R Programming For Natural Resource Professionals

Week 14/15

Simulations: resampling/bootstrapping

Learning objectives for this “week”

1. Understand use cases for multivariate statistics in natural resource management/research
2. Learn to run common multivariate analyses
3. Learn to interpret the results of common multivariate analyses

What is multivariate statistics?

- We’ve been working with univariate stats
  - Response ~ predictor1 + predictor2….
- Multivariate stats involve many response variables
  - Without predictors: Interdependence methods
  - With predictors: Dependence methods

Interdependence methods

Dimension reducing methods

- Principal component analysis
- Correspondence analysis
- Factor analysis

Cluster analysis

- K-means clustering
- Hierarchical clustering

Multidimensional scaling

- Non-metric multidimensional scaling

Dependence methods

Discrimination and classification methods

- Canonical variate analysis
- Neural networks
- Random forest

Constrained ordination

- Canonical correlation analysis
- Redundancy analysis

Multivariate regression

Principal component analysis (PCA)

Would take awhile to explain the math behind PCA..

- Linear algebra and eigensystems. Interested? Take a multivariate stats course!

Goal: simplify large and complex data sets

- “Reduce dimensionality”

All data can be perfectly summarized (all variance explained) if the number of dimensions used to explain it equals number of observations

- Not really in practice though

Principal component analysis (PCA)

- Explaining data variation in “dimensions”
- Consider a data set with three variables: x, y, z
- PCA calculates a covariation matrix
  - In what ways are x, y, and z correlated/related?

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

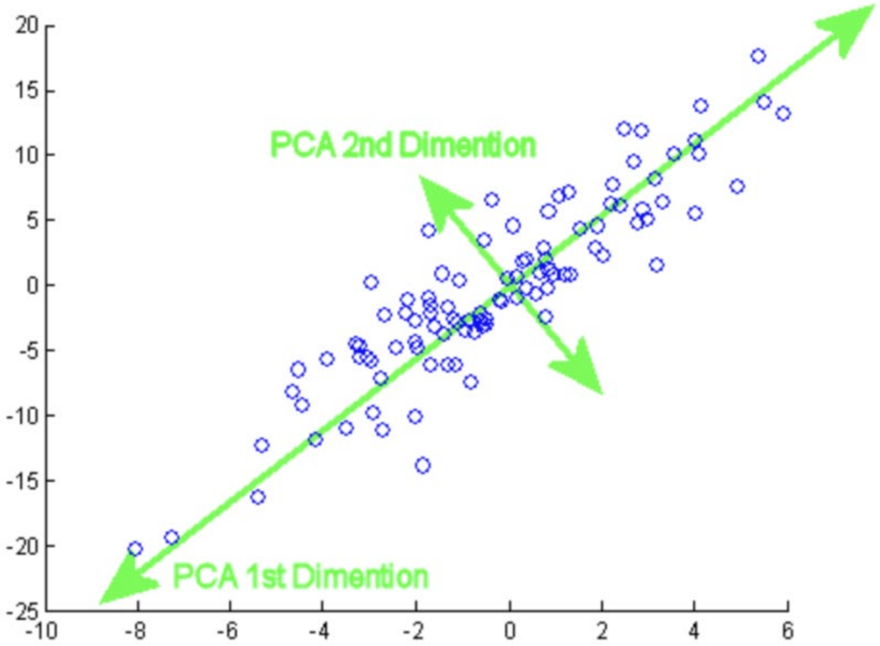
Principal component analysis (PCA)

$$\left[\begin{array}{ccc} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{array}\right]$$

- Covariation is summarized into eigenvectors
- Patterns of covariation are assigned to “principal component axes” or “principal components” or “axes” or “PCs”

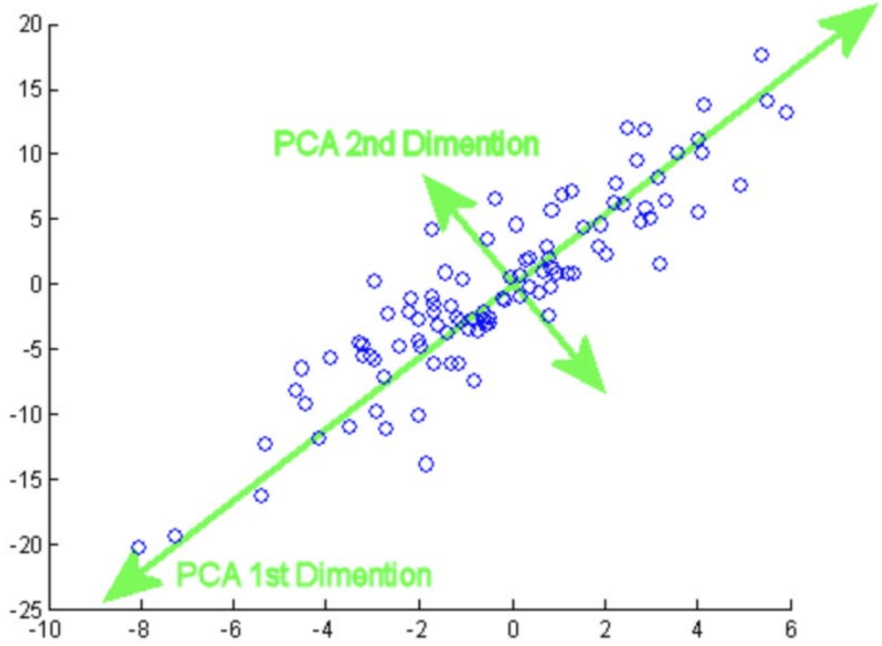
Principal component analysis (PCA)

- Principal component axes are orthogonal to each other
  - They characterize different "dimensions" of the variation
    - E.g., 10 principal components = 10 dimensions



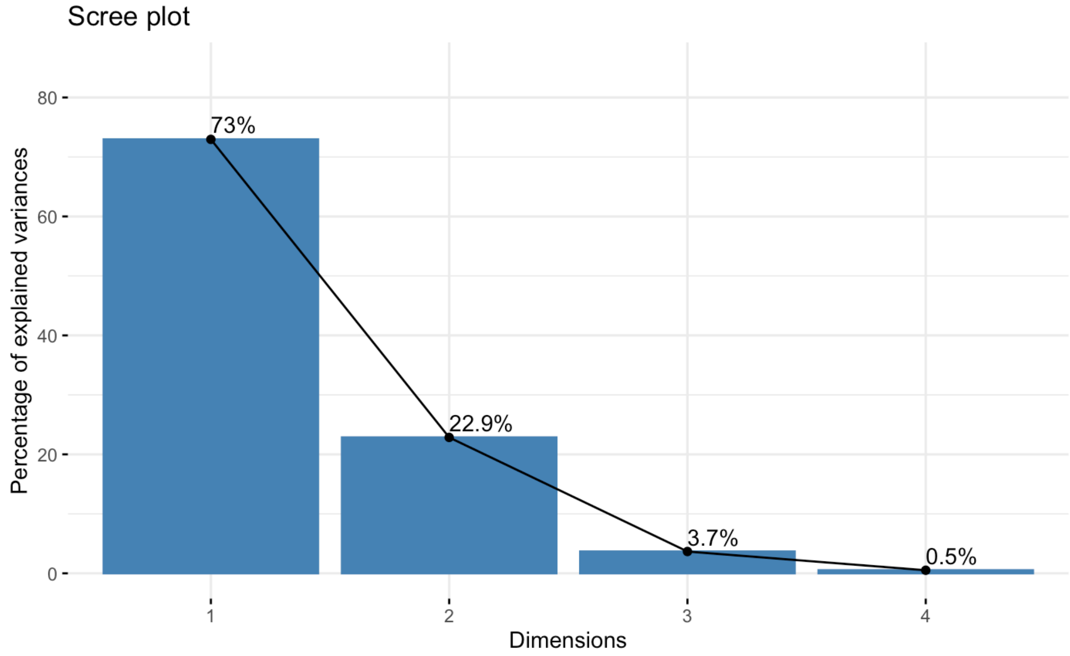
Principal component analysis (PCA)

- Principal component axes are orthogonal to each other
  - They characterize different “dimensions” of the variation
    - E.g., 10 principal components = 10 dimensions
- First axis explains the most variation
  - Rest in descending order



Principal component analysis (PCA)

- Principal component axes are orthogonal to each other
  - They characterize different “dimensions” of the variation
    - E.g., 10 principal components = 10 dimensions
- First axis explains the most variation
  - Rest in descending order
- Variance explained often decays quickly
  - Visualized with a scree plot

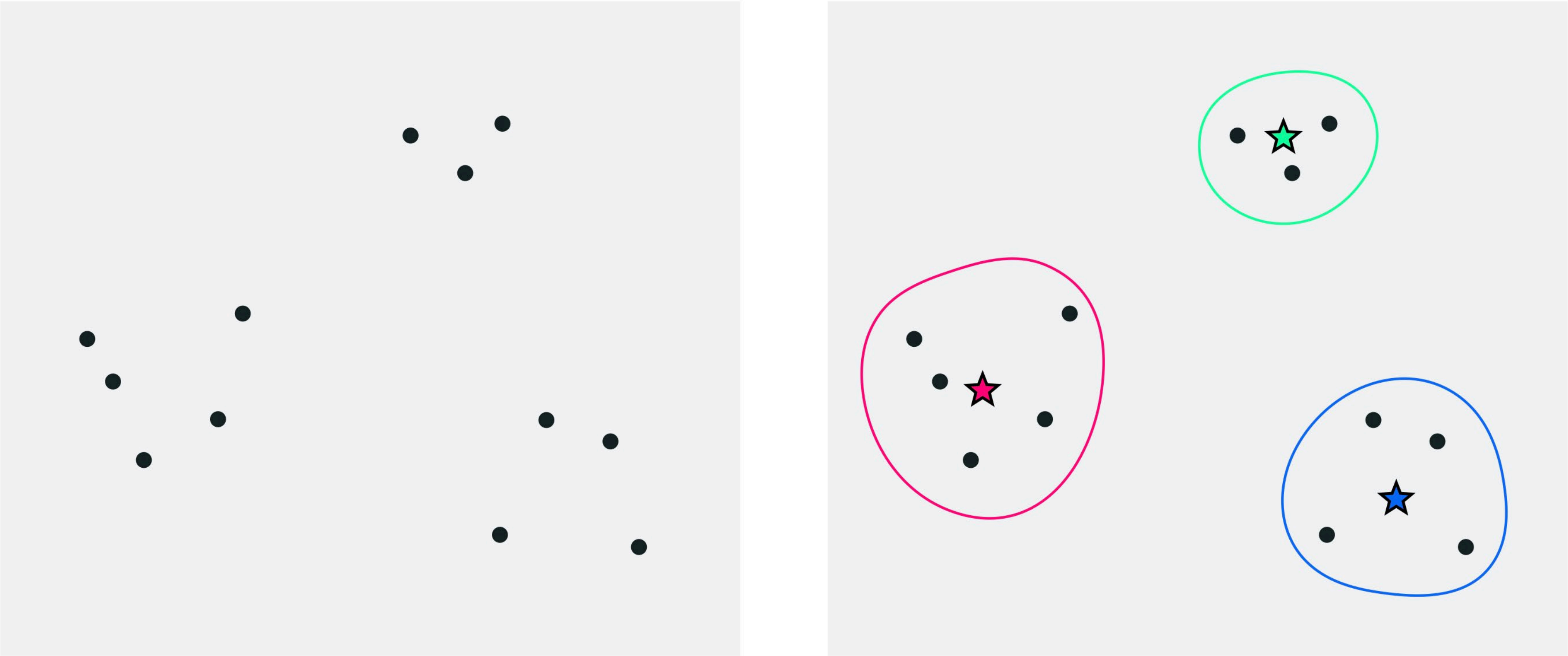


Principal component analysis (PCA)

- Best practice is usually to scale and center data before PCA
- In prcomp() use arguments ‘scale’ and ‘center’
  - Scale: standardize range of variable values
  - Center: center variable values on zero

K-means clustering

- Goal: group observations based on their overall similarity

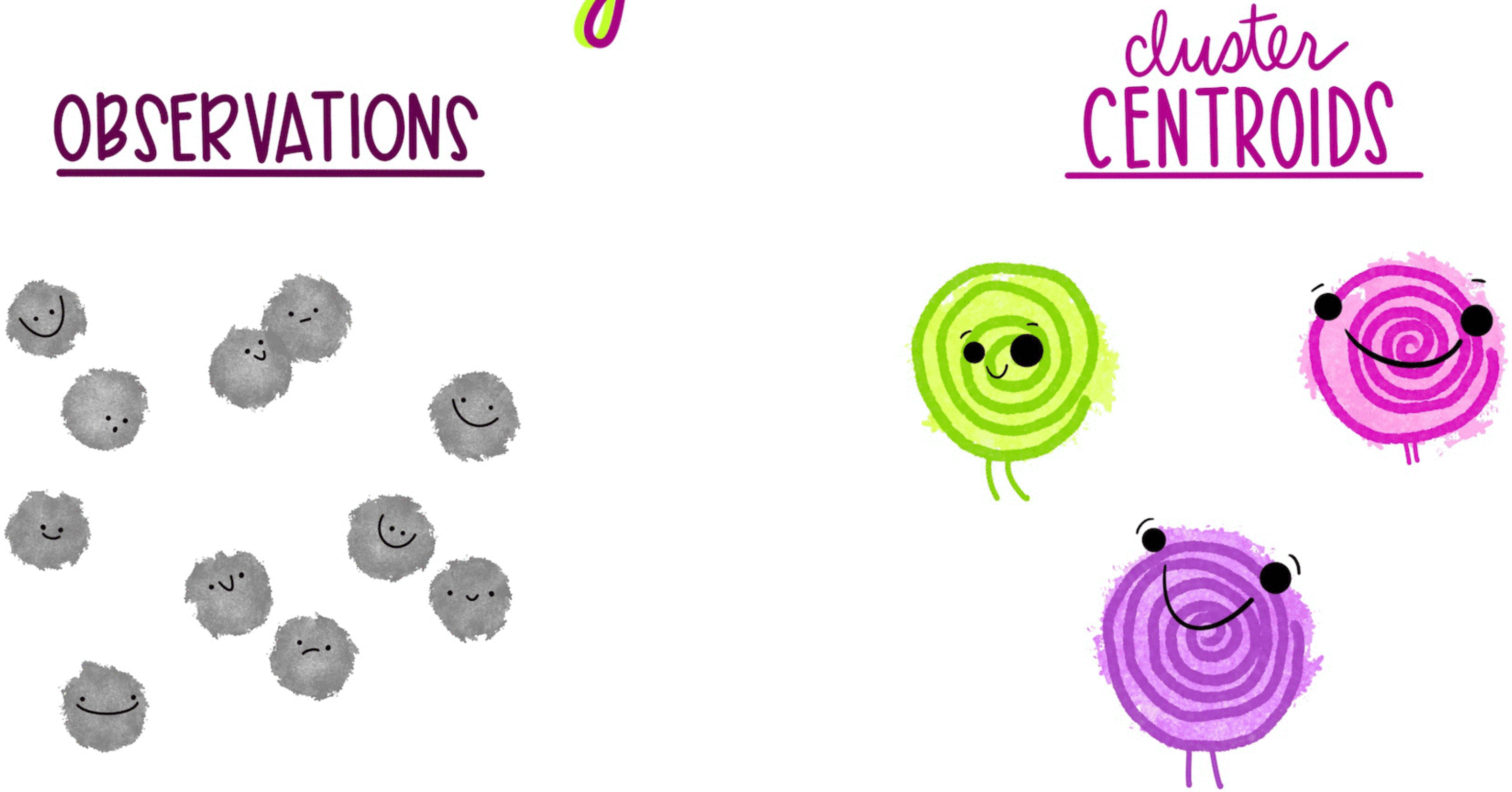


K-means clustering

- Goal: group observations based on their overall similarity
- Algorithm:
  - Place “centroids” among a scatterplot
    - Assessed number of centroids determined by either biological hypothesis or identified statistically
  - Measure distance from each point to each centroid
  - Move centroids to locations where distances are minimized
  - Repeat until optimal solution found (smallest average distance from a point to its nearest centroid).

k-means clustering

: assign each observation to one of k clusters based on the nearest cluster centroid.

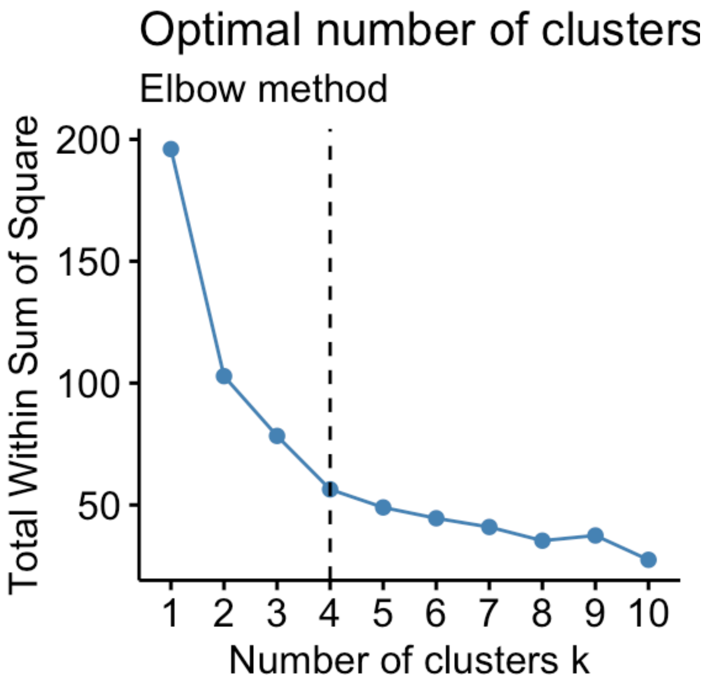


@allison_horst

K-means clustering

How many clusters (K) are there?

- Within-cluster sum of squares
- Goal is to minimize within-cluster distance from each point to a centroid
- “Elbow method” for determining best fit
- Lots of other methods for determining best K



Nonmetric multidimension scaling

- Goal: Use distance measures to identify differences among count data
  - Often used to identify correlates of differences in biotic communities

- ‘Distances measures’ include a variety of methods for calculating covariation among observations in a matrix
  - Bray-Curtis often used for community data

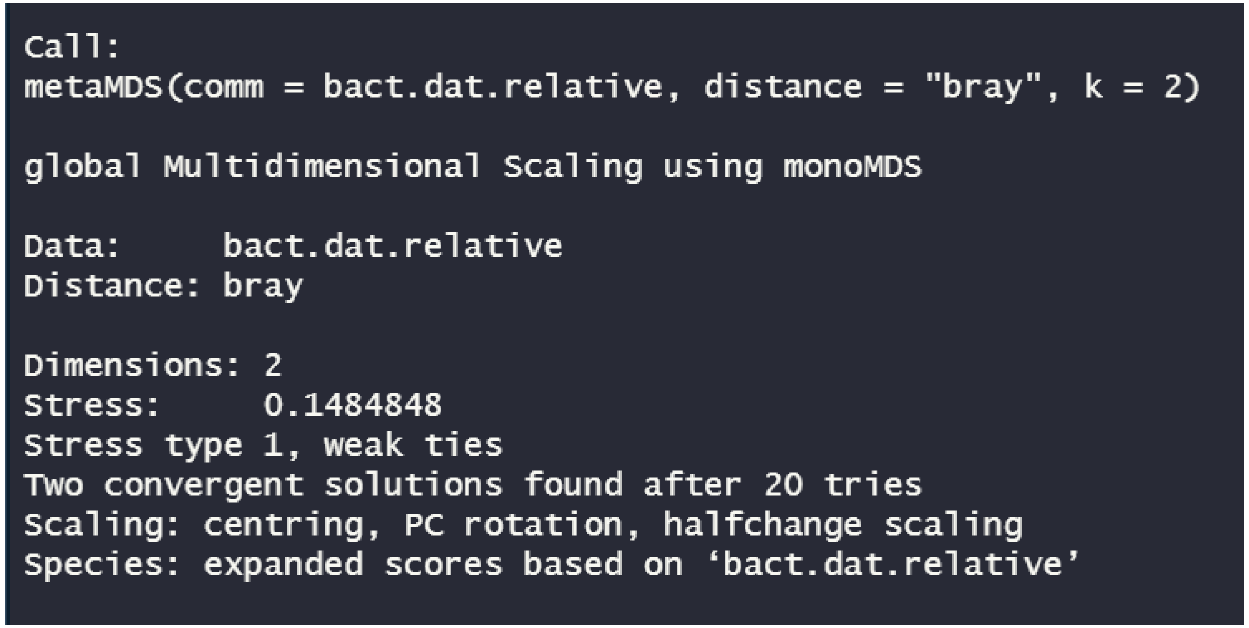
Nonmetric multidimension scaling

- vegan::metaMDS is a common approach in R
- metaMDS can calculate the distance matrix for you

- Environmental metrics can be fit to data using vegan::envfit

metaMDS output

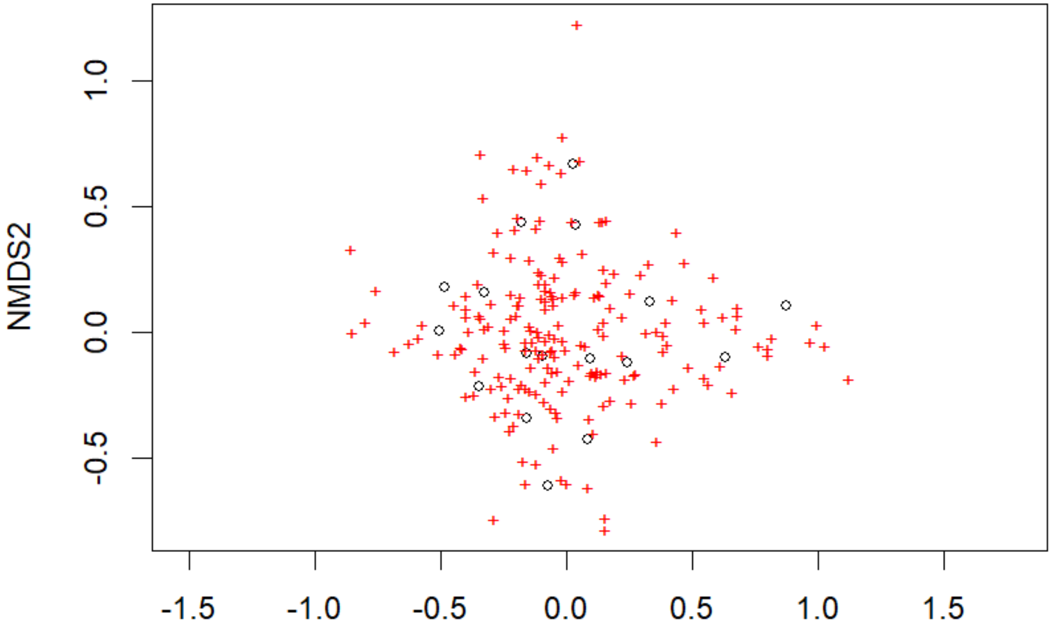
- “Stress” is a key metric
  - How well does the NMDS model fit the data?
  - <0.05 is excellent
  - 0.05-0.1 is great
  - 0.1-0.2 is good/ok
  - 0.2-0.3 is poor

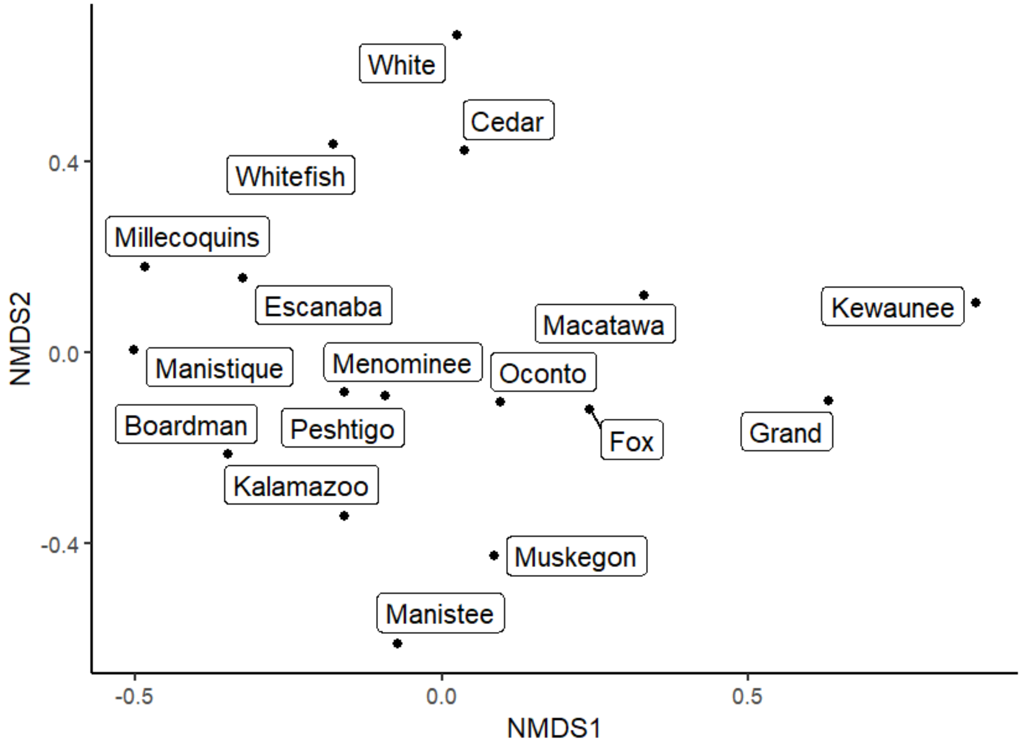


```
Call:
metaMDS(comm = bact.dat.relative, distance = "bray", k = 2)

global Multidimensional Scaling using monoMDS

Data:     bact.dat.relative
Distance: bray

Dimensions: 2
Stress:     0.1484848
Stress type 1, weak ties
Two convergent solutions found after 20 tries
Scaling: centring, PC rotation, halfchange scaling
Species: expanded scores based on ‘bact.dat.relative’
```

Plotting NMDS

- Default plot isn't great
- Can use ggplot if point coordinates are extracted
- scores(NMDS, display = "sites")


NMDS2


White
Cedar
Whitefish
Millecoquins
Escanaba
Macatawa
Kewaunee
Manistique
Menominee
Oconto
Boardman
Peshtigo
Fox
Grand
Kalamazoo
Muskegon
Manistee
NMDS2
NMDS1