Session 3A:
Statistical considerations for ‘omic data analysis

Jacob Price, PhD
Maria Sevillano, PhD student

@Jake_in_the_Lab @katty_sevi

AEESP | Association of Environmental Engineering & Science Professors

Prelude & Motivation

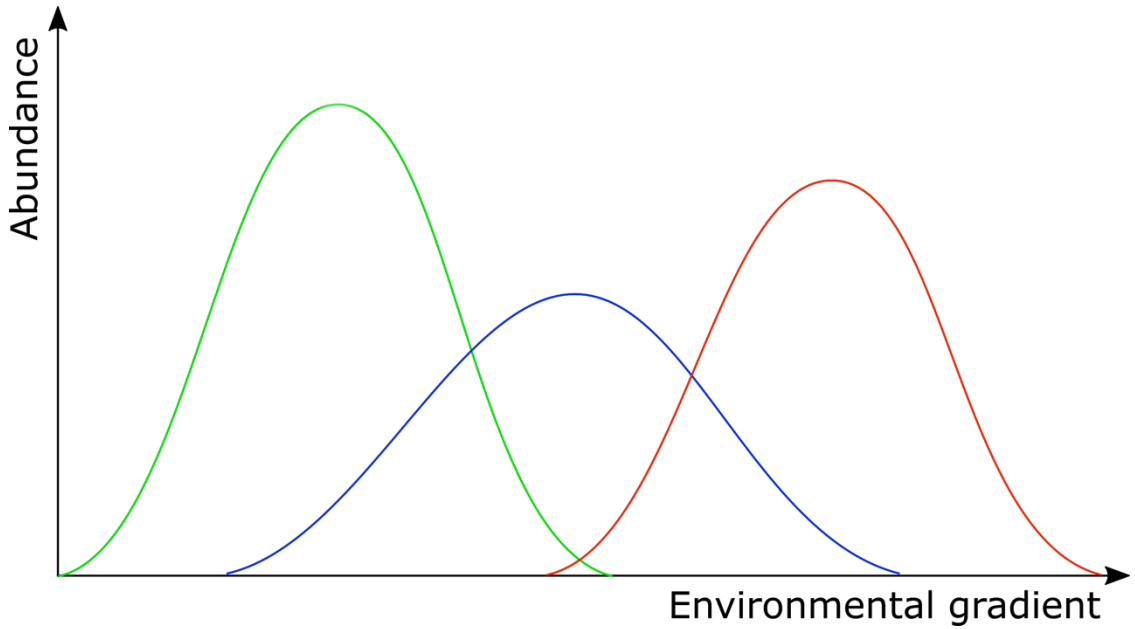


Figure: ourcodingclub.github.io

Prelude & Motivation

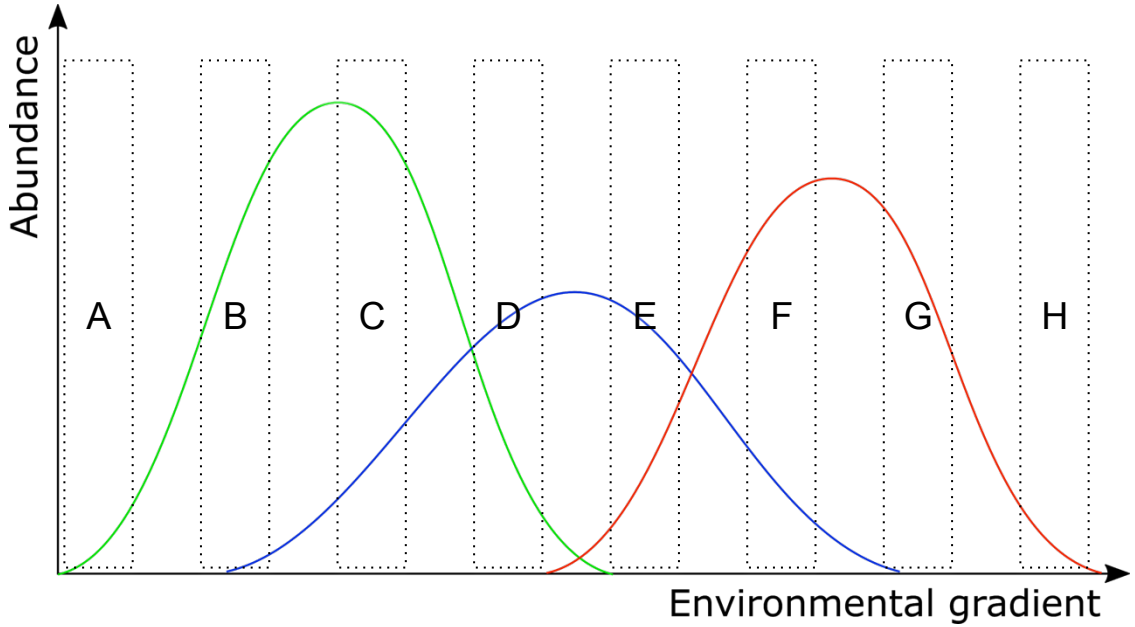


Figure: ourcodingclub.github.io

Prelude & Motivation

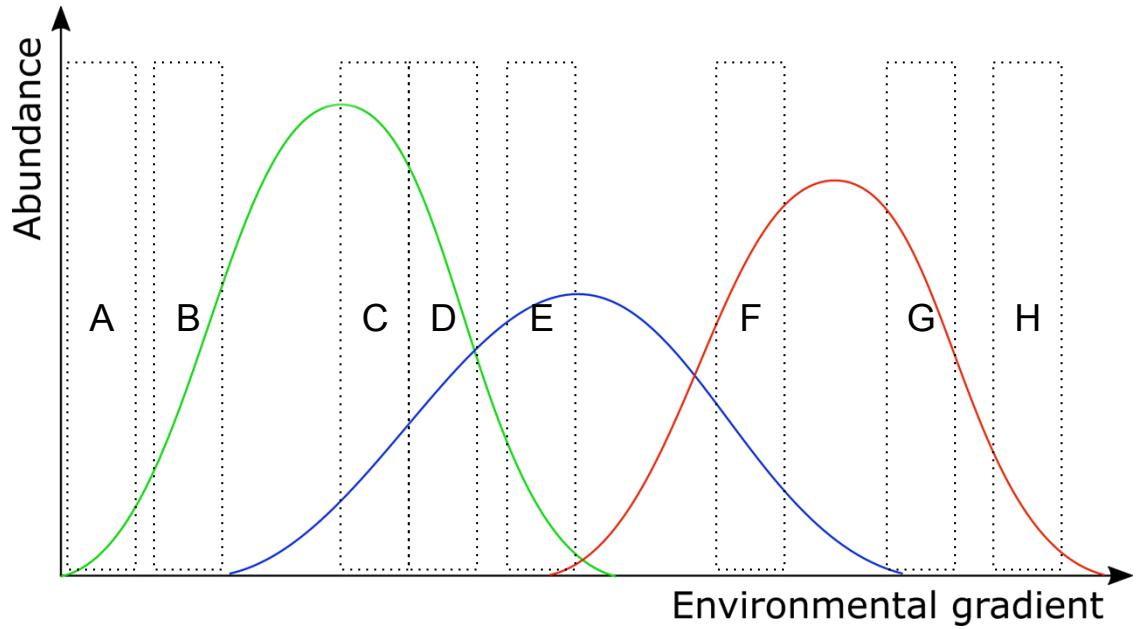


Figure: ourcodingclub.github.io

Prelude & Motivation

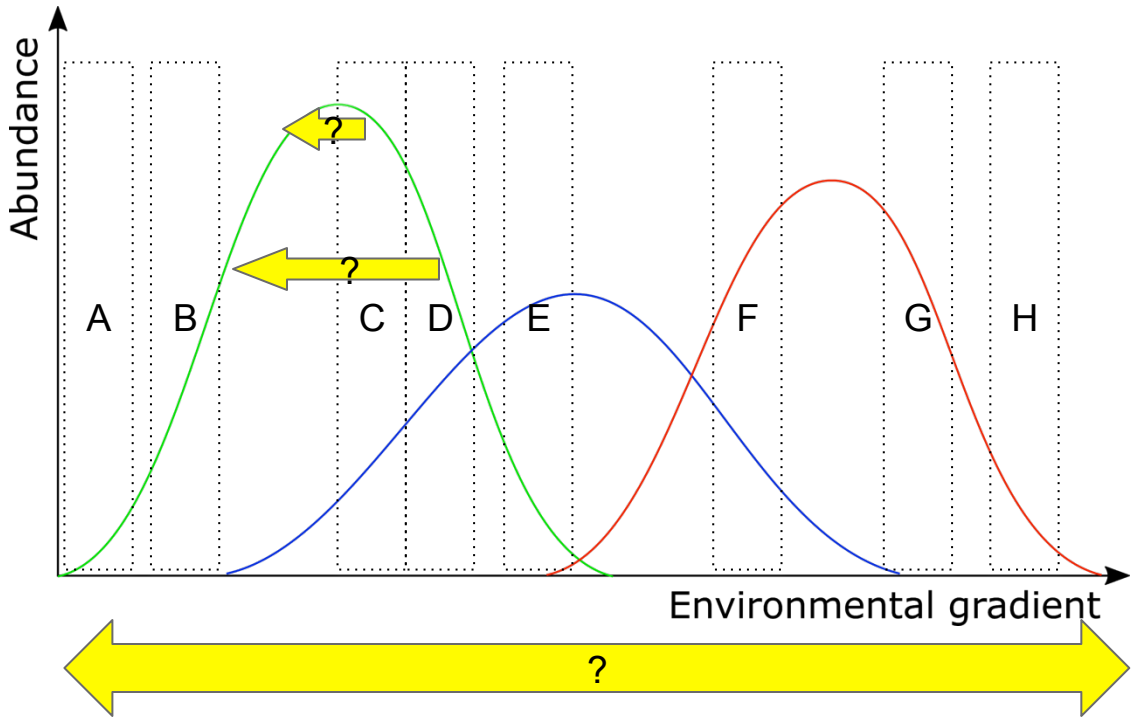


Figure: ourcodingclub.github.io

Prelude & Motivation

Exploratory

- Indirect Gradient Analysis
- Unconstrained

Hypothesis-driven

- Direct Gradient Analysis
- Constrained
  - using environmental data
- Hypothesis testing is available
  - many (? most ?) are permutational

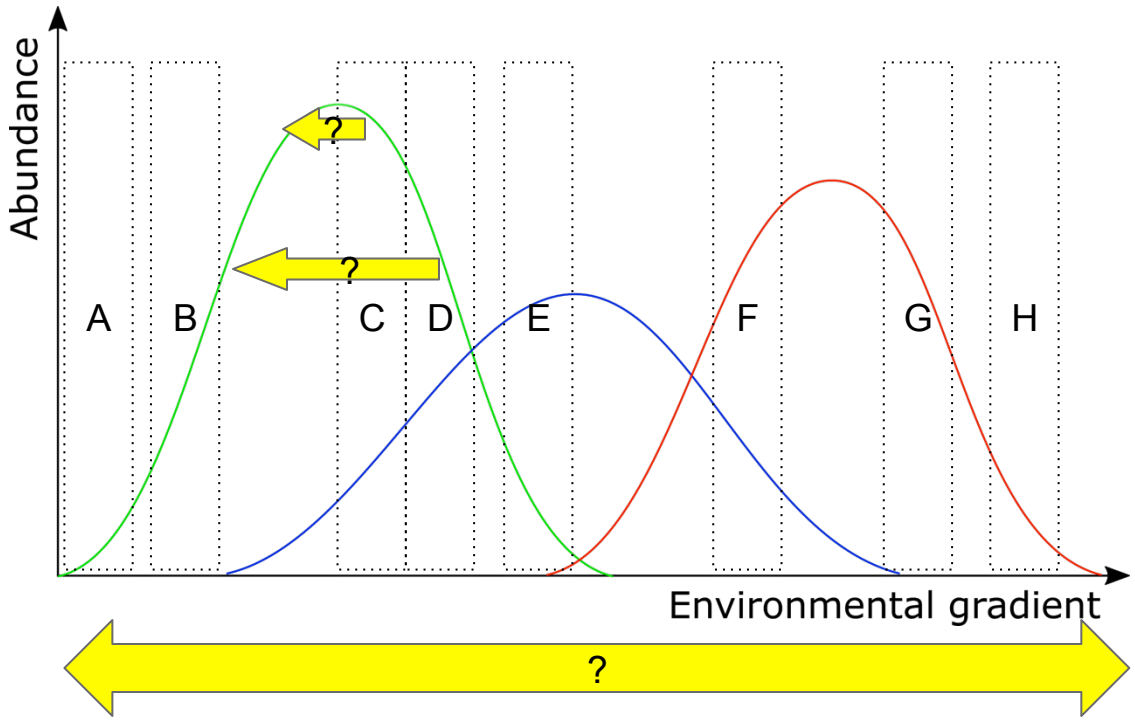


Figure: ourcodingclub.github.io

Prelude & Motivation

Outline

- Data structures
- Transformations & Distances
- Ordination
- Clustering
- Hypothesis testing
  - PERMANOVA
  - ANOSIM
  - Differential abundance testing

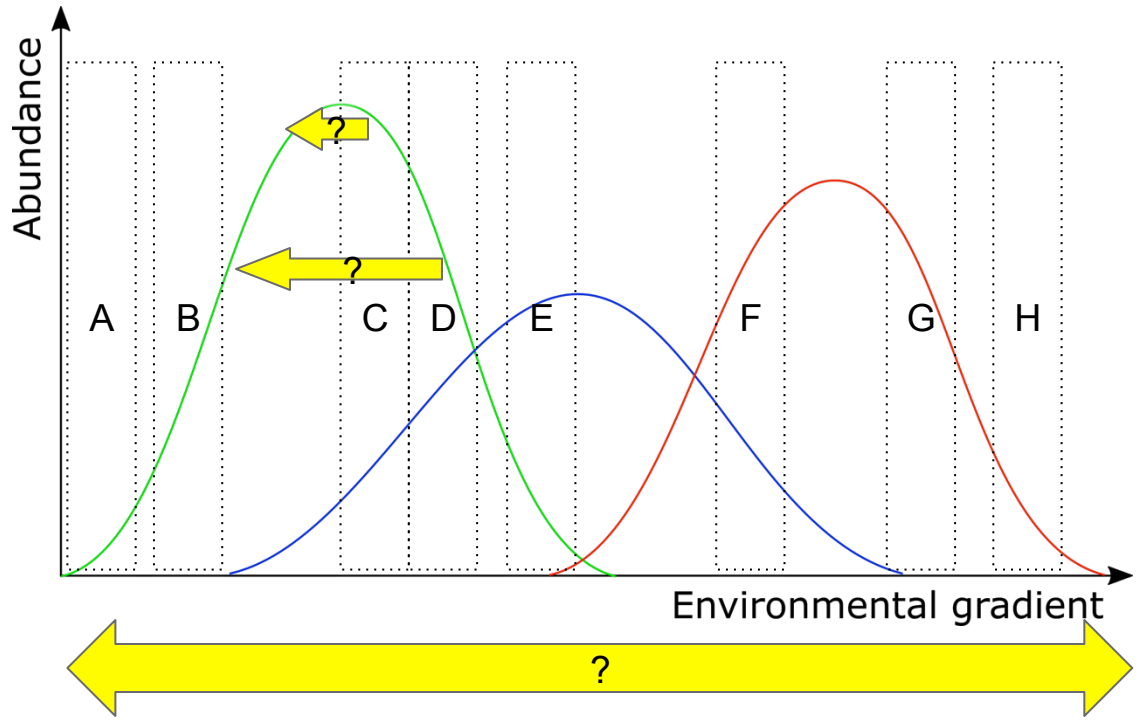


Figure: ourcodingclub.github.io

Data Structures

OTU data table

OTU ID	Taxonomy
1	species A
2	species A
3	species B
4	species C

Abundance Matrix

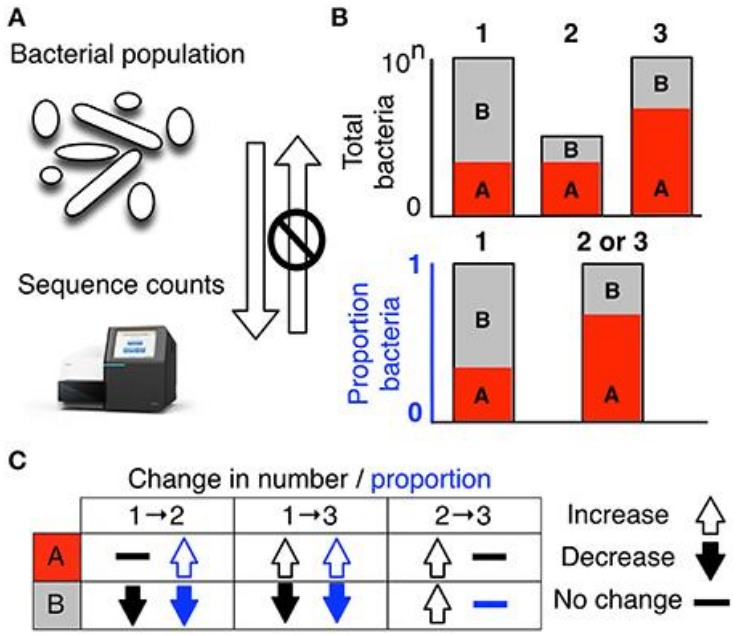
OTU ID	Sample 1	Sample 2	Sample 3
1	0	3	0
2	43	0	2
3	153	23	8
4	0	0	0

Sample data table

Sample ID	Treatment	Host
Sample 1	treated	A
Sample 2	treated	B
Sample 3	control	A
Sample 4	control	B

Challenges:

1. High-dimensional
2. Sparse
3. Data is compositional → → →
4. Variation in (per sample) sequencing depth
5. Individual OTU counts are highly variable, both in scale and variance



Gloor et. al. (2017) DOI: 10.3389/fmicb.2017.02224

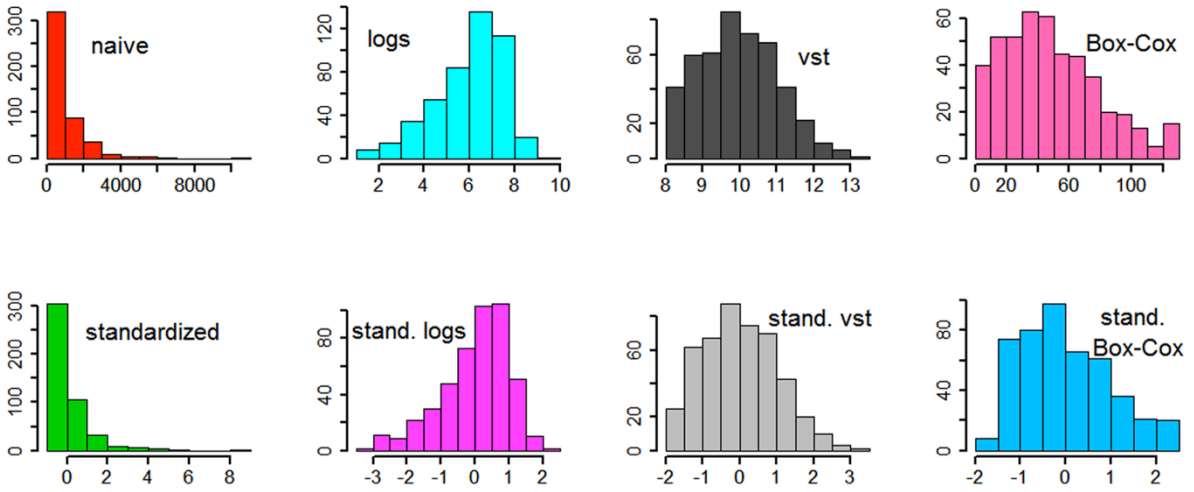
Transformations

Statistically Motivated

1. Normalizing transformations
    - reduce the effects of skew, kurtosis
2. Standardizing
    - places all descriptors on the same scale

Ecologically Motivated

1. Hellinger
    a. de-emphasizes the effects of low abundance species
2. Chi-square
    a. de-emphasizes the effects of high abundance species



Distance Measures

Distance measures quantify how similar or dissimilar two objects (samples) are.

	Categorical	Phylogenetic
Presence/ Absence	**Jaccard**	UniFrac
Quantitative abundance	Bray-Curtis	Weighted UniFrac

Others:
Manhattan
Euclidian
Canberra
Jaccard
Chao
Morisita-Horn
Shannon
...

For two objects a, b, c	Metric	Semimetric	Nonmetric
Minimum Distance = 0 if (a == b) D(a, b) = 0	X	X	X
Positive Distance if (a != b) D(a, b) > 0	X	X	
Symmetry D(a, b) = D(b, a)	X	X	X
Triangle Inequality D(a, b) + D(b, c) >= D(a, c)	X		

Allow the statistical method (and it's limitations) to guide selection

Ordination

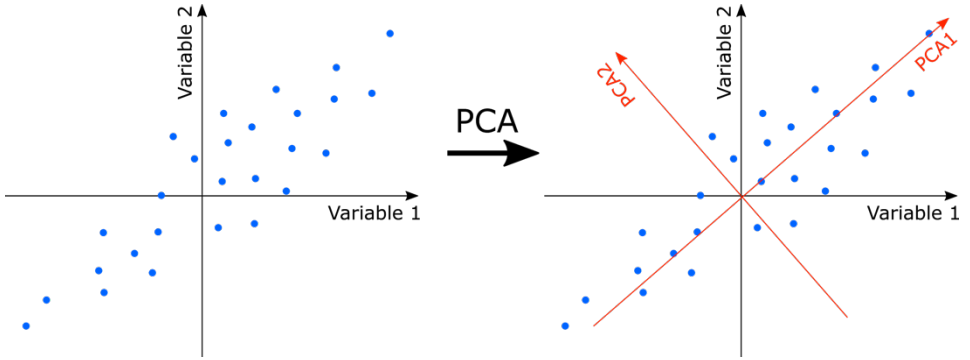
Identify relationships between species abundances and samples/sites

<u>Indirect Gradient Analysis ⇔ Exploratory</u>

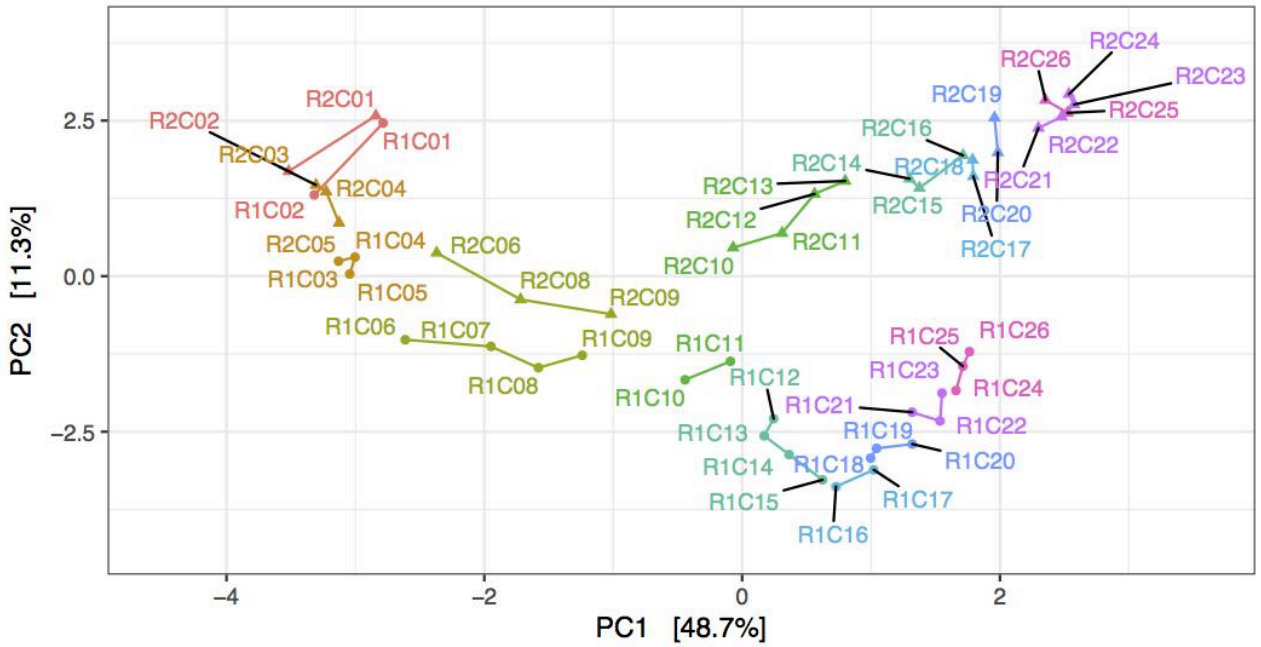
- Unconstrained Ordination
- Maximize the ecological gradient identification
- Analog: Scatterplot

<u>Direct Gradient Analysis ⇔ Hypothesis Driven</u>

- Constrained Ordination
- Species responses are limited to combinations of predicting variables
- Analog: Multivariate Multiple Regression
- Significance test:
  - permutation-based ANOVA → vegan::anova()



ourcodingclub.github.io

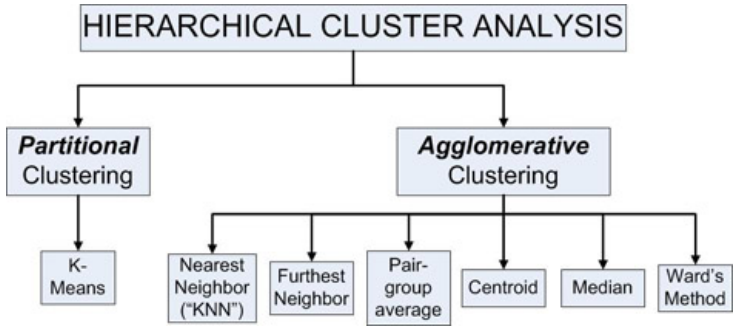


Price & Sales (2019) In Purgatory

Clustering

Identifying groups of objects

- Partitional
- Agglomerative

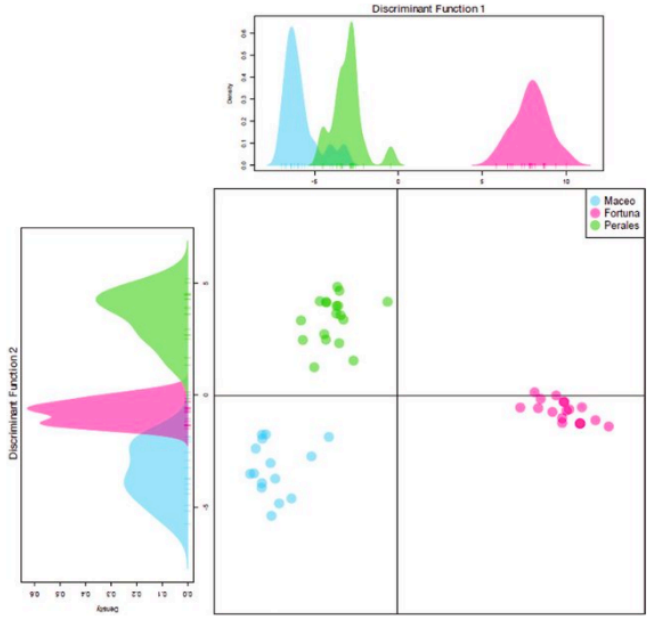


eigenvector.com

Unconstrained ⇔ Exploratory

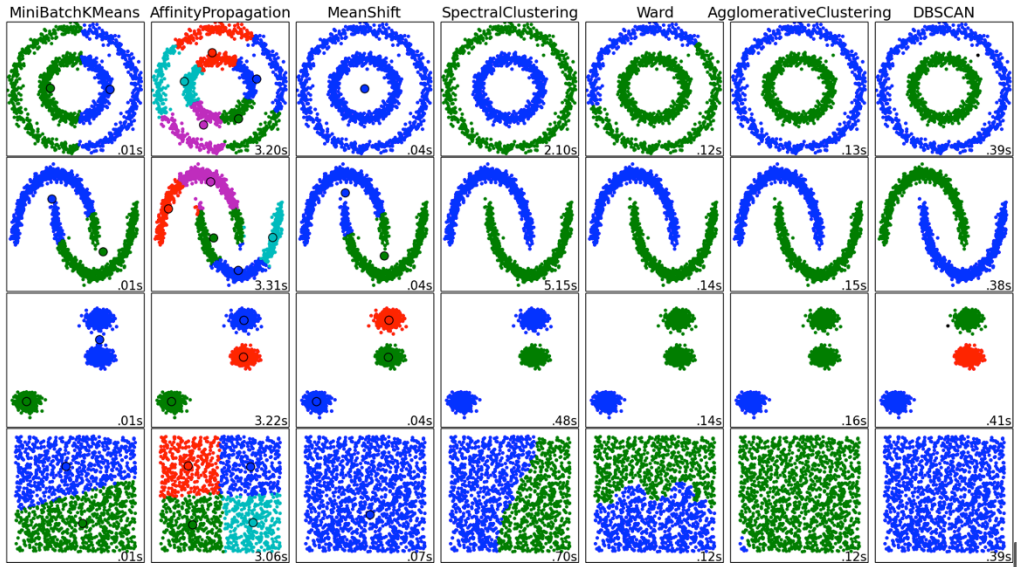
Constrained ⇔ Hypothesis Driven

- Spatial
- Temporal
- mvpart::mvpart()



* Aristizabal et. al. (2018) DOI:10.1093/jhered/esx110

* NOT the result of clustering. Used for illustrative purposes.



scikit-learn.org

Hypothesis Testing

Warnings

- observational studies cannot imply causation
- avoid data dredging
- be aware of highly correlated variables
- beware overfitting of model
- correct for multiple testing

Example Tests

- PERMANOVA
- ANOSIM
- Differential abundance testing

PERMANOVA - PERmutational Multivariate ANalysis of VAriance

Test whether there is a significant difference between two or more groups.

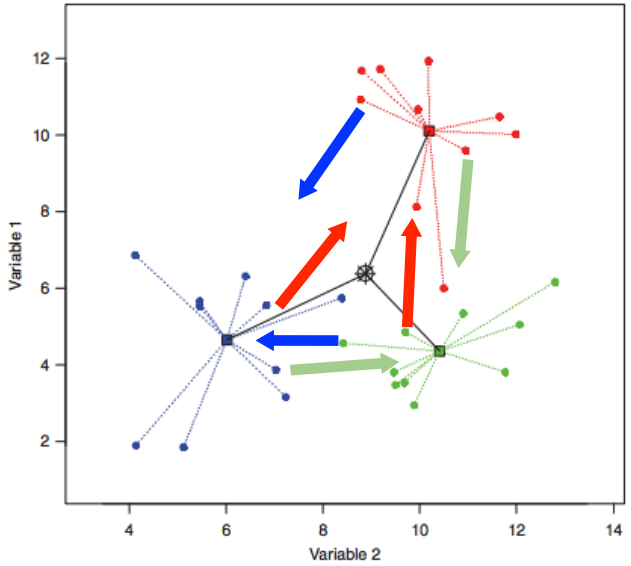
Input: distance matrix

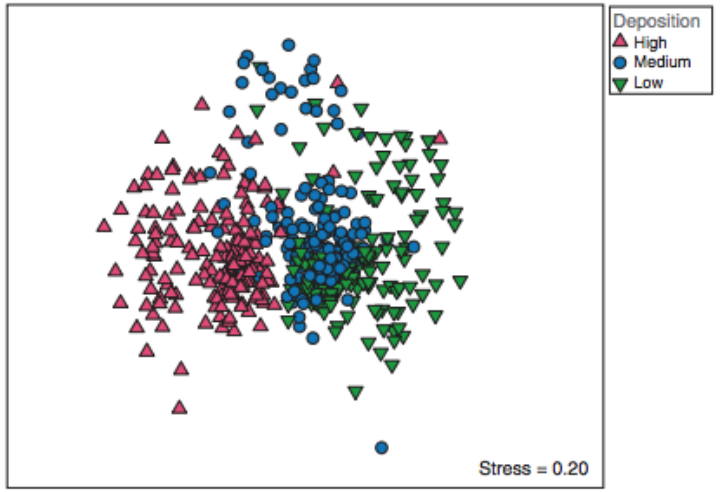
Pseudo F-statistic $$F = \frac{SS_A \div (a - 1)}{SS_W \div (N - a)}$$

Statistical significance is obtained via permutation of objects within the dataset.

Warnings:

- very sensitive to permutational scheme.





All figures adapted from: Anderson, M.J. (2017) DOI: 10.1002/9781118445112.stat07841

ANOSIM - Analysis of Similarities

Test whether there is a significant difference between two or more groups.

Input: distance matrix

- Calculated on ranked Distances

Test Statistic → R

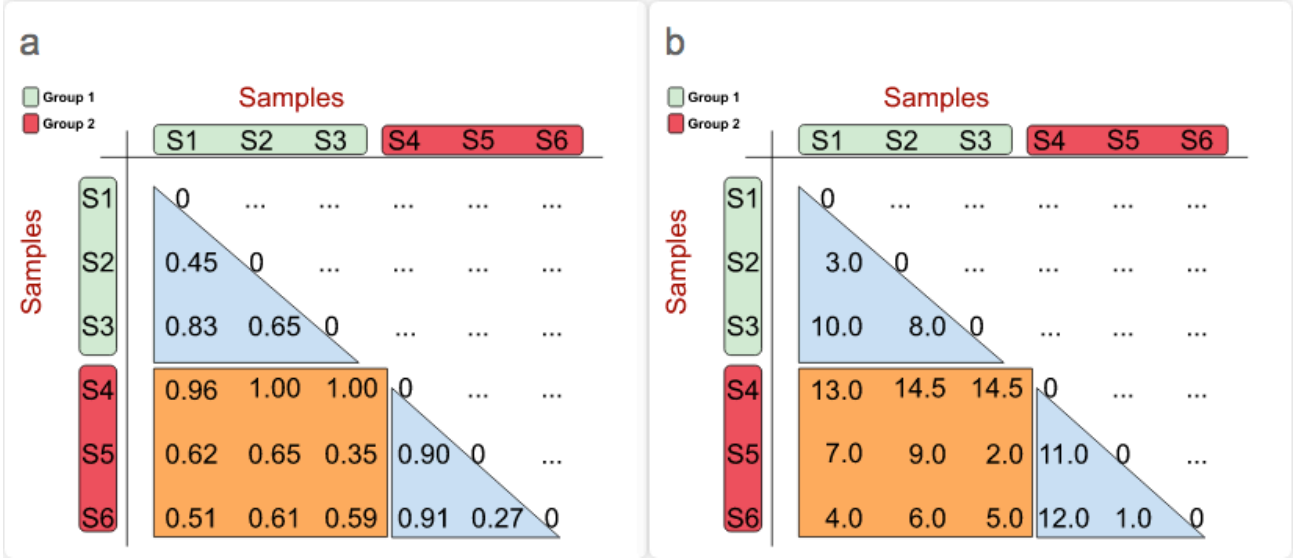
$$R = \frac{\bar{r}_A - \bar{r}_W}{M/2}$$

- relative within-group similarity.
- range [-1, 1]
- interpreted similarly to correlation coefficient

Statistical significance is obtained via permutation of objects within the dataset.

Warnings:

- ANOSIM provides more output than PERMANOVA
  - but PERMANOVA is more robust



Buttigieg & Ramette (2014) DOI: 10.1111/1574-6941.12437

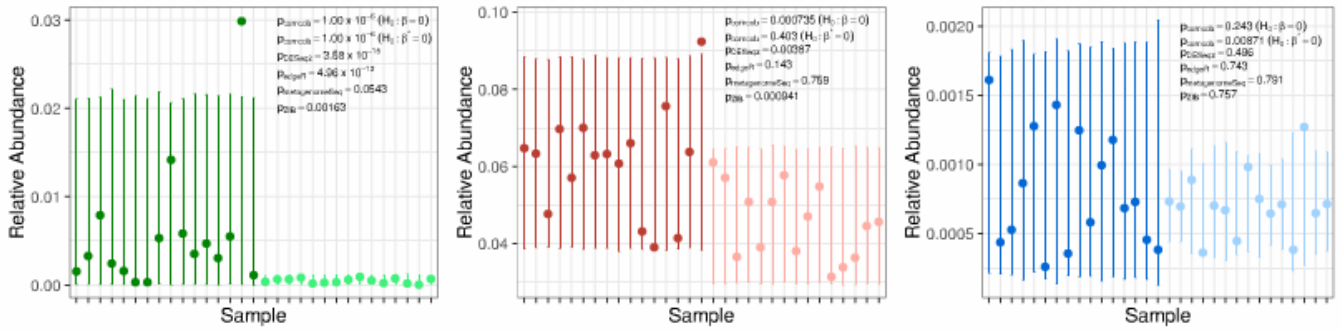
Differential Abundance Testing

Tests whether OTU abundances are different between two or more groups.

	Differential	
	Abundance	Variability
CORNCOB	X	X
DESeq2	X	
EdgeR	X	
metagenomeSeq	X	

Input and statistics depend upon precise method

- Must account for multiple testing



CORNCOB detects taxa that are :

- Diff abund and diff in variance
- Diff abund but no diff in variance
- Not diff abund but diff in variance

Highly useful in identifying dysbiosis

Martin et. al. (2019) arXiv:1902.02776

Thank-you

Jacob Price, PhD
Maria Sevillano, PhD student

@Jake_in_the_Lab @katty_sevi

AEESP | Association of Environmental Engineering & Science Professors