpaul
@swampthingpaul
swampthingecology.org
pauljulianphd@gmail.com

Ordination Analysis

A family of statistical analyses used to order multivariate data.

Some common analyses include:

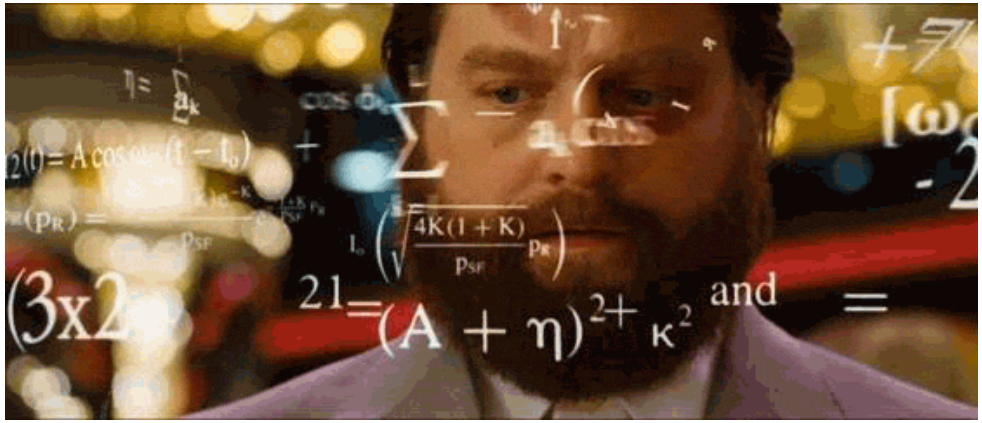
- **Principal Component Analysis (PCA)**
- Correspondance analysis (CA) and its derivatives
    - detrended CA
    - canonical CA
- Redundancy Analysis (RDA)
- Non-Metric Multidimensional Scaling (NMDS)
- Bray-Curtis Ordination

Principal Component Analysis

"A rose by any other name ..."

I have heard PCA called many names:

- ***unsupervised feature extraction***
- ***dimensionality reduction***
- statistical hand waving
- mass plotting
- magic



- Rooted in linear algebra, its the simplest of the true eigenvector-based multivariate analyses.
- Creates weighted linear combination of the original varables to capture as much variance in the dataset whilest eliminating correlations/redundancies.
- Reveals the internal structure of the data in a way that best explains the variance in the data.

Principal Component Analysis

Typically when talking about PCA you hear terms like *loading*, *eigenvectors* and *eigenvalues*.

- **Eigenvectors** are unit-scaled loadings. Mathematically, they are the column sum of squared loadings for a factor. It conceptually represents the amount of variance accounted for by a given factor.

- **Eigenvalues** is the measure of variation in the total sample accounted for by each factor. Computationally, a factor's eigenvalues are determined as the sum of its squared factor loadings for all the variables. The ratio of eigenvalues is the ratio of explanatory importance of the factors with respect to the variables (remember this for later).

- **Factor Loadings** is the correlation between the original variables and the factors. Analogous to Pearson's r, the squared factor loadings is the percent of variance in that variable explained by the factor

Principal Component Analysis

Imagine a multivariate dataset...lets say lake water quality data

```
##   Station.ID              LAKE   Date.EST Alk Chla    SRP    TP   TN  DIN
## 1        A03 East Tohopekaliga 2005-05-17  17  4.0 0.0015 0.024 0.71 0.04
## 2        A03 East Tohopekaliga 2005-06-21  22  4.7 0.0015 0.024 0.68 0.03
## 3        A03 East Tohopekaliga 2005-07-19  16  5.1 0.0015 0.020 0.63 0.02
## 4        A03 East Tohopekaliga 2005-08-16  17  3.0 0.0015 0.021 0.55 0.03
```

- or any other dataset with several different variables

PCA is a way to reduce the dimensionality of the data and determine what *statistically* matters.

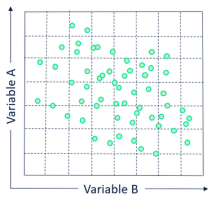
Its beyond a data winnowing technique it also shows similarity (or difference) between groups and relationships between variables.

Principal Component Analysis

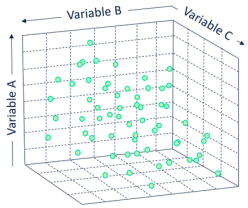
Disadvantages

- PCA is data hungry

- Fall victim to the curse of dimensionality.
  - As dimensionality increases, effectiveness of the data decreases.
  - As dimensions are added to a data set, the distance between points increases in the multivariate space.



2-Dimensional Problem Space



3-Dimensional Problem Space

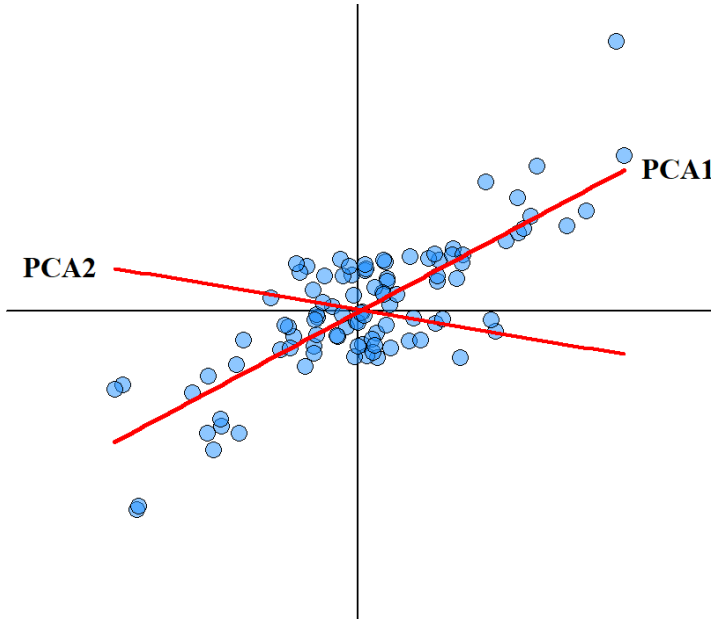
PCA Assumptions

- **Multiple Variables:** An obvious assumption, a multivarate analysis needs multiple variables. Meant for continuous variables, but ordinal varables are frequently used.
- **Sample Adequacy:** Size matters!! A general rule of thumbs has been a minimum of 150 cases (ie rows), or 5 to 10 cases per variable.
- **Linearity:** It is assumed that the relationships between variables are linearly related. The basis of this assumption is rooted in the fact that PCA is based on Pearson correlation coefficients and therefore the assumptions of Pearson's correlation also hold true. Generally, this assumption is somewhat relaxed.
- **Outliers:** Outliers can have a disproportionate influence on the resulting component computation. Since principal components are estimated by essentially re-scaling the data retaining the variance outlier could skew the estimate of each component within a PCA.

PCA Assumptions

- One more thing about outliers.
- Another way to visualize how PCA is performed is that it uses rotation of the original axes to derive a new axes, which maximizes the variance in the data set.

In 2D this looks like this:



R you ready?

PCA Analysis can be done through a variety of `R` Packages. Each have there own nuisances...

- `prcomp()` and `princomp()` are from the bases `stats` package. The quickest, easiest and most stable method
- `PCA()` in the `FactoMineR` package.
- `dubi.pca()` in the `ade4` package.
- `acp()` in the `amap` package.
- `rda()` in the `vegan` package. More on this later.

Personally, I only have experience working with `prcomp`, `princomp` and `rda` in the following examples we will be using `rda()` but this can be adapted to the other analyses.

- `rda()` performs redundancy analysis. Normally RDA is used for *"constrained ordination"* but without predictors, functionally RDA == PCA

- R packages used today include
  - AnalystHelper
  - reshape
  - vegan
  - REdaS

```
library(AnalystHelper)
library(reshape)
library(vegan)
library(REdaS)
```

- For demonstration purposes we will use the dataset I introduced earlier .../data/lake_data.csv.
  - dat<-read.csv(".../data/lake_data.csv")
- If you are playing the home game you can follow along with .../PCA_rawcode.R.

[1] https://github.com/SwampThingPaul/AnalystHelper

[2] https://github.com/SwampThingPaul/PCA_Workshop

Currently the data is in rows, we need it in columns (data massaging).

```
# Cross tabulate the data based on parameter name
dat.xtab <- cast(dat,Station.ID+LAKE+Date.EST~param,value="HalfMDL",mean)

# Cleaning up/calculating parameters
dat.xtab$TN <- with(dat.xtab,TN_Combine(NOx,TKN,TN))
dat.xtab$DIN <- with(dat.xtab, NOx+NH4)

# More cleaning of the dataset
vars <- c("Alk","Cl","Chla","DO","pH","SRP","TP","TN","DIN")
dat.xtab <- dat.xtab[,c("Station.ID","LAKE","Date.EST",vars)]

head(dat.xtab[,c("Station.ID","LAKE",vars)],4L)
```

```
##   Station.ID              LAKE Alk   Cl Chla  DO  pH    SRP    TP   TN  DIN
## 1        A03 East Tohopekaliga  17 19.7  4.0 7.9 6.1 0.0015 0.024 0.71 0.04
## 2        A03 East Tohopekaliga  22 15.4  4.7 6.9 6.4 0.0015 0.024 0.68 0.03
## 3        A03 East Tohopekaliga  16 15.1  5.1 7.1 NaN 0.0015 0.020 0.63 0.02
## 4        A03 East Tohopekaliga  17 14.0  3.0 6.9 6.3 0.0015 0.021 0.55 0.03
```

- NA values are a no go in PCA analyses...some more cleaning.
- How many NAs do we have?

```
#How many rows of data do we have?
nrow(dat.xtab)
```

```
## [1] 725
```

```
#How many rows contain NAs
nrow(na.omit(dat.xtab))
```

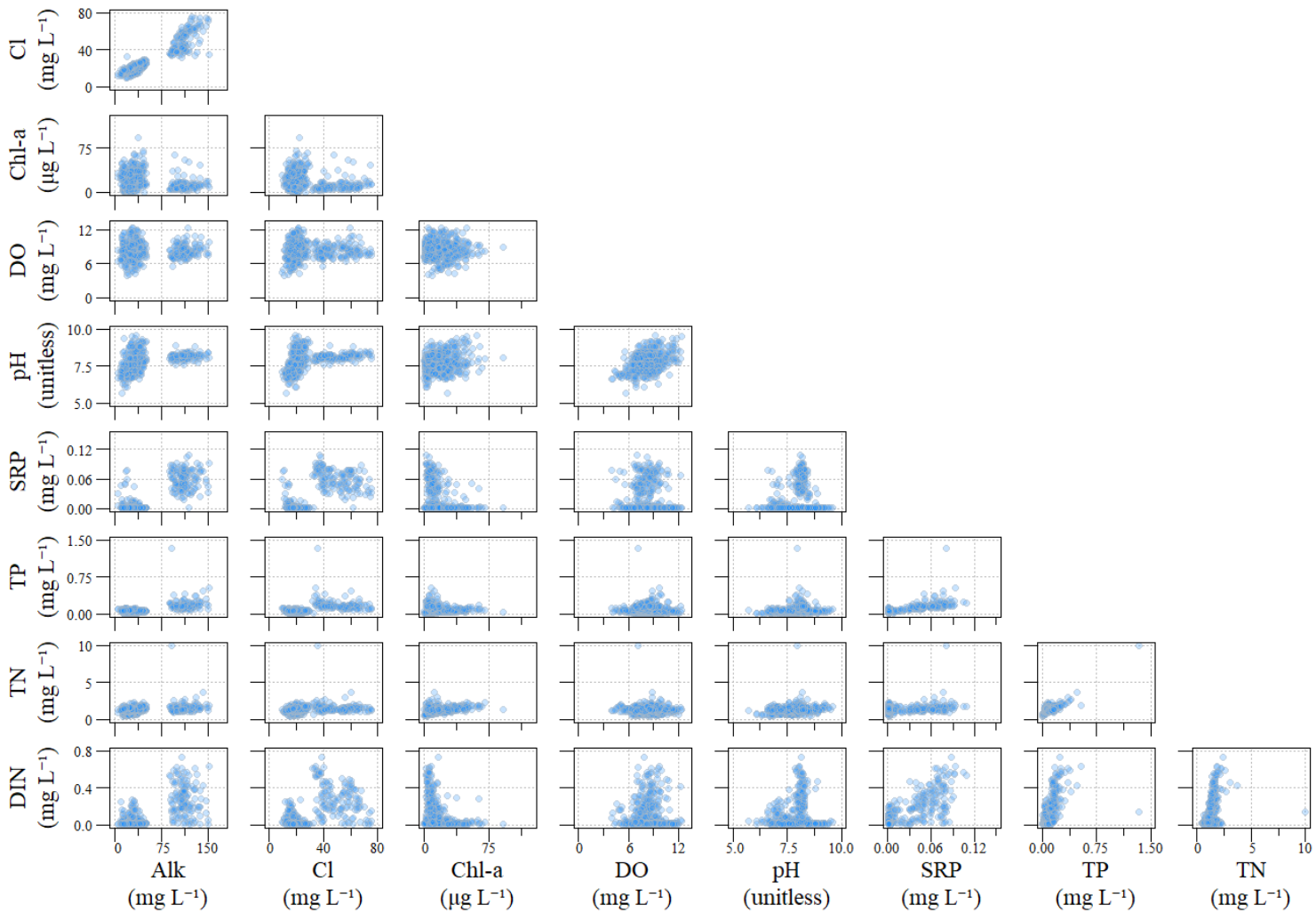
```
## [1] 515
```

That is 210 rows removed due to incomplete data.

- could narrow the parameters you want to look at to avoid excessive data culling.

```
dat.xtab <- na.omit(dat.xtab)
```

Lets take a quick look at the data...



Scatterplot of all data for the example $dat.\ xtab$ dataset.

- Lets check the measure of sampling adequacy.
- Some have suggested to perform a sampling adequacy analysis such as Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy.
    - However, KMO is less a function of sample size adequacy as its a measure of the suitability of the data for factor analysis, which leads to the next point.

```
KMOS(dat.xtab[,vars])
```

```
##
## Kaiser-Meyer-Olkin Statistics
##
## Call: KMOS(x = dat.xtab[, vars])
##
## Measures of Sampling Adequacy (MSA):
##       Alk        Cl      Chla        DO        pH       SRP        TP        TN
## 0.7274872 0.7238120 0.5096832 0.3118529 0.6392602 0.7777460 0.7524428 0.6106997
##       DIN
## 0.7459682
##
## KMO-Criterion: 0.6972786
```

Based on the KMO analysis, the KMO-Criterion of the dataset is 0.7, well above the suggested 0.5 threshold.

- Lets do another check of the data using Bartlett's Test of Sphericity.
    - Bartlett's Test of Sphericity $\neq$ Bartlett's Test for Equality of Variances
- Test of Sphericity tests whether the data comes from a multivariate normal distribution with zero covariances.
    - compares an observed correlation matrix to the identity matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

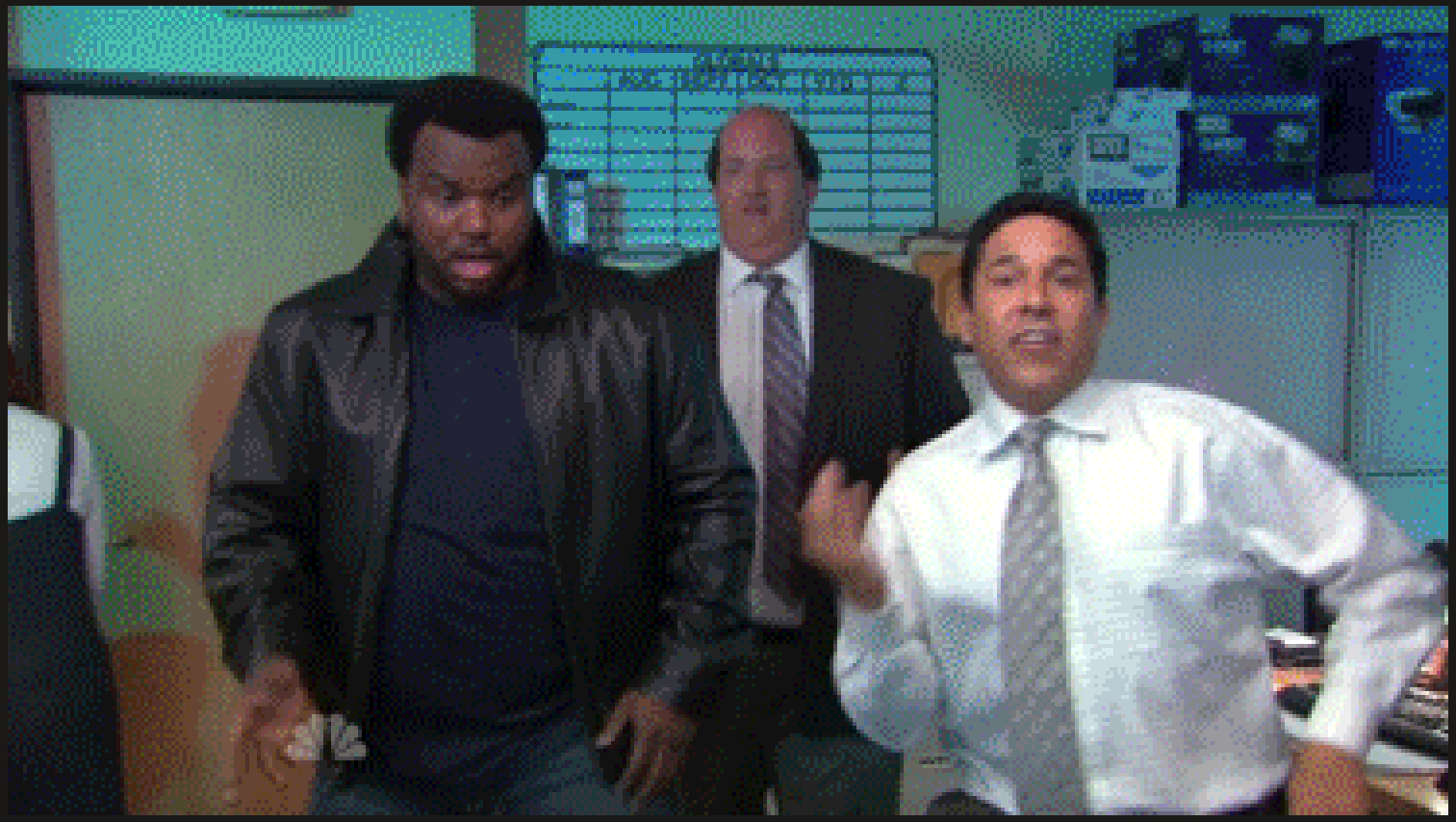
```
# Bartlett's Test Of Sphericity
bart_spher(dat.xtab[,vars])
```

```
##     Bartlett's Test of Sphericity
##
## Call: bart_spher(x = dat.xtab[, vars])
##
##      X2 = 4616.865
##      df = 36
## p-value < 2.22e-16
```

- The data is significantly different from an identity matrix (H_0 : all off-diagonal correlations are zero) and suitable for PCA.

Now that the data is cleaned and checked ...

... lets do some PCA!!



PCA analysis is pretty straight forward.

```
dat.xtab.pca <- rda(dat.xtab[,vars],scale = T)
```

- Using `rda(... scale = T)` without predictors and scaled functionally produces a PCA. Can also compare usin `princomp()`.

Now lets see the importance and variance explained by each component by extracting some important information.

- The quickest way is to use `summary(dat.xtab.pca)$cont`.

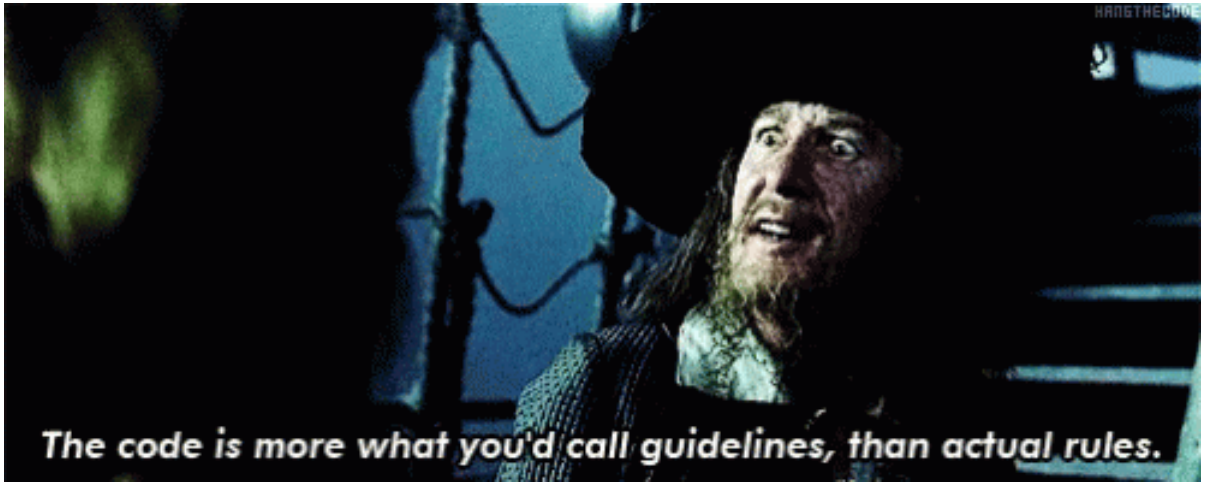
```
## $importance
## Importance of components:
##                            PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Eigenvalue            4.2963 1.8811 1.3700 0.7110 0.34429 0.18096 0.12522
## Proportion Explained  0.4774 0.2090 0.1522 0.0790 0.03825 0.02011 0.01391
## Cumulative Proportion 0.4774 0.6864 0.8386 0.9176 0.95586 0.97597 0.98988
##                            PC8      PC9
## Eigenvalue            0.058289 0.032800
## Proportion Explained  0.006477 0.003644
## Cumulative Proportion 0.996356 1.000000
```

To understand what all this means lets extract the information ourselves.

```
#Extract eigenvalues (see definition above)
eig <- dat.xtab.pca$CA$eig

# Percent of variance explained by each compoinent
variance <- eig*100/sum(eig)

# The cumulative variance of each component (should sum to 1)
cumvar <- cumsum(variance)

# Combine all the data into one data.frame
eig.pca <- data.frame(eig = eig, variance = variance,cumvariance = cumvar)
```

- To double check, compare `summary(dat.xtab.pca)$cont` with `eig.pca` ... they should be the same.
- What does the component eigenvalue and percent variance mean?
- More importantly what does it tell us about our data?

This information helps tell us how much variance is explained by the components. It also helps identify which components should be used moving forward.

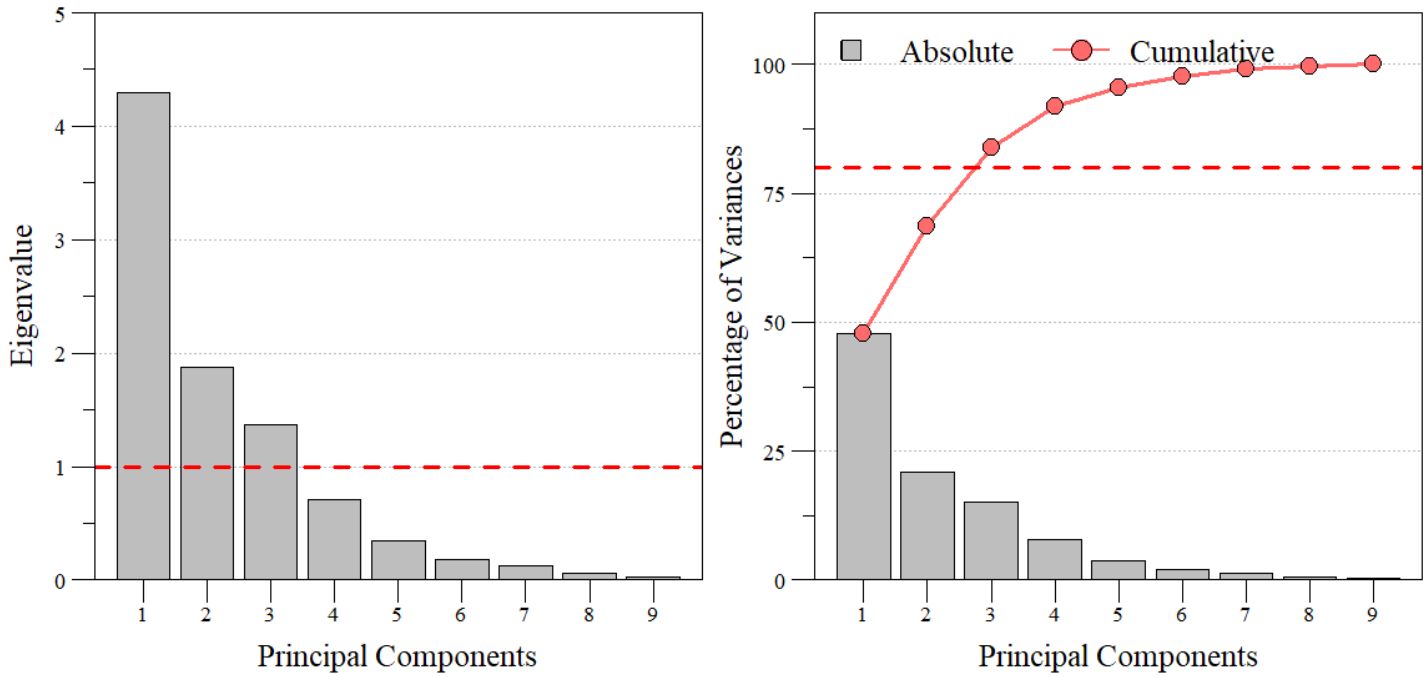


Generally there are two general rules:

1. Pick components with eignvalues of at least 1.
    - This is called the Kaiser rule.
2. The selected components should be able to describe at least 80% of the variance.
    - If you look at `eig.pca` you'll see that based on these criteria component 1, 2 and 3 are the components to focus on as they are enough to describe the data.

A scree plot displays these data and shows how much variation each component captures from the data.

Scree plots



Left: Scree plot of eigenvalues for each prinicipal component with the Kaiser threshold identified. Right: Scree plot of the variance and cumulative variance for each priniciple component.

Biplot

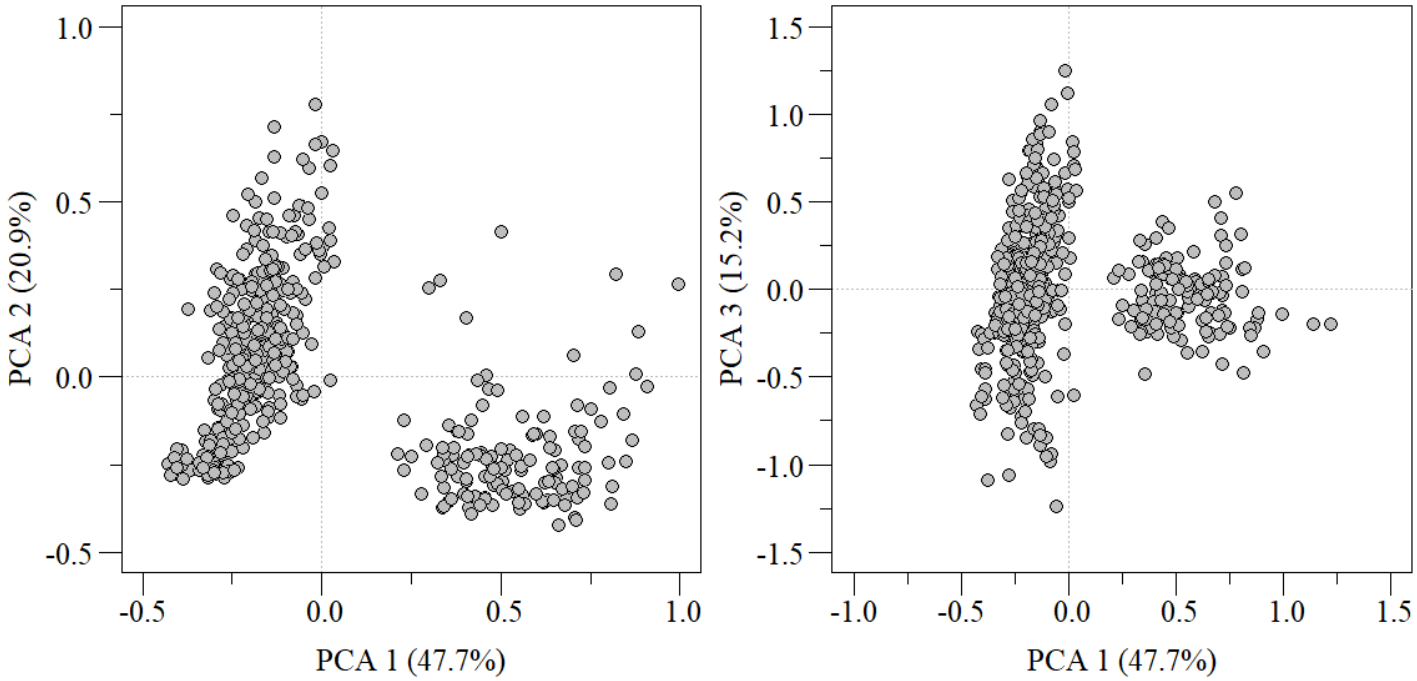
Now that we know which components are important, lets put together our biplot and extract components. To extract out components and specific loadings we can use the `scores()` function in the `vegan` package.

- It is a generic function to extract scores from `vegan` oridination objects such as RDA, CCA, etc.
  - This function also seems to work with `prcomp` and `princomp` PCA functions in `stats` package.

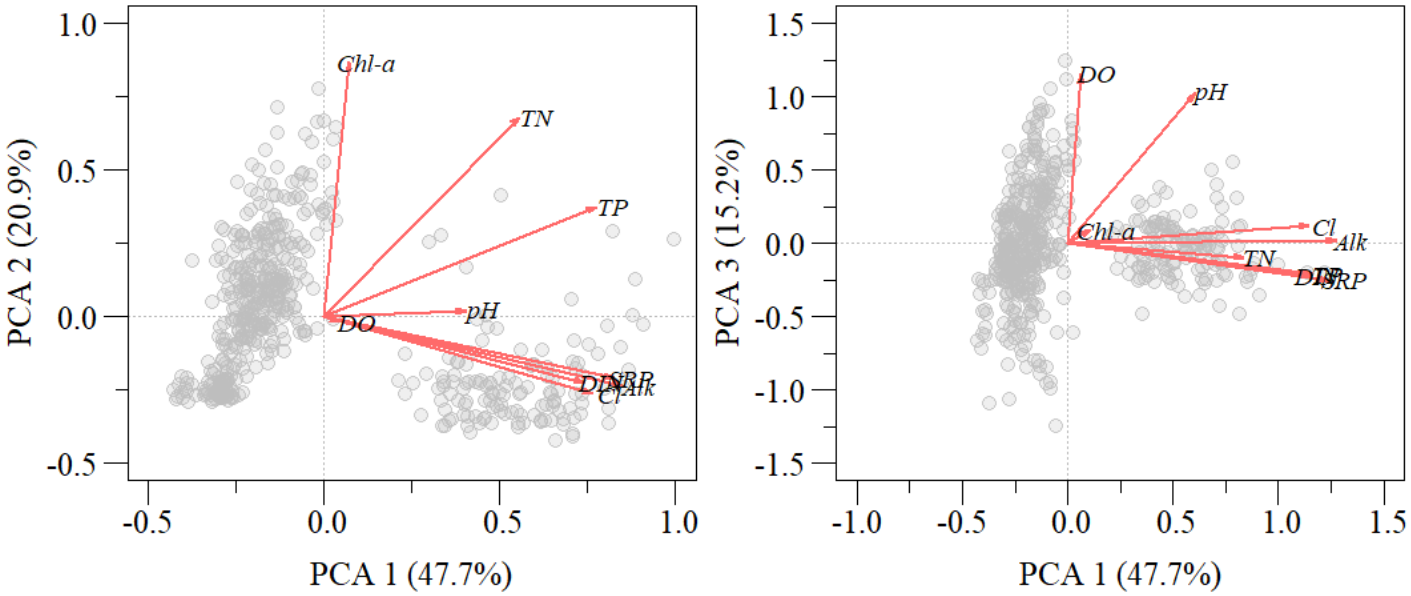
```
scrs <- scores(dat.xtab.pca,display=c("sites","species"),choices=c(1,2,3));
```

- `scrs` is a list of two item, species and sites. Species corresponds to the columns of the data and sites correspond to the rows.
  - Use `choices` to extract the components you want, in this case we want the first three components. Now we can plot the scores.

Biplot



PCA biplot of two component comparisons from the $data.\ xtab.\ pca$ analysis.



PCA biplot of two component comparisons from the $data.\ xtab.\ pca$ analysis with rescaled loadings.

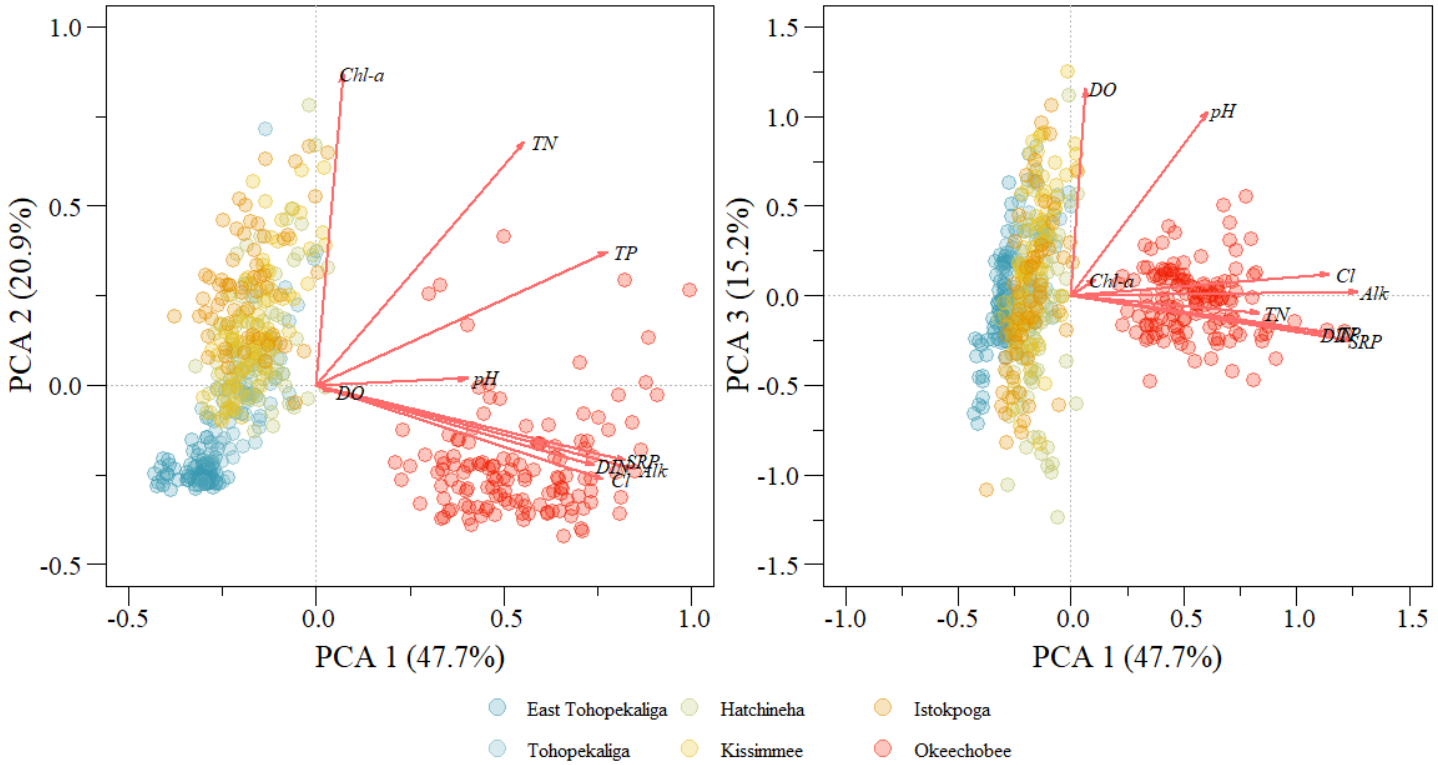
Typically when you see a PCA biplot, you also see arrows of each variable. This is commonly called loadings and can interpreted as:

- When two vectors are close, forming a small angle, the variables are typically positively correlated.
- If two vectors are at an angle 90° they are typically not correlated.
- If two vectors are at a large angle say in the vicinity of 180° they are typically negatively correlated.

You can take this one even further by showing how each lake falls in the ordination space by joining the `sites` to the original data frame. This is also how you use the derived components for further analysis.

```
dat.xtab <- cbind(dat.xtab,scrs$sites)

head(dat.xtab,3L)
```

```
##   Station.ID              LAKE   Date.EST Alk   Cl Chla  DO  pH    SRP    TP
## 1        A03 East Tohopekaliga 2005-05-17  17 19.7  4.0 7.9 6.1 0.0015 0.024
## 2        A03 East Tohopekaliga 2005-06-21  22 15.4  4.7 6.9 6.4 0.0015 0.024
## 4        A03 East Tohopekaliga 2005-08-16  17 14.0  3.0 6.9 6.3 0.0015 0.021
##     TN  DIN        PC1        PC2        PC3
## 1 0.71 0.04 -0.3901117 -0.2240239 -0.5666993
## 2 0.68 0.03 -0.3912797 -0.2083258 -0.6284024
## 4 0.55 0.03 -0.4290627 -0.2486860 -0.6599207
```



PCA biplot of two component comparisons from the $data.\ xtab.\ pca$ analysis with rescaled loadings and Lakes identified.

- You can extract a lot of great information from these plots and the underlying component data but immediately we see how the different lakes are group and how differently the lakes are loaded with respect to the different variables.

pauljulianphd@gmail.com

github.com/SwampThingPaul/PCA_Workshop