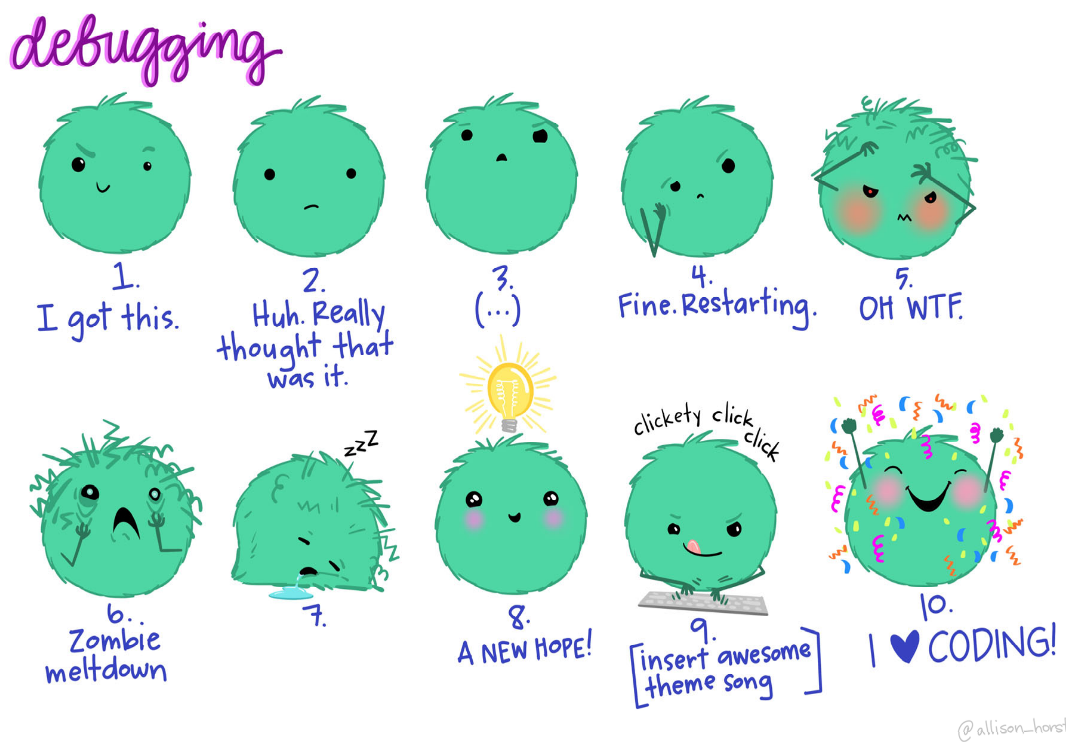
Last week: how are we feeling?

- Common types of multivariate data…what is dimensionality?
- Think about what we want to do with these data
- Talk about some of the common methods
- Learn to do Principle Components Analysis (PCA) and interpret a biplot
- Learn about how to measure distances in multidimensional space



Stats meme/post of the week

How to measure distance in species space

Bray-Curtis dissimilarity (distance)

	Site 1	Site 2	Site 3	Site 4
Site 1	0			
Site 2	0.78	0		
Site 3	1	0.45	0	
Site 4	0.5	0.45	0.33	0

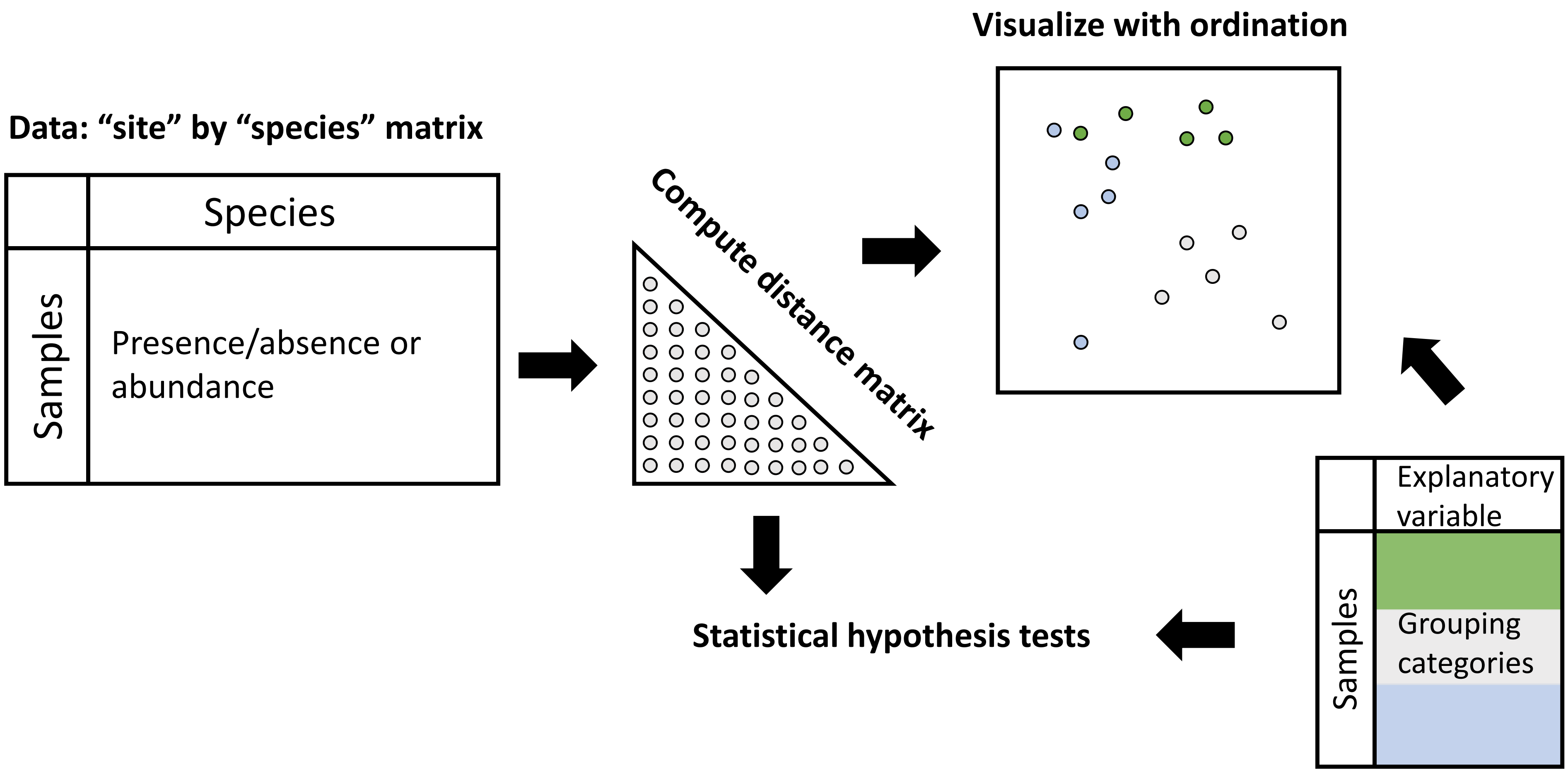
	Species A	Species B
Site 1	1	0
Site 2	3	5
Site 3	0	3
Site 4	1	2

Ranges 0 (the same) to 1 (no shared species)

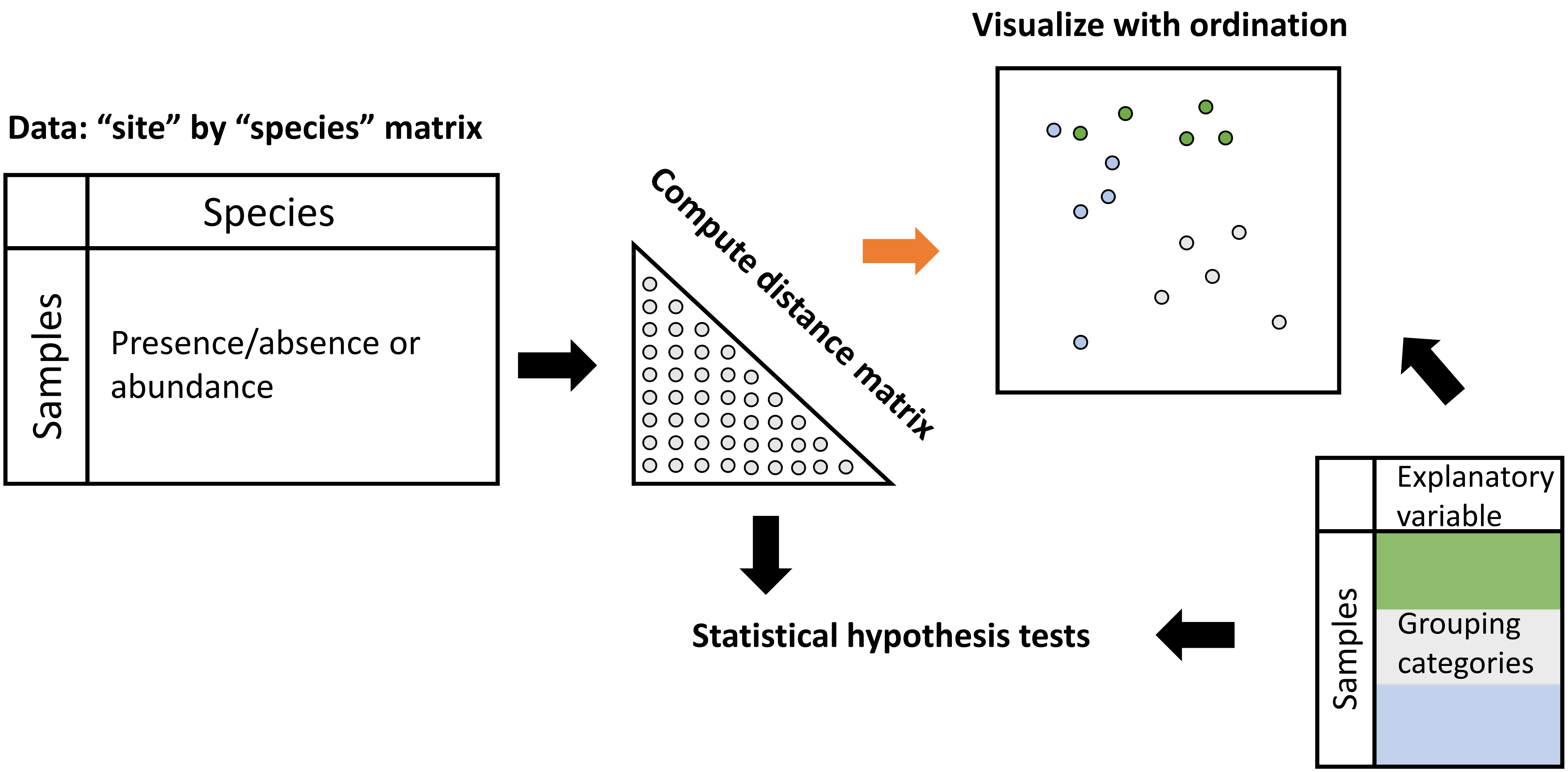
What can we do with a distance matrix?

1. Ordination: visualize and find patterns
2. Clustering: group samples based on distances
3. Analysis and hypothesis testing:
    a) Are distances between groups greater than distances within groups? (PERMANOVA)
    b) Are distances between samples within groups homogenous among groups? (PERMDISP)

General workflow for distance-based ordination and hypothesis testing



General workflow for distance-based ordination and hypothesis testing



Two major options for distanced-based, unconstrained ordination

1. PCoA: Principle Coordinates Analysis
   - Assumes *linear* relationship between distance matrix and ordination distance
     - Metric (eigen-based), calculated
   - Is a generalized version of PCA
     - PCoA with Euclidean distance matrix is the same as PCA
     - But can handle any other distance metrics (e.g. Bray-Curtis) suited to various data types

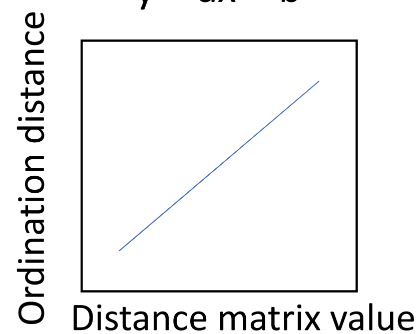
2. NMDS: Non-metric Multi-Dimensional Scaling
   - Assumes *monotonic* relationship between distance matrix and ordination distance
     - Not metric (rank-based), iterative
   - Can use with any type of distance metric
   - User predetermines the number of ordination axes

Two major options for distanced-based, unconstrained ordination

1. PCoA: Principle Coordinates Analysis
   - Assumes *linear* relationship between distance matrix and ordination distance
     - Metric (eigen-based), calculated
   - Is a generalized version of PCA
     - PCoA with Euclidean distance matrix is the same as PCA
     - But can handle any other distance metrics (e.g. Bray-Curtis) suited to various data types

What does “linear” mean?

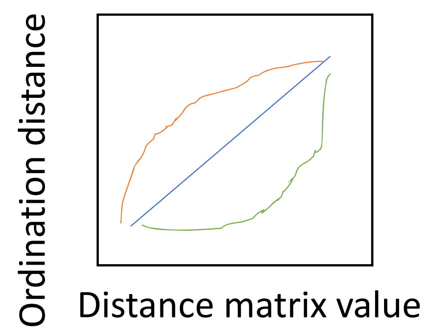
y = ax + b



2. NMDS: Non-metric Multi-Dimensional Scaling
   - Assumes *monotonic* relationship between distance matrix and ordination distance
     - Not metric (rank-based), iterative
   - Can use with any type of distance metric
   - User predetermines the number of ordination axes

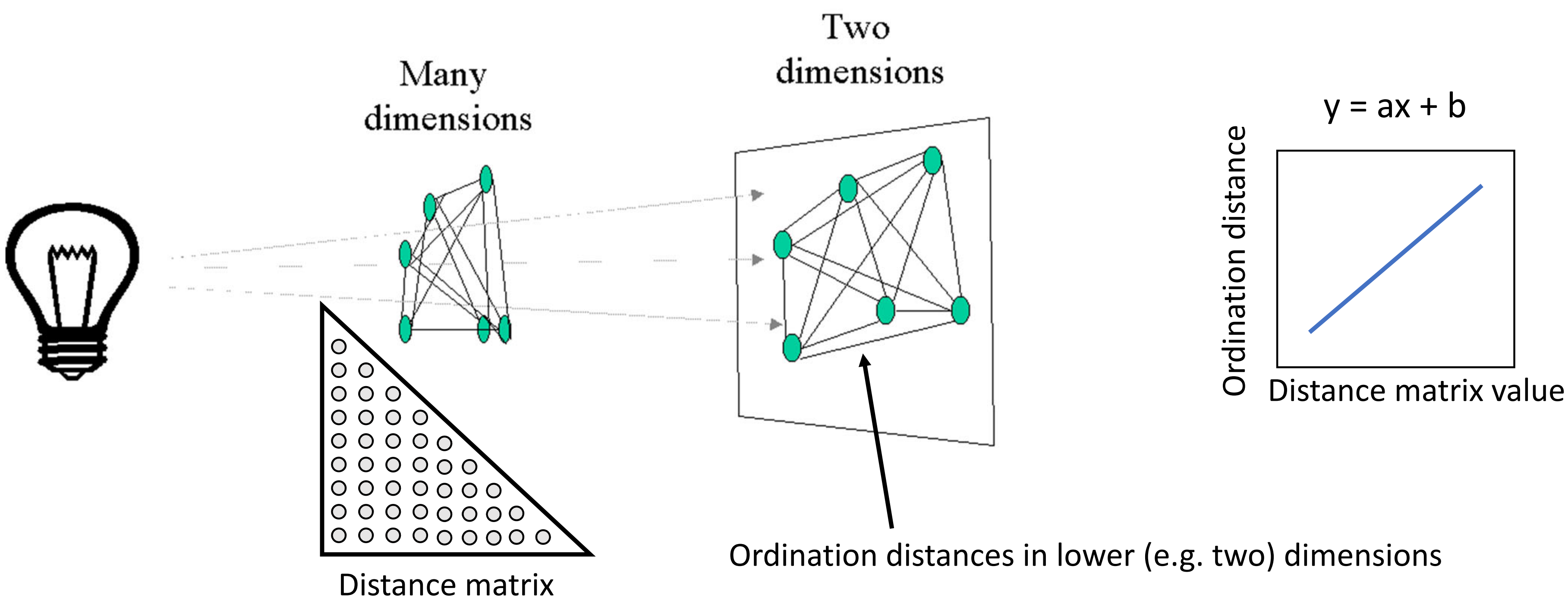
What does “monotonic” mean?

y increases with x



PCoA: Principle coordinates analysis

Maximizes the *linear correlation* between the distances in the distance matrix, and the distances in a space of low dimension

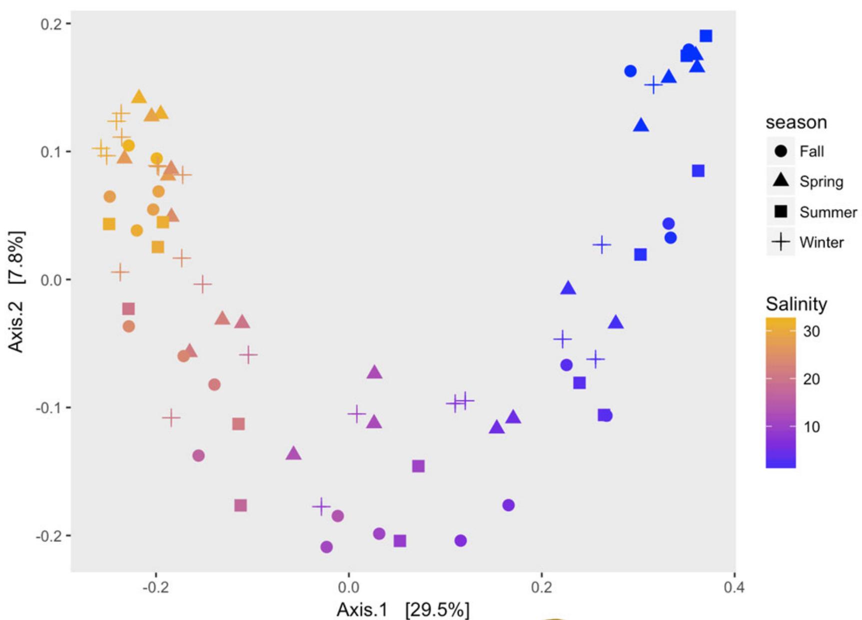


PCoA: Principle coordinates analysis

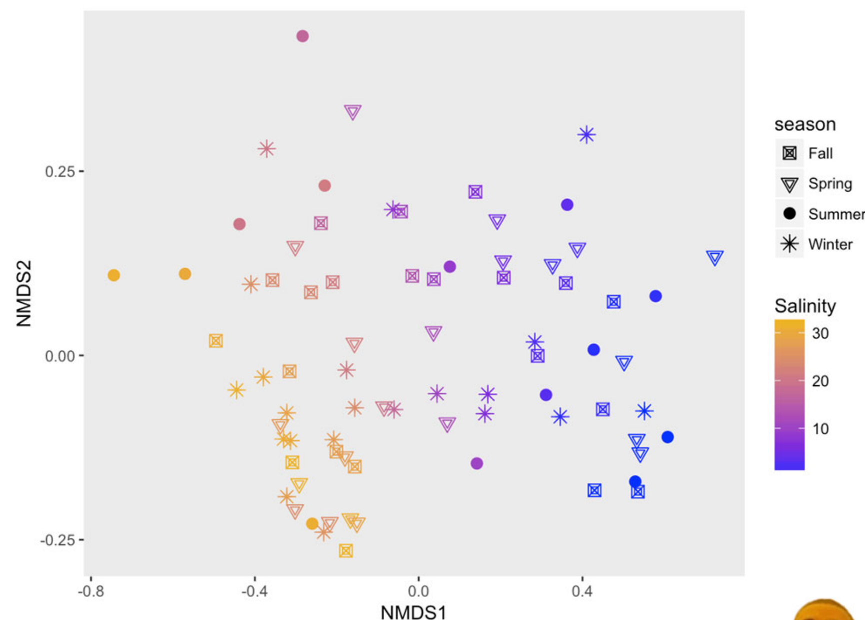
- You get as many orthogonal “principle coordinates” as there are dimensions in the data (just like PCA)
- Principle coordinates do not necessarily decrease in sequence of %variance explained
- Pro: more flexible than PCA
- Con: if using non-Euclidean distances, PCoA may spit out negative eigenvalues which can’t be mapped onto real ordination axes (usually not huge issue)
- Used to not be very popular
  - if you have data that’s normal, might as well do PCA (easier interpretation)
  - if you have zero-inflated count/proportional data, NMDS has fewer caveats
- Popularized in recent years for microbiome data because it is the default ordination option in QIIME 😒

One sign that your PCoA is not great and you should probably try NMDS instead

PCoA: “Arch” or “Horseshoe” effect

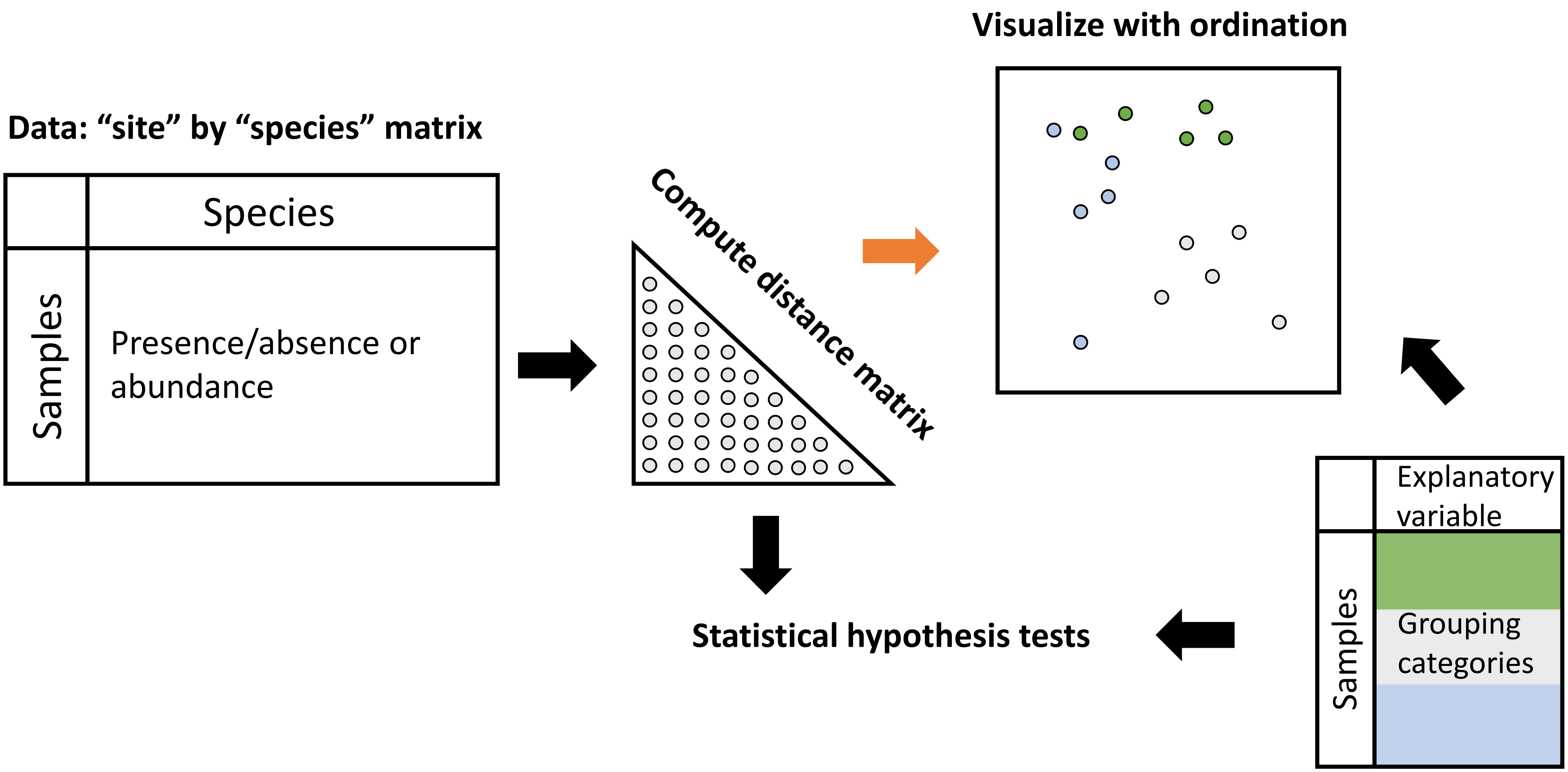


NMDS on the same dataset



Break

NMDS: Non-metric Multi-Dimensional Scaling



NMDS: Non-metric Multi-Dimensional Scaling

Instead of using the actual values in the distance matrix we use only their rank

	Sp A	Sp B
Site 1	1	0
Site 2	3	5
Site 3	0	3
Site 4	1	2

Data: “site” by “species” matrix

	Site 1	Site 2	Site 3	Site 4
Site 1	0			
Site 2	0.78	0		
Site 3	1	0.45	0	
Site 4	0.5	0.45	0.33	0

Bray-Curtis distance matrix

	Site 1	Site 2	Site 3	Site 4
Site 1	0			
Site 2		0		
Site 3			0	
Site 4				0

Rank of distance

NMDS: Non-metric Multi-Dimensional Scaling

Step 1: Rank calculated distances from smallest to largest distance

	Sp A	Sp B
Site 1	1	0
Site 2	3	5
Site 3	0	3
Site 4	1	2

Data: “site” by “species” matrix

	Site 1	Site 2	Site 3	Site 4
Site 1	0			
Site 2	0.78	0		
Site 3	1	0.45	0	
Site 4	0.5	0.45	0.33	0

Bray-Curtis distance matrix

	Site 1	Site 2	Site 3	Site 4
Site 1	0			
Site 2	5	0		
Site 3	6	2	0	
Site 4	4	2	1	0

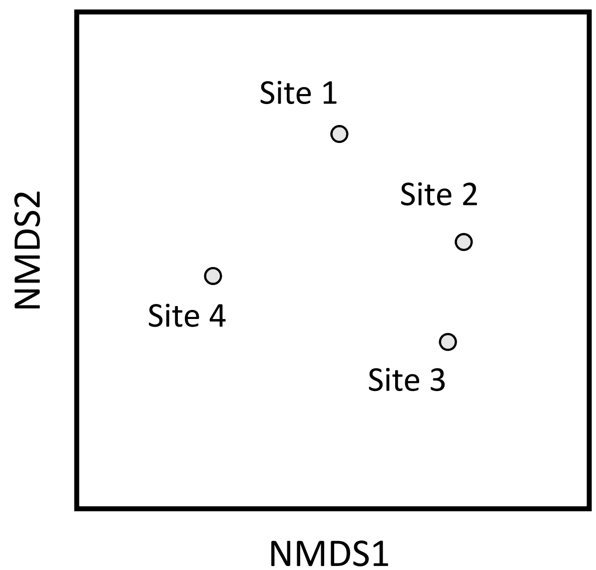
Rank of distance

NMDS: Non-metric Multi-Dimensional Scaling

Step 2: Organize points on *predetermined*, lower-dimensional space (usually 2D or 3D) in some sort of *random* starting configuration

	Site 1	Site 2	Site 3	Site 4
Site 1	0			
Site 2	5	0		
Site 3	6	2	0	
Site 4	4	2	1	0

Rank of distance

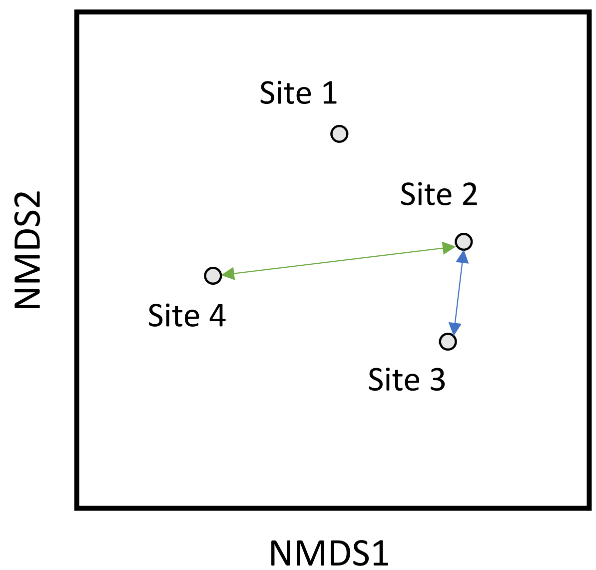


NMDS: Non-metric Multi-Dimensional Scaling

Step 3: Compare the ranked ordination distances of the configuration to their original ranks in multidimensional space.

	Site 1	Site 2	Site 3	Site 4
Site 1	0			
Site 2	5	0		
Site 3	6	2	0	
Site 4	4	2	1	0

Rank of distance



	Site 1	Site 2	Site 3	Site 4
Site 1	0			
Site 2	2	0		
Site 3	4	1	0	
Site 4	3	6	5	0

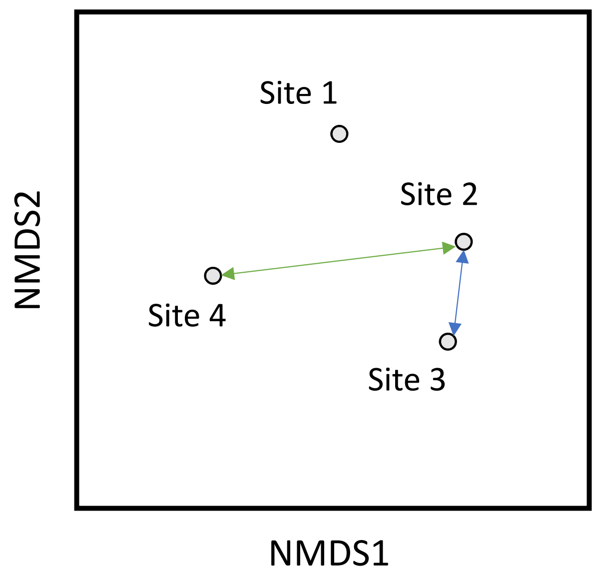
Rank of ordination distance

NMDS: Non-metric Multi-Dimensional Scaling

Step 3: Compare the ranked ordination distances of the configuration to their original ranks in multidimensional space. Calculate stress (goodness of fit between original and ordination distances).

	Site 1	**Site 2**	**Site 3**	**Site 4**
Site 1	0			
Site 2	5	0		
Site 3	6	2	0	
Site 4	4	2	1	0

Rank of distance



	Site 1	**Site 2**	**Site 3**	**Site 4**
Site 1	0			
Site 2	2	0		
Site 3	4	1	0	
Site 4	3	6	5	0

Rank of ordination distance

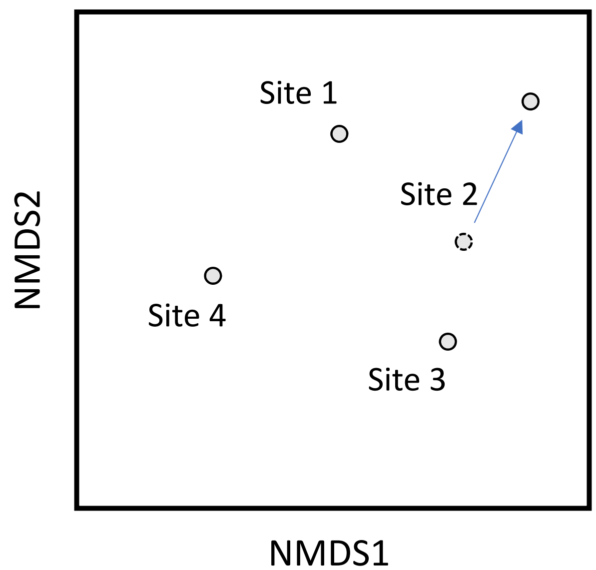
NMDS: Non-metric Multi-Dimensional Scaling

Repeat steps 2 and 3: Iteratively move the points around in ordination space until the calculated stress value for that configuration is *as low as possible*.

	Site 1	**Site 2**	**Site 3**	**Site 4**
Site 1	0			
Site 2	5	0		
Site 3	6	2	0	
Site 4	4	2	1	0

Rank of distance

Breakout groups: try to ordinate these sites!

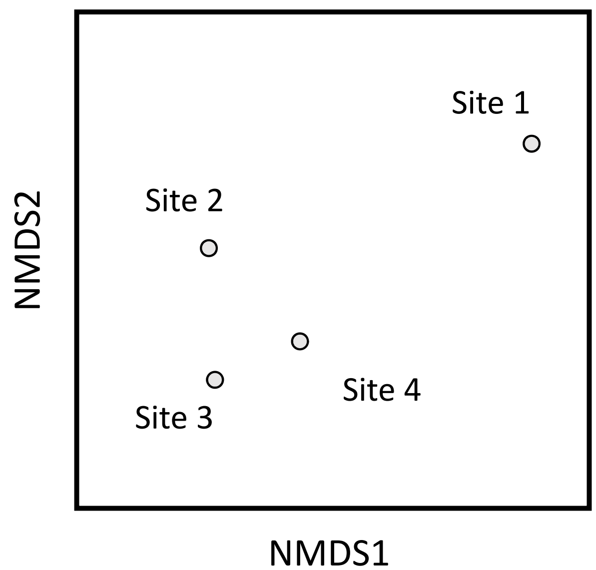


NMDS: Non-metric Multi-Dimensional Scaling

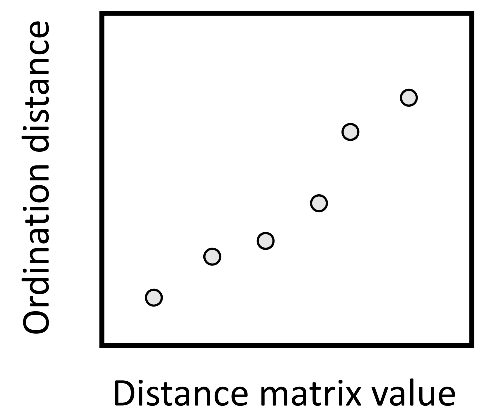
Repeat steps 2 and 3: Iteratively move the points around in ordination space until the calculated stress value for that configuration is *as low as possible*.

	Site 1	Site 2	Site 3	Site 4
Site 1	0			
Site 2	5	0		
Site 3	6	2	0	
Site 4	4	2	1	0

Rank of distance



Monotonic relationship
= low stress values

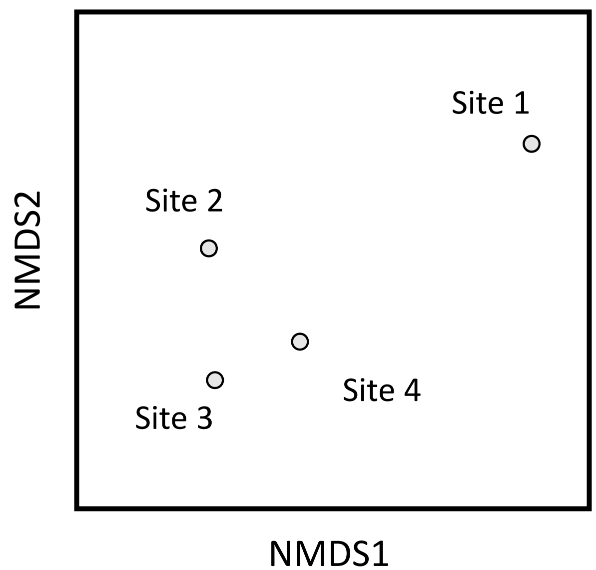


NMDS: Non-metric Multi-Dimensional Scaling

Step 4: Huzzah! Now you have your final NMDS ordination!

On an NMDS ordination, points closer to each other (in ordination space) are more similar to each other in composition (multidimensional "species" space).

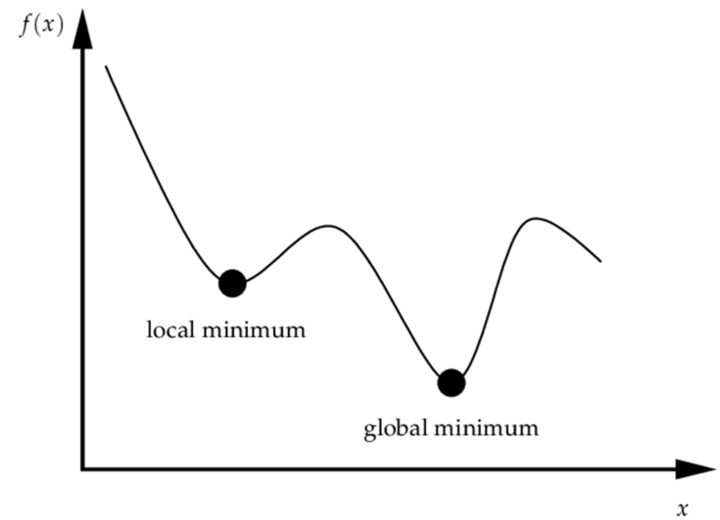
	Sp A	**Sp B**
Site 1	1	0
Site 2	3	5
Site 3	0	3
Site 4	1	2



NMDS: Non-metric Multi-Dimensional Scaling

The final ordination solution depends on:

1. The number of lower dimensions/axes (*k*) chosen
2. But even given the same *k,* may still vary depending on whether the algorithm found the truly best (lowest stress) solution possible
3. There may be more than one solution with the same, lowest, stress value.



A general rule for stress values: <0.2 is great, 0.2-0.3 is iffy, stress values > 0.3 means the ordination solution is not a good reflection of distance ranks in the original data.

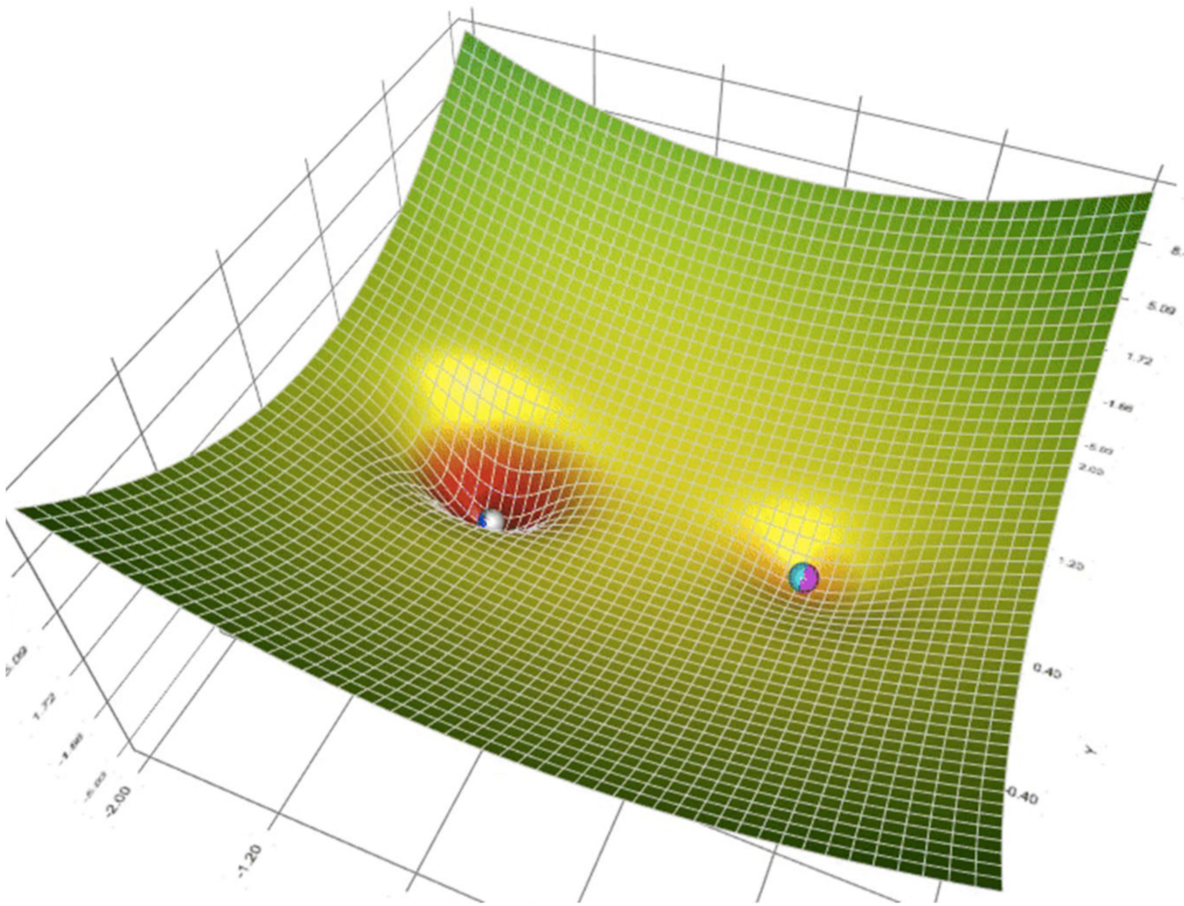
NMDS: Non-metric Multi-Dimensional Scaling

The final ordination solution depends on:

1. The number of lower dimensions/axes (*k*) chosen
2. But even given the same *k*, may still vary depending on whether the algorithm found the truly best (lowest stress) solution possible
3. There may be more than one solution with the same, lowest, stress value.

A general rule for stress values: <0.2 is great, 0.2-0.3 is iffy, stress values > 0.3 means the ordination solution is not a good reflection of distance ranks in the original data.

Make sure to find global best solution by using many random starts and iterations

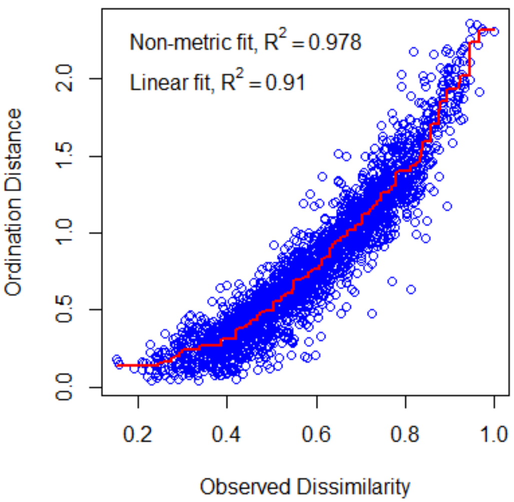


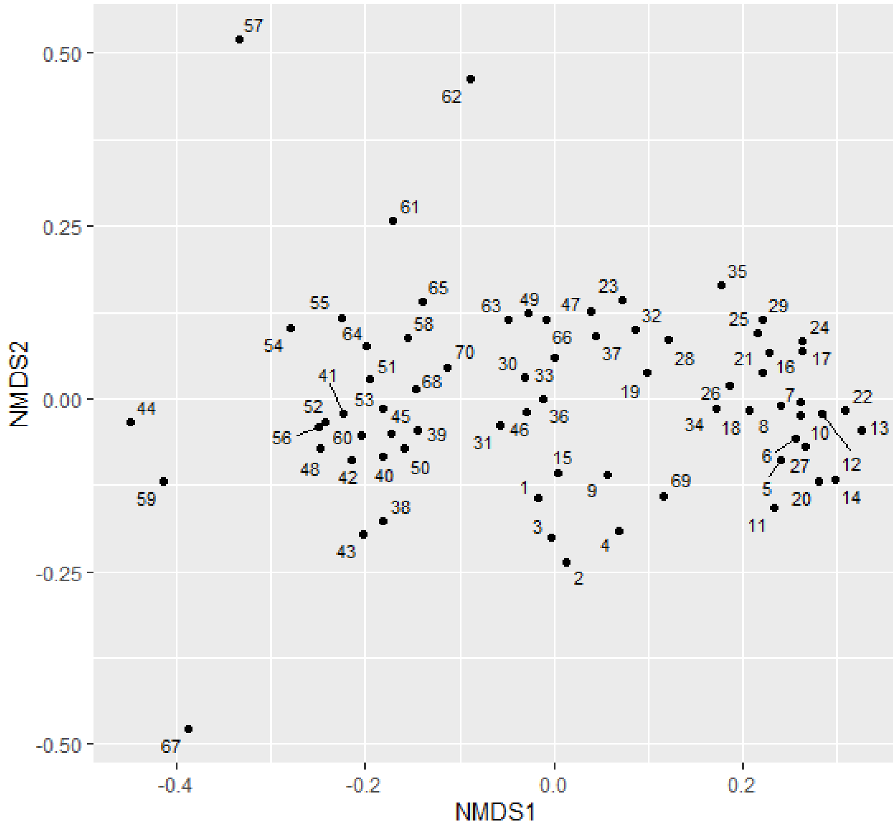
Example dataset

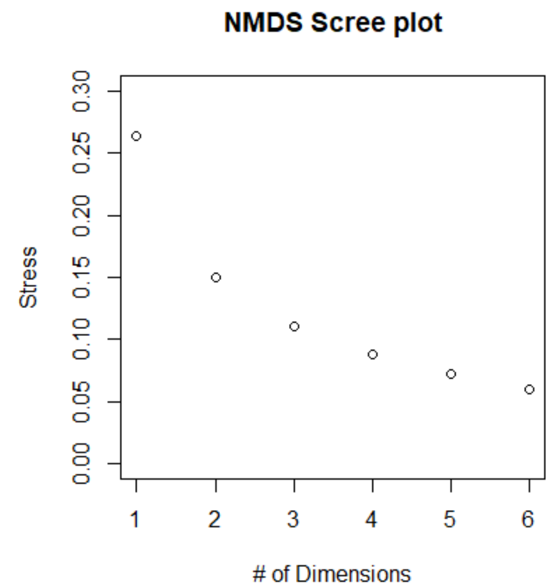
- mite_abund_matrix.csv: 70 samples and their mite composition
- mite_explain_var.csv: some information about the different properties of each sample that might explain difference in composition

(Borcard and Legendre 1994)

NMDS initial outputs







NMDS (final) output

It looks like the abundance of shrubs is an important factor in determining mite community composition

How can we test this hypothesis statistically?