Multivariate statistics for microbial community analysis

Dr Christopher Quince

NERC Metapop 2014

Introduction

- Multivariate stats are applicable to any high dimensional matrix ***X*** with elements x_{ij}
- *N* samples *i* are rows and *S* columns *j* are taxa, OTUs, SEED subsystems, COGS, …
- Now we want to answer the hypotheses that your sampling was designed to address

Multivariate Statistics for Community Comparisons

- Multivariate statistics - machine learning
  - clustering (discrete unsupervised)
  - classification (discrete supervised)
  - ordination (continuous unsupervised)
  - constrained ordination or regression (continuous supervised)

Multivariate Analyses in Microbial Ecology (Ramette FEMS Microbiol. Ecol. 62 2007)

- Distinction between exploratory (unsupervised):
  - Principal component analysis
  - Cluster analysis
  - Non-metric multidimensional scaling
- and hypothesis driven techniques (supervised):
  - Redundancy analysis
  - Canonical correspondence analysis
  - Permutation multivariate analysis of variance
  - Mantel tests
- Microbial ecologists prefer the former

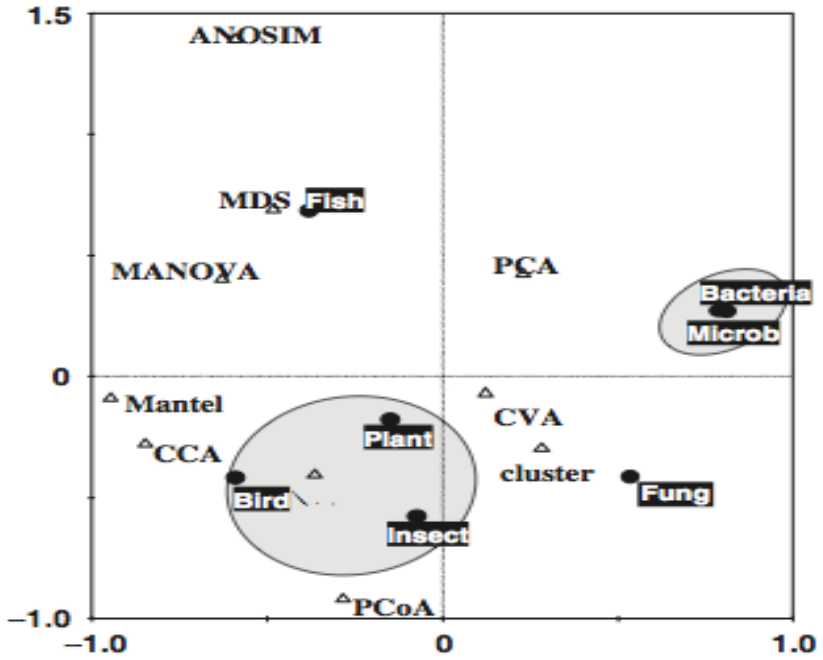


Table 1. Usage (%) of multivariate methods in different fields

	Exploratory analysis				Hypothesis-driven analysis						
Keywords†	Cluster	PCA	MDS	PCoA	CCA	RDA	MANOVA	Mantel	ANOSIM	CVA	Total number‡
Bacter*	48.5	38	4.5	0.4	3.2	1.8	1.3	0.4	0.9	1.1	1141
Microb*	45.8	40.2	3.9	1.1	2.2	2.2	1.1	1.7	0.6	1.1	179
Plant*	40.3	28.5	4.6	1.7	15.5	3.7	1.9	2.3	0.6	0.9	3335
Fung*	54	27.2	2.8	1.1	8.5	2.8	0.9	1.1	0.2	1.4	563
Fish*	30.1	33.7	9.8	0.3	13.5	2.7	3.6	2.9	2.3	1.2	1464
Bird*	41	20.5	5.4	0.7	21.2	3.5	2.1	4.2	0.5	0.9	429
Insect*	54.3	13.7	6.1	0.8	11.5	4.4	3.5	3	1.1	1.7	637

Overview

- Data pre-processing
- Apha diversity
- Beta diversity
  - Clustering
  - Classification
  - Ordinations
- Linking specific taxon to the environment

Data preprocessing

- Data matrix ***X*** with elements x_{ij} (*N* samples *i* are rows and *S* columns)

  + distance measure between variables possibly derived from tree or graph

- Preprocessing:
  - Remove low frequency OTUs or low abundance phyla or genes (MultiCoLA – Gobet et al. 2010)
  - Subsample so that each data set is the same size x_{i+}:

    or use relative abundances $y_{ij}=x_{ij}/x_{i+}$:
  - Standardization to dimensionless z scores
  - Log transform
  - Further transformations (Legendre and Gallagher Oecologia 2001)

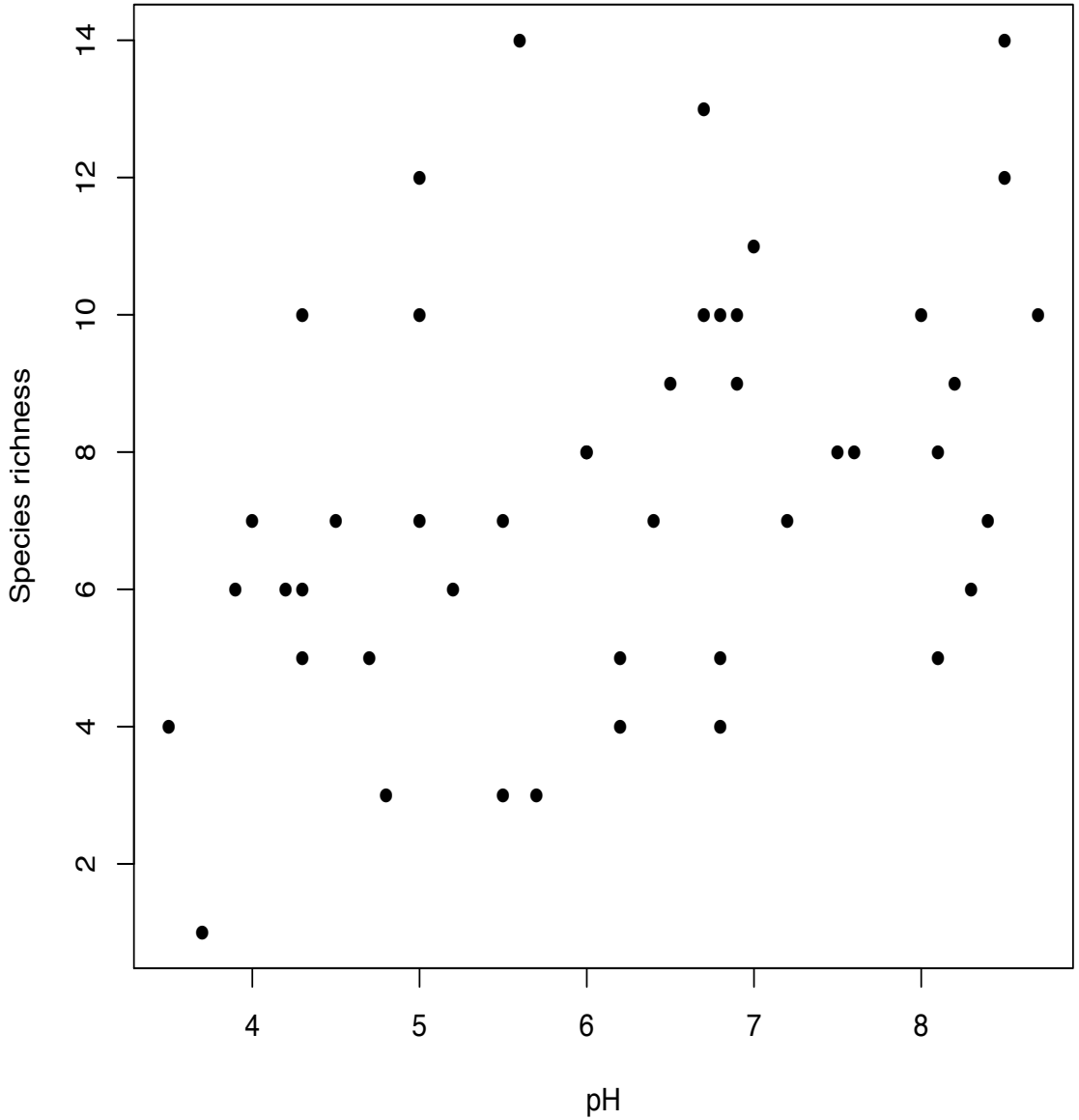
$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

Examples

- Example data set comprising archaeal amoA gene from 46 soils “Niche specialization of terrestrial archaeal ammonia oxidizers “ (Gubry-Rangin et al. PNAS 2011)

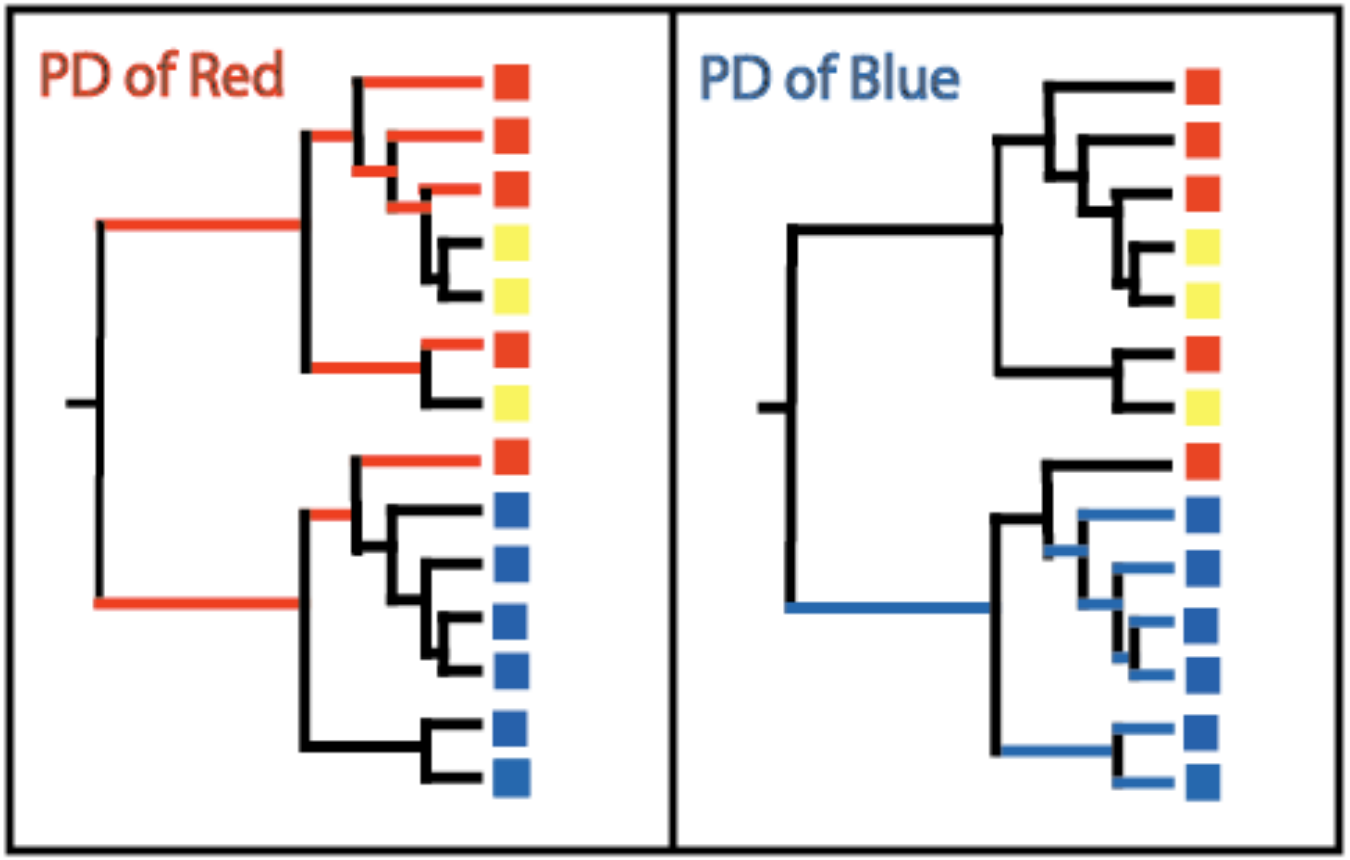
Diversity

- How many different types of object are observed in a sample
  - richness
  - Simpson index
  - Shannon's entropy
  - Renyi entropy (Jost, L. (2006) Entropy and diversity. *Oikos*)
- Subsampling



- p-value = 0.005569
- p-value is the probability of obtaining a test statistic at least as extreme as the observed given the null hypothesis
- Test statistic Pearson's correlation – follows a t-distribution if the data is Gaussian distributed

Phylogenetic Diversity (PD):
a qualitative, phylogenetic α-diversity metric

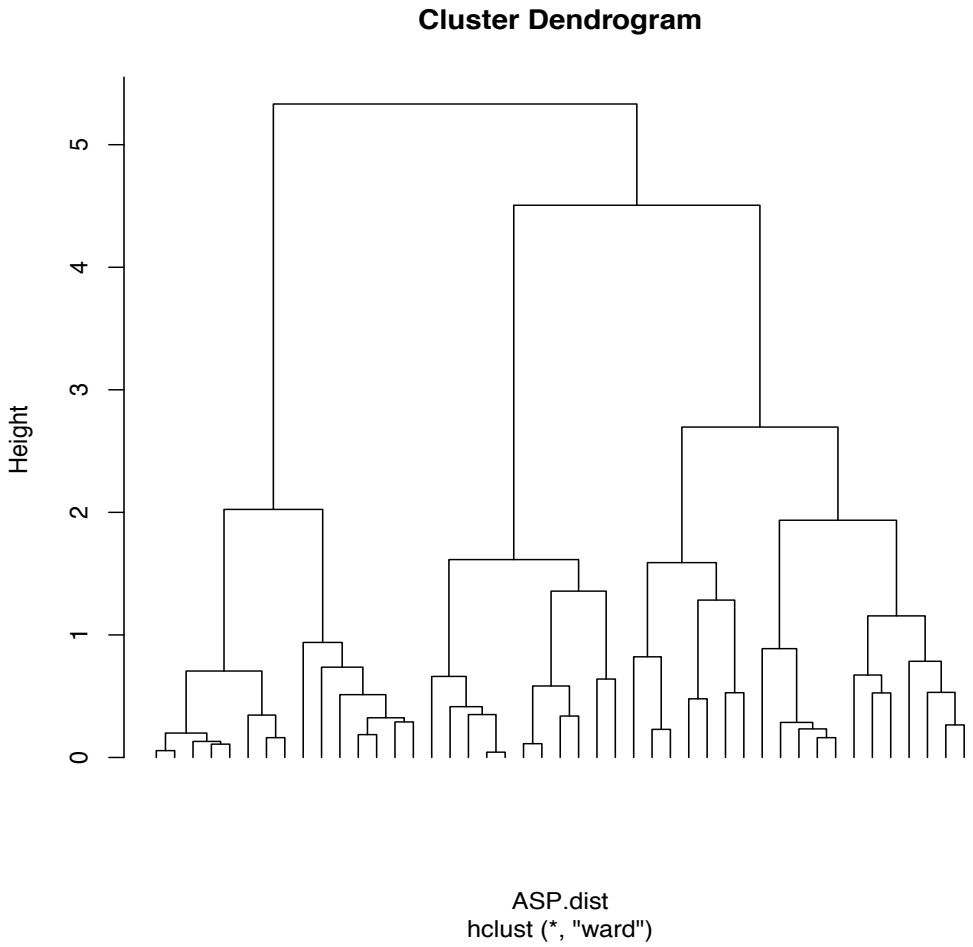


Sum of branch length covered by a sample

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation. 61:1-10.

Cluster analysis

- Find natural groupings in data
  - hierarchical or partitional
- R mode analysis – similarities amongst variables, the taxa
- Q mode analysis – similarities amongst the samples
- Key is the choice of ecological meaningful distances



Distances

R mode - species

- Jaccard

$$d_{lm} = \frac{A + B - 2J}{A + B - J}$$

A is the number of sites where:

$$y_{il} > 0$$

B is the number of sites where:

$$y_{im} > 0$$

J is the number of sites where:

$$y_{il} > 0 \quad \text{and} \quad y_{im} > 0$$

Q mode - taxa

- Euclidean $d_{ij} = \sqrt{\sum_{l=1}^{S} \left(y_{il} - y_{jl} \right)^2}$
- Chi-squared $d_{ij} = \sqrt{y_{++} \left(\sum_{l=1}^{S} \frac{1}{y_{+j}} \left[\frac{y_{il}}{y_{i+}} - \frac{y_{jl}}{y_{j+}} \right]^2 \right)}$
- Overlap based (Bray-Curtis):

$$d_{ij} = 1 - \sum_{l=1}^{S} \min(y_{il}, y_{jl})$$

- Hellinger: $d_{ij} = \sqrt{\left(\sum_{l=1}^{S} \left[\sqrt{y_{il}} - \sqrt{y_{jl}} \right]^2 \right)}$

What is wrong with Euclidean distance?

		Species 1	Species 2	Species 3
Species abundance paradox data (three sites, three species)	Site 1	0	1	1
	Site 2	1	0	0
	Site 3	0	4	8
Distance function		*D*(site 1, site 2)	*D*(site 1, site 3)	*D*(site 2, site 3)
$D_{\text{Euclidean}}$		1.7321	7.6158	9.0000
D_{chord}		1.4142	0.3204	1.4142
$D\chi^2_{\text{metric}}$		1.0382	0.0930	1.0352
$D\chi^2_{\text{distance}}$		4.0208	0.3600	4.0092
$D_{\text{species profiles}}$		1.2247	0.2357	1.2472
$D_{\text{Hellinger}}$		1.4142	0.1697	1.4142

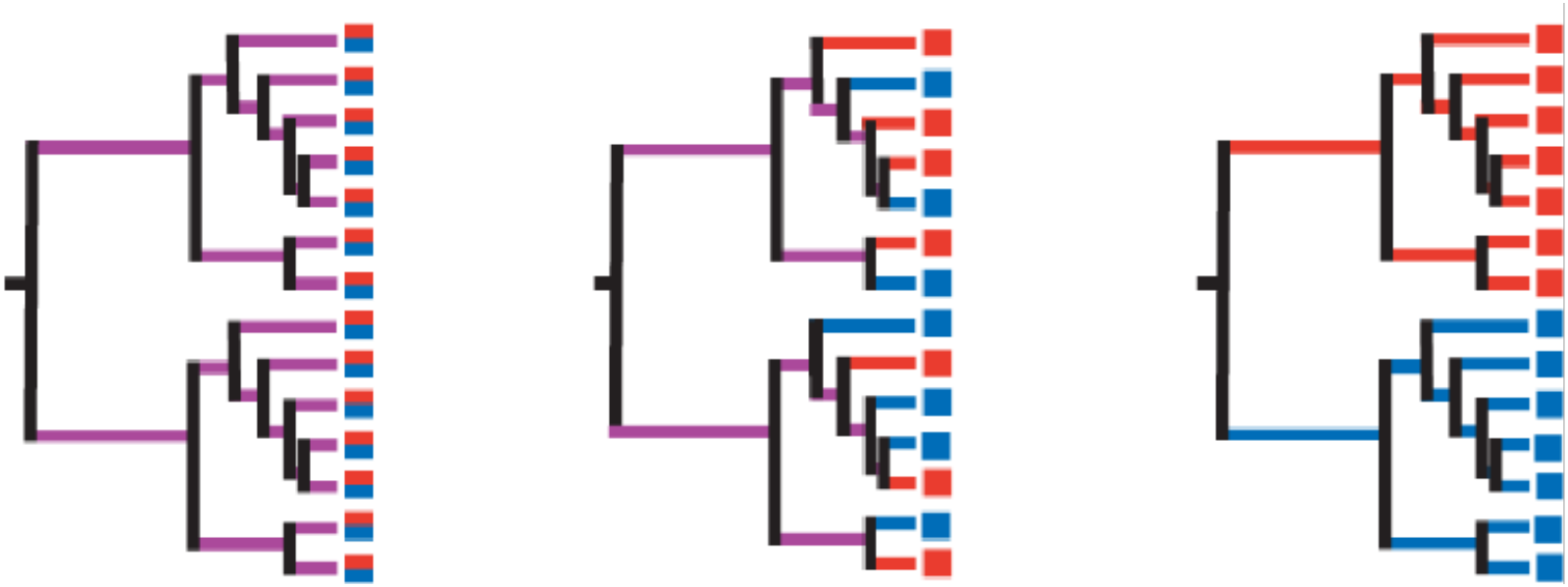
- Transforming data can help (Legendre and Gallagher Oecologia 2001)
- Using relative abundances gives species profiles

UniFrac: a qualitative, phylogenetic β-diversity metric

Identical communities
D = 0.0

Related communities
D ~ 0.5

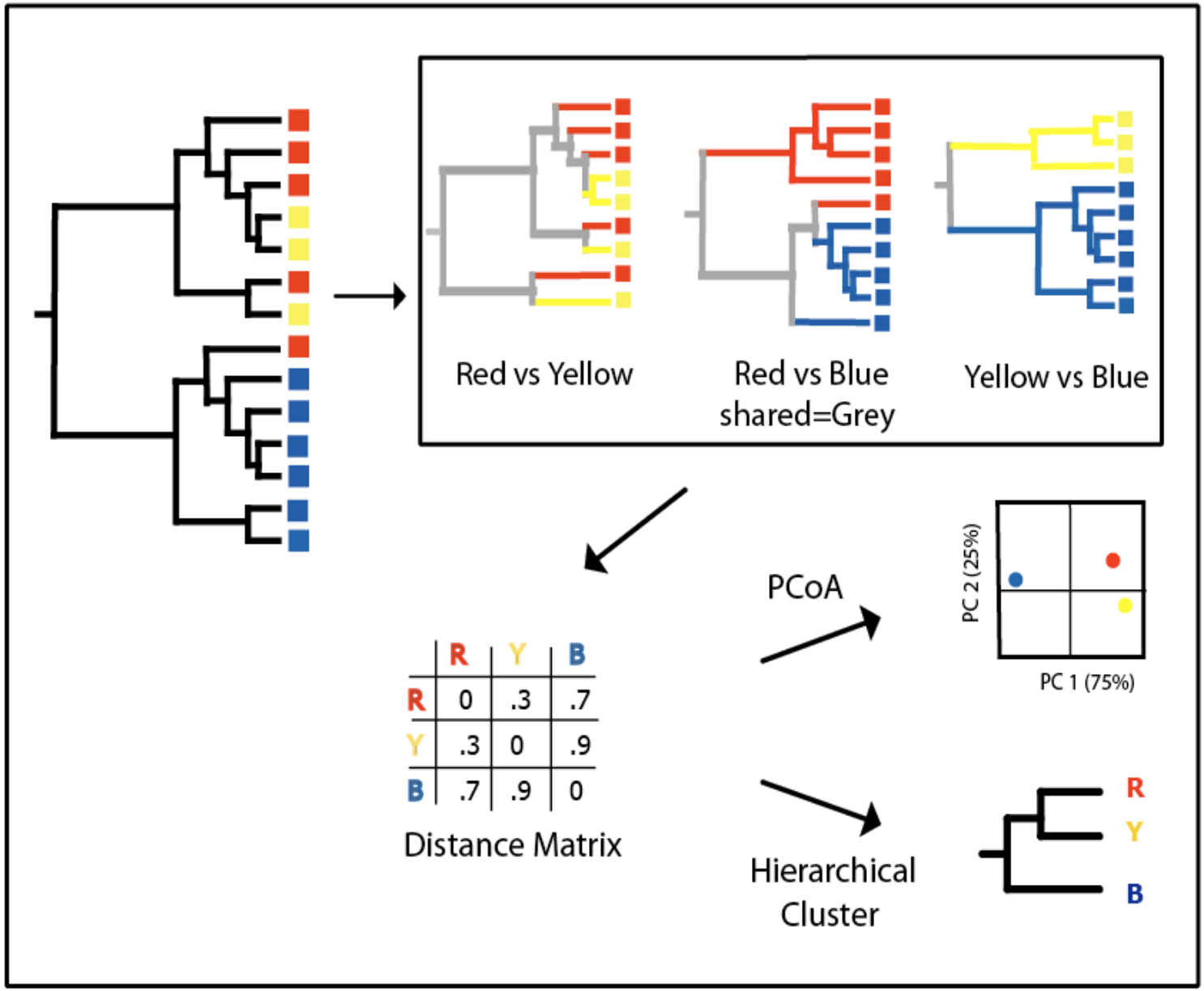
Unrelated communities
D = 1.0



Percent of observed branch length that is unique to either sample

Lozupone and Knight, 2005, Appl Environ Microbiol 71:8228

Clustering by UniFrac distance



Phylogenetically Aware Community Metrics

- Weighted Unifrac is the quantitative equivalent of Unifrac
- Unifrac is not the only choice:
  - comdist: mean pairwise distances separating species in two communities, may or may not be weighted by species abundances
  - phylogenetic community dissimilarity: explicit evolutionary model (Ives and Helmus Am. Nat. 2010)
  - generalised Unifrac (Chen et al. Bioinformatics 2012)

Ordination

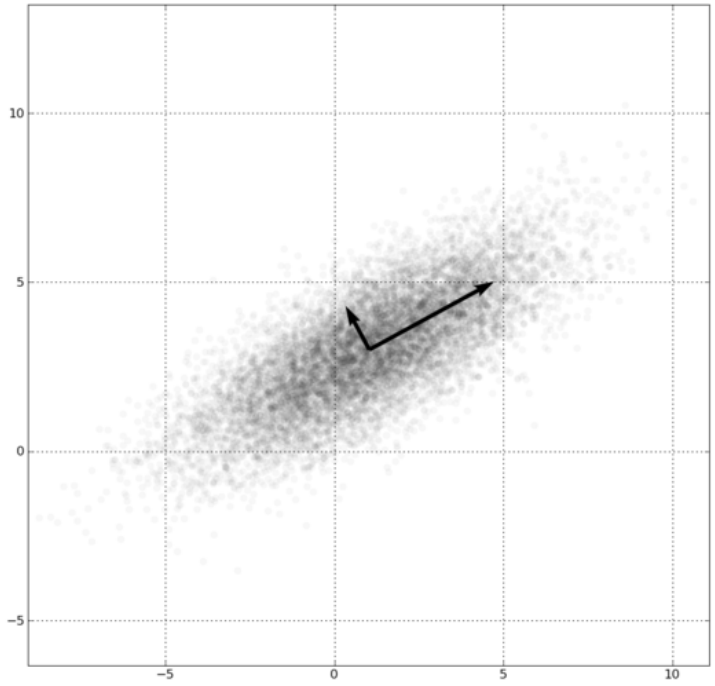
- Clustering is a discrete latent variable approach
- Ordination map samples to continuous latent variables
- Allows representation in a reduced dimension D

Ordination Methods – Principal Components Analysis

- Represent samples in new space by a linear map:

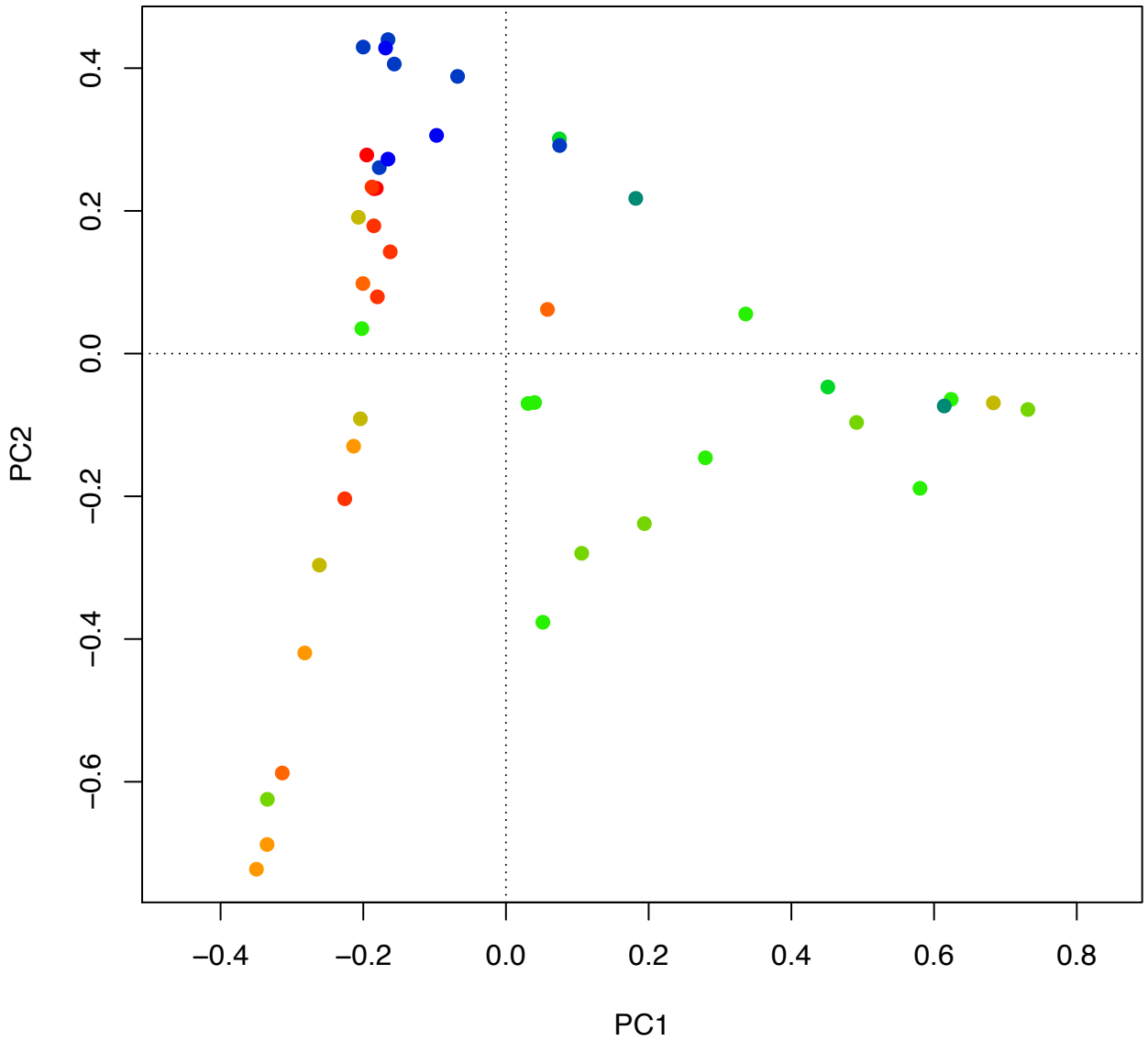
$$\mathbf{z}_i = \mathbf{M}\mathbf{y}_i$$

- Where **M** is a D times S matrix of coefficients
- Principal components does this so as to maximise the variance in the projected data
- Preserves Euclidean distances in the transformed space and assumes Gaussian noise
- Example of an eigenvector methods because found by solving eigenvector equation



$$C\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

Where C is the SXS sample covariance matrix



Eigenvalues are decreasing

Sum of eigenvalues is total variance

Proportion explained by each eigenvector

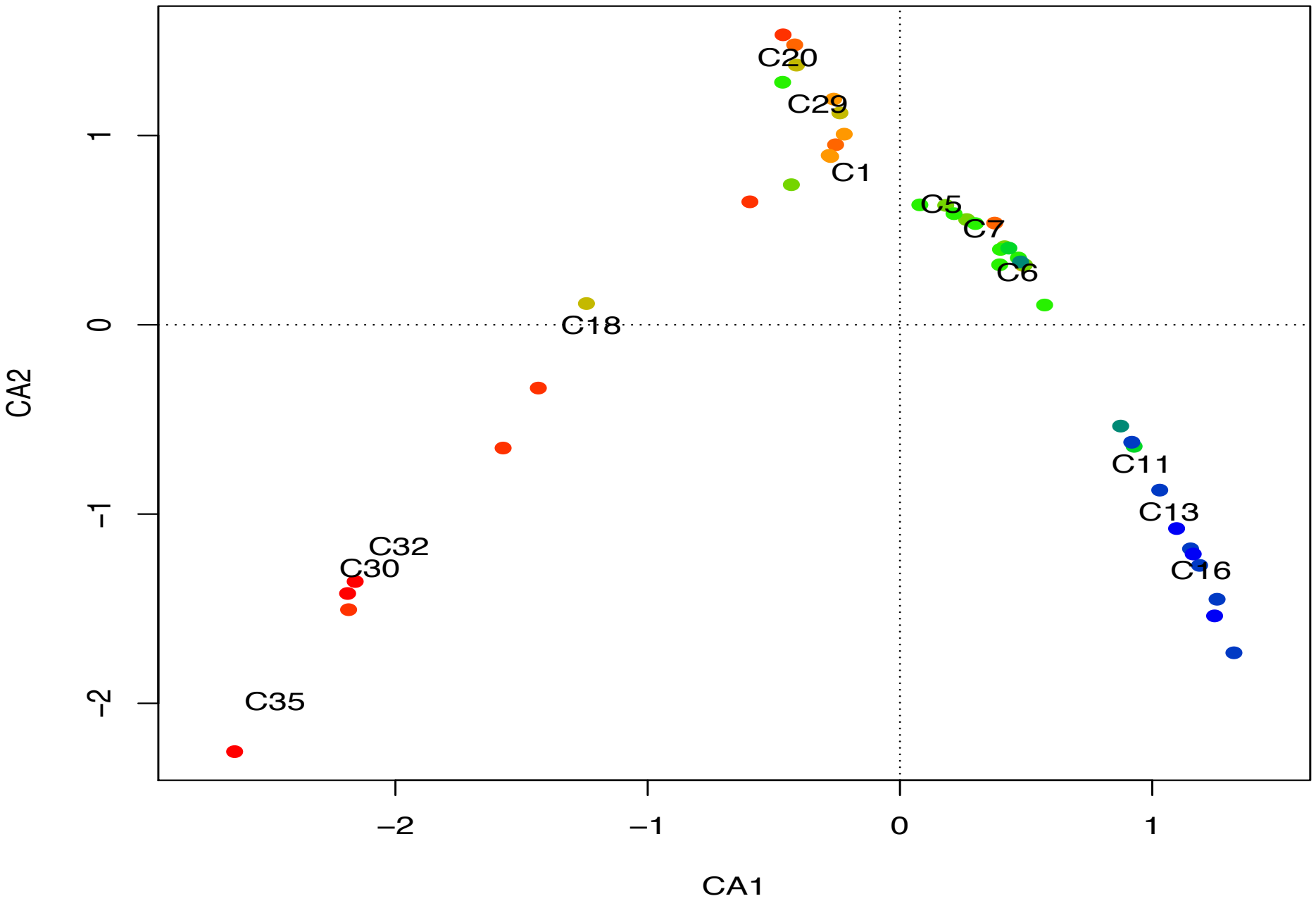
$$\lambda_i / \sum_{i=1}^{N} \lambda_i$$

PCA (cont.)

- Drawback of PCA is the restriction to Euclidean distance but can address this with transformations (Legendre and Gallagher Oecologia 2001)
  - Example: square root transformation gives Hellinger distances

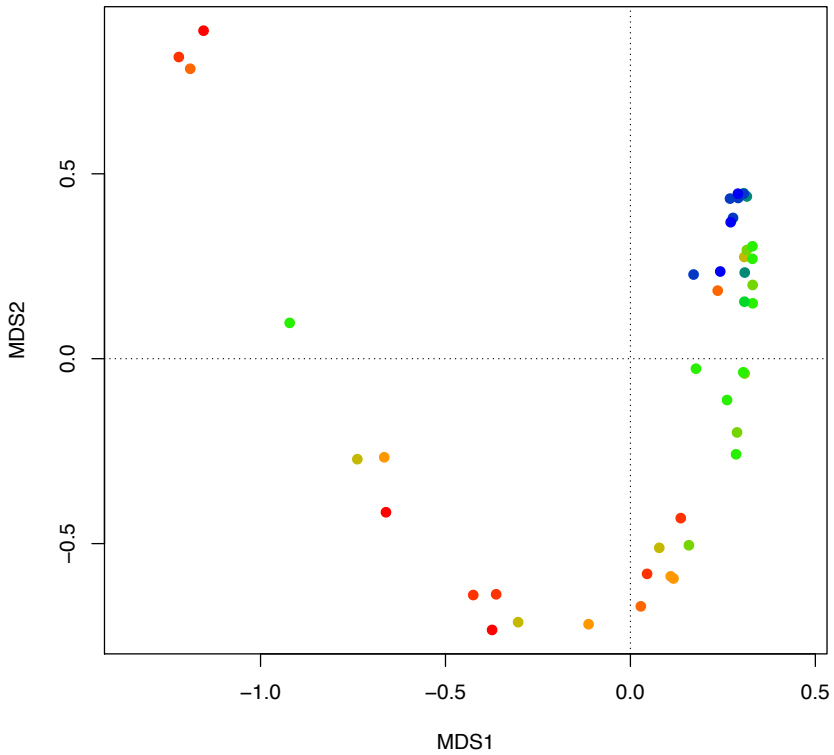
Correspondence Analysis

- Each species has a preferred environment represented by a score
- Each site has a score in the same metric
- Abundance of species in each site is distributed as a Gaussian on that metric with equal standard deviation
- Actually comes down to solving eigenvector equation again
- Unimodal
- Preserves chi-squared distances
- Both PCA and correspondence analysis allow biplots
- Eigenvalues related to amount of variation explained by each transformed dimension


C20
C29
C1
C5
C7
C6
C18
C11
C13
C16
C32
C30
C35
CA1
CA2
−2
−1
0
1
−2
−1
0
1

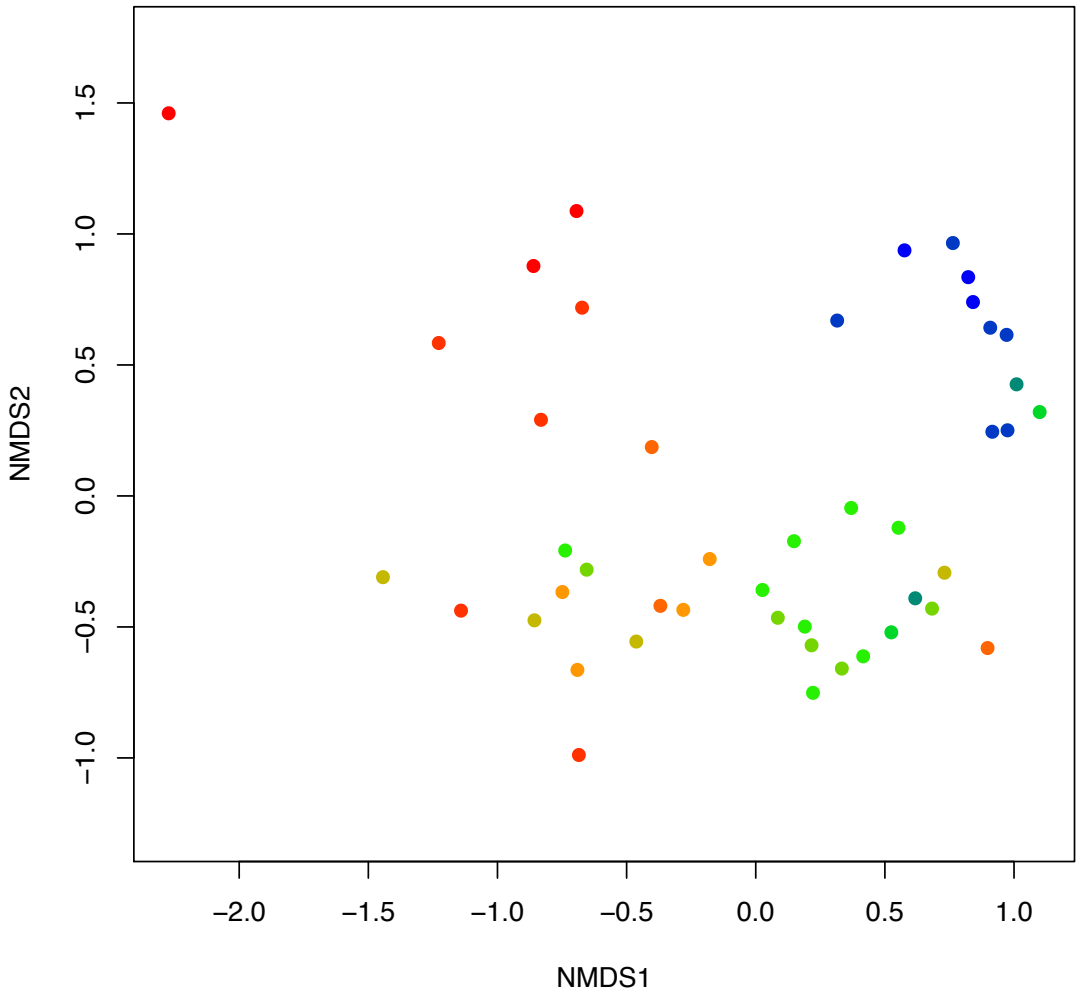
Metric Multi-Dimensional Scaling

- Aim to map samples from *S* dimensional space to lower dimension *D* often just 2
- Any distance can be used in MDS
- Preserve distances in new space
- Non-linear map
- Example – principal coordinates analysis (PCoA)
- Eigenvalue method operating on NXN distance matrix
- Biplots are possible
- Phylogenetic distances e.g. Unifrac can be used (right) – Qiime

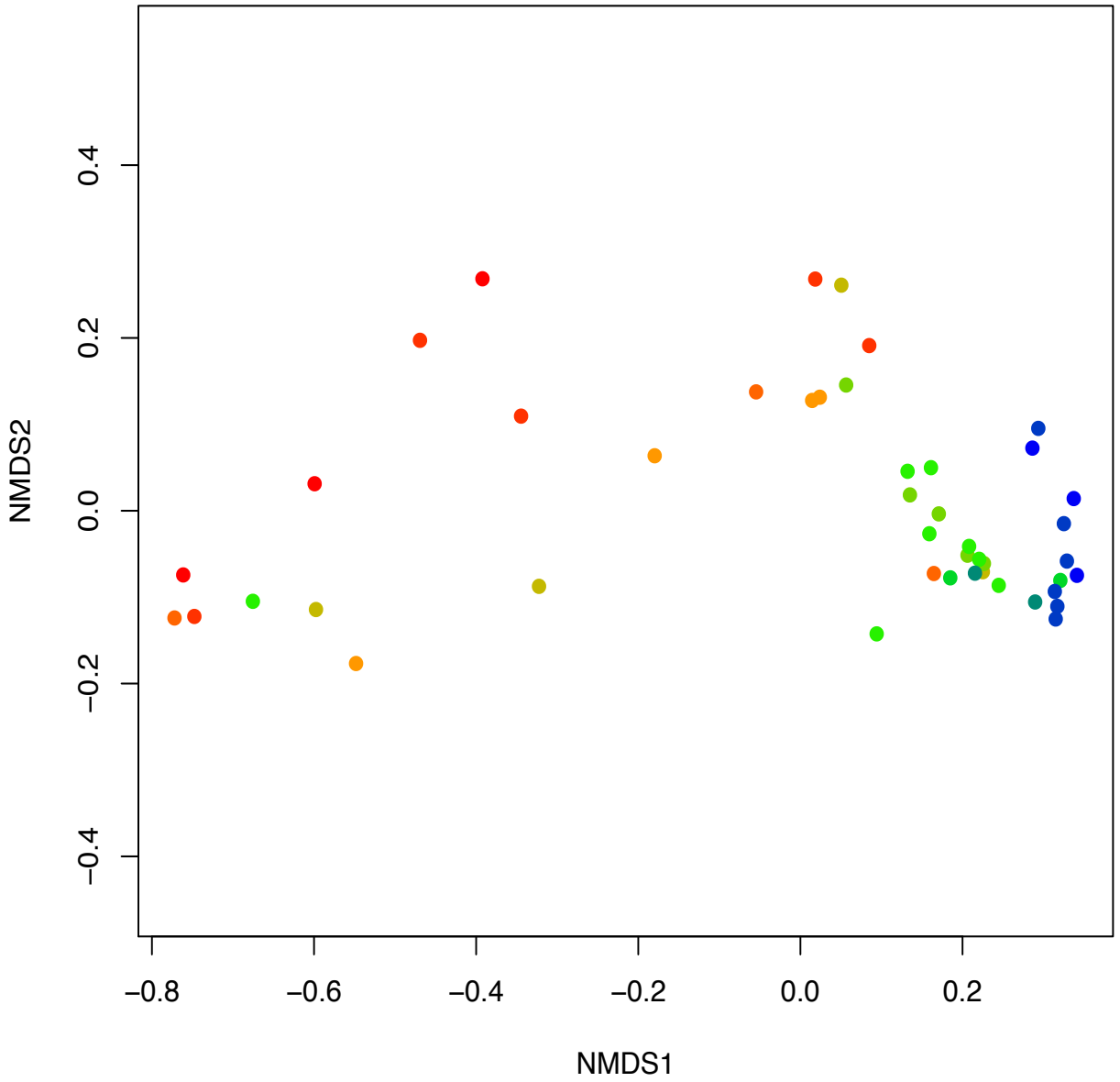


Non-Metric Multidimensional Scaling

- Differs from metric multidimensional scaling in that we only try to preserve order of distances in mapping
- New distances are not equal to old but related monotonically
- Stress is a measure of quality of fit
- Biplots are possible dimensions predetermined



- Phylogenetic distances can be used in NMDS too e.g. comdist

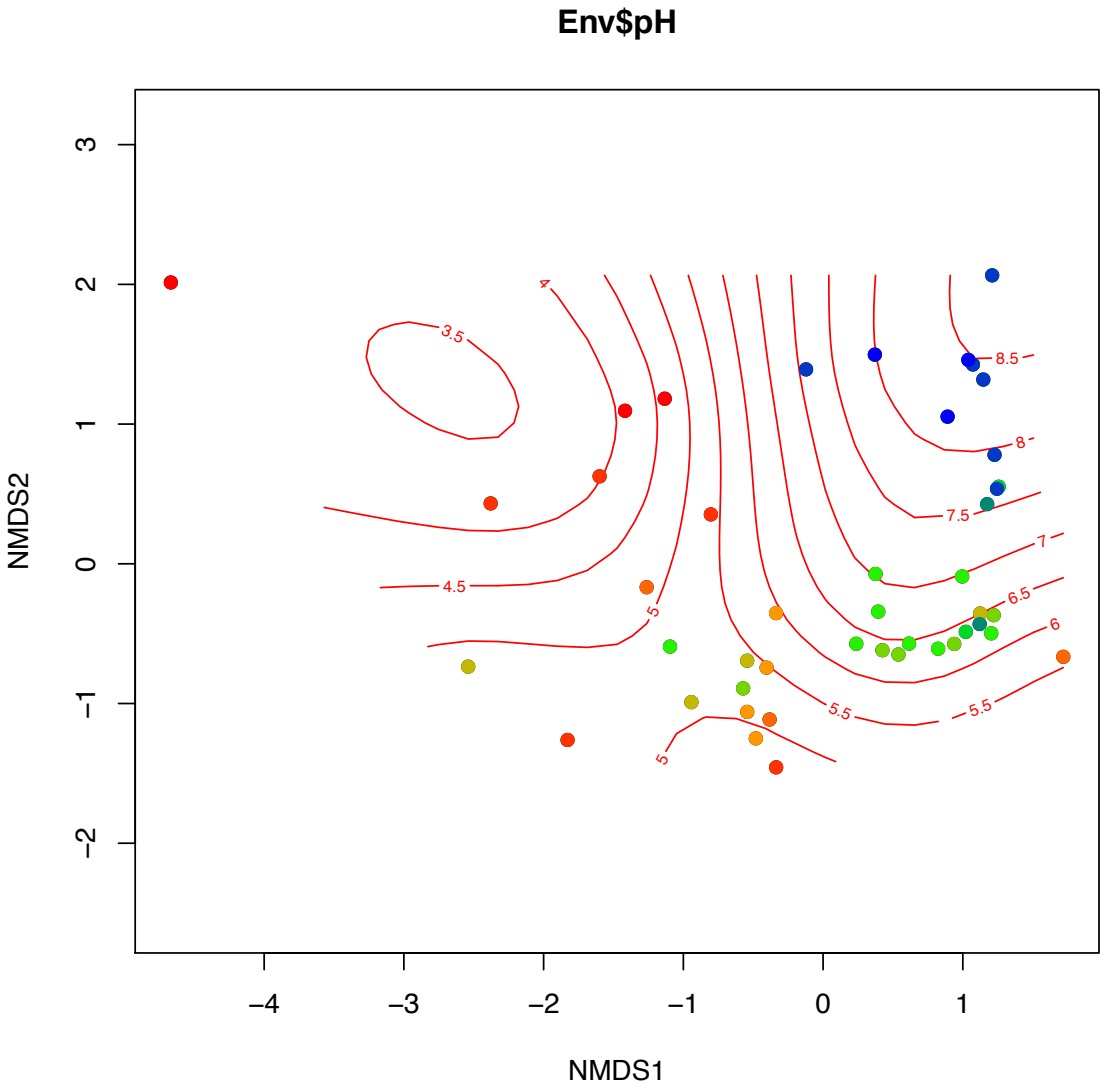


Which ordination method should you use?

- Depends on the range of environments in your data sets
- Over short environmental gradients PCA suffices since species response should be linear
- Over longer ranges then correspondence analysis is preferred because species will appear unimodal
- Use whatever looks better or NMDS instead!
- Use principal coordinates or NMDS for Unifrac

Incorporating environment variables into ordination diagrams

- We can do this indirectly by coloring samples
- We can also plot environmental gradients onto ordination as vectors or contours



Constrained Ordination

- Incorporating environmental variables directly can reveal hidden patterns
- Ordination axes become a linear combination of the environmental variables
  - Constrained PCA is called redundancy analysis
  - Constrained correspondence analysis is called canonical correspondence analysis (CCA
  - Constrained principal coordinates analysis is also possible (Anderson and Willis Ecology 2003) becomes canonical analysis of principal coordinates
- Eigenvalues now give proportion of variance explained by each combination of environmental variables
- Choice between the RDA and CCA should be guided by same rules as in unconstrained case: long gradients imply CCA

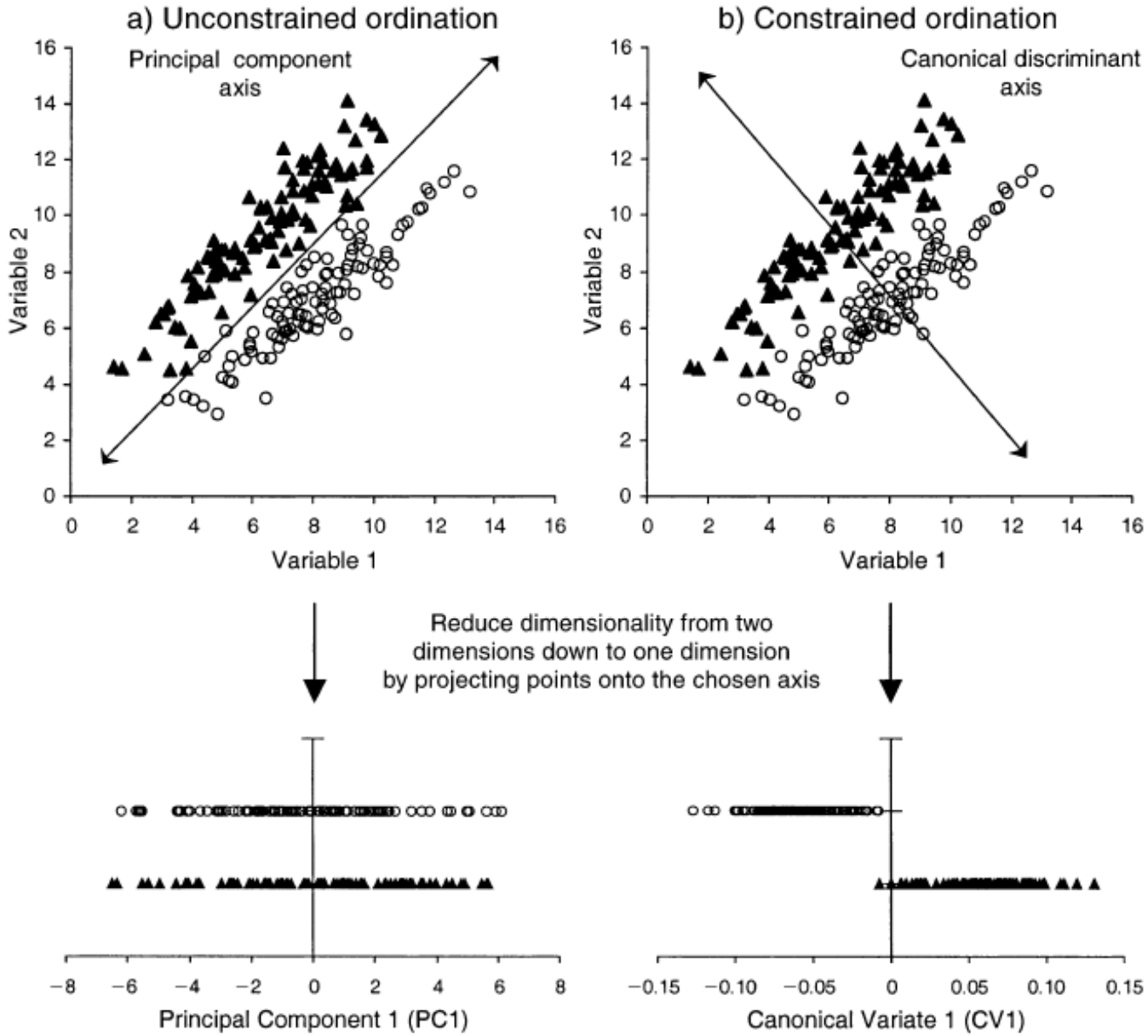
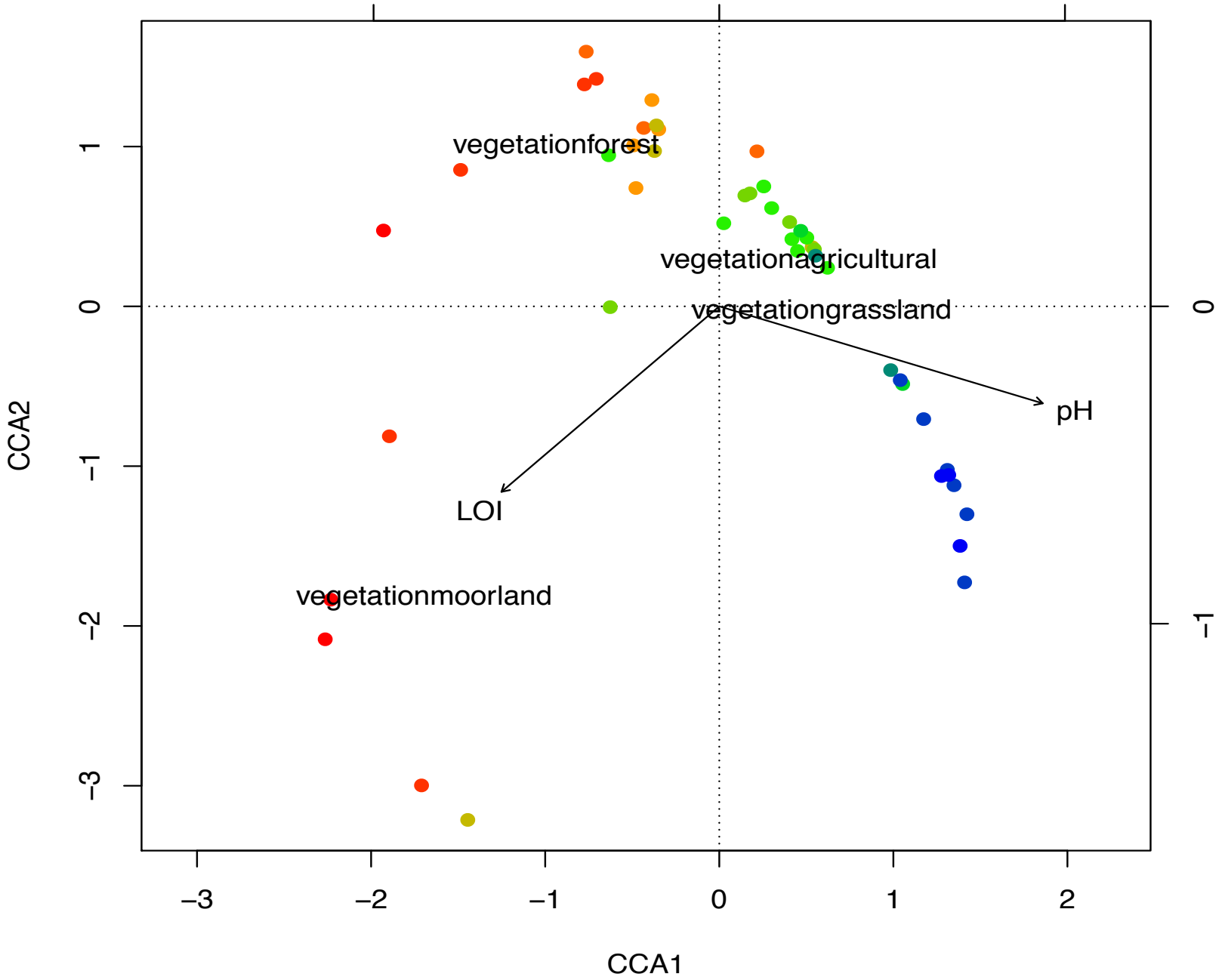


FIG. 1. Visual comparison of the method used to reduce dimensions in (a) an unconstrained and (b) a constrained ordination procedure. Data were simulated from a multivariate normal distribution with the two groups having different centroids (6, 9) and (9, 7), but both variables had a standard deviation of 2, and the correlation between the two variables was 0.9. Note the difference in scale between the first canonical axis (CV1) and the first principal component (PC1).

Anderson and Willis Ecology 2003


vegetationforest
vegetationagricultural
vegetationgrassland
pH
LOI
vegetationmoorland
CCA1
CCA2
−3
−2
−1
0
1
2
−3
−2
−1
0
1
0
−1

Testing significance of constrained ordination variables

- A common question to ask is what variables are significant
- Can apply permutation ANOVA to constrained ordinations
  - Simple idea, permute the identity of the samples whilst keeping the environmental variable matrix fixed
  - Compute a test F-statistic, calculate proportion of permuted matrices with F-statistic smaller than the true value this is the p-value: $$F = \frac{SS_A/(a-1)}{SS_W/(N-a)}$$
- Gives for the CCA that everything but CN is significant

Multivariate Analysis of Variances

- Permutation approach can be applied to any community distance matrix - non-parametric multivariate analysis of variance (Anderson Austral. Ecology 2001, McArdle and Anderson Ecology 2001)
- We can use Bray-curtis or Unifrac for example:

	BC – R2	BC - p	UF – R2	UF -p
pH	0.138	0.001***	0.359	0.001***
C	0.065	0.001***	0.016	0.276
N	0.029	0.056.	0.004	0.693
CN	0.032	0.064.	0.012	0.373
Moisture	0.017	0.420	0.011	0.429
LOI	0.031	0.056	0.014	0.350
vegetation	0.087	0.015*	0.098	0.050*
Residuals	0.598		0.484	

Mantel tests

- This is confirmed by Mantel tests which give correlations between matrices
- Compare Bray-curtis community matrix with environment ($r = 0.3946$, $p < 0.001$)
- Or for Unifrac ($r = 0.2368$, $p < 0.029$)
- What if we want to know which species are responding to the significant environmental variables…

Discrete groups

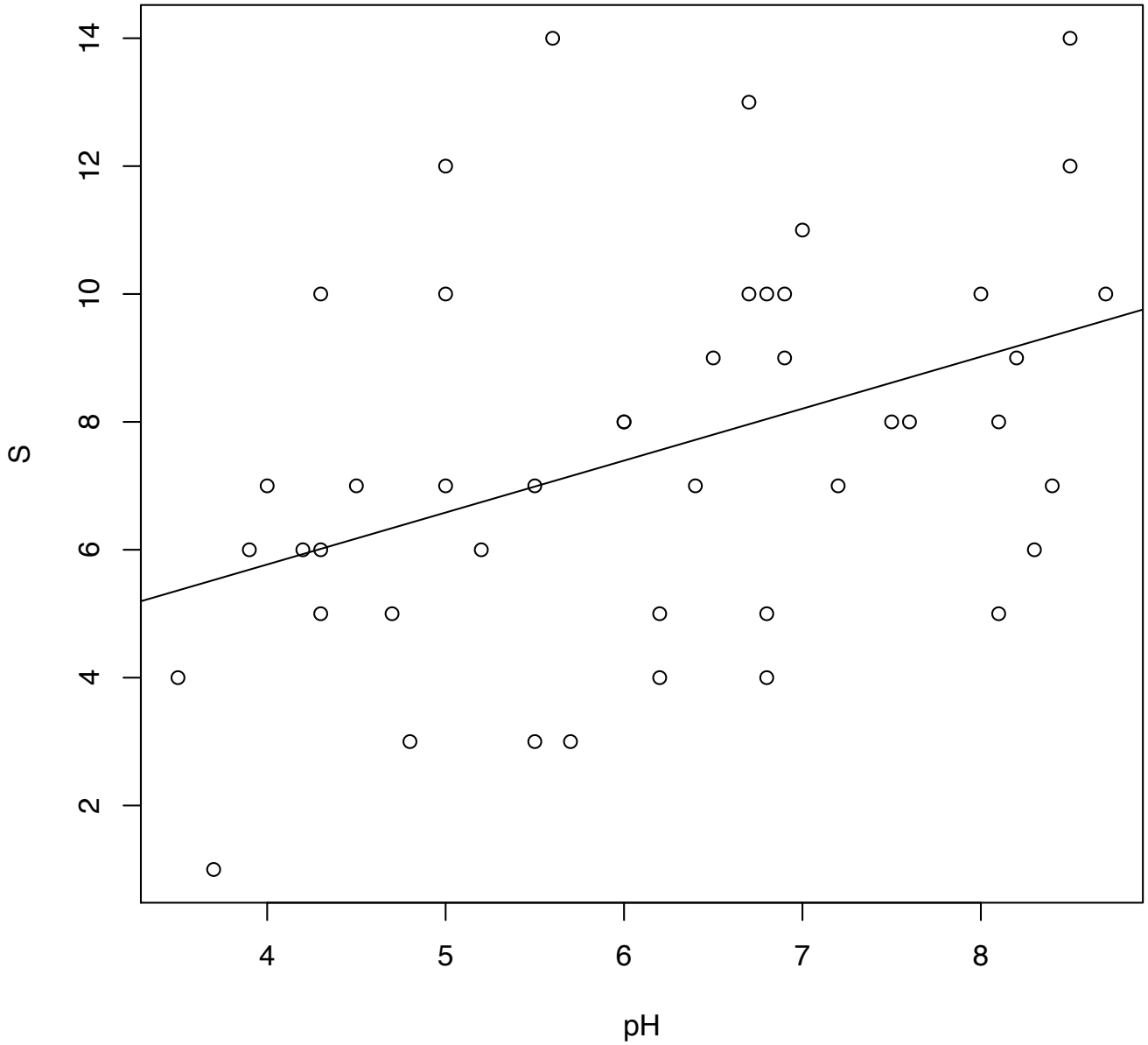
- If you are considering abundance differences between two groups then a t-test can be used (Metastats – White et al. PLoS Comp Biol 2009)
- Multiple groups then ANOVA
- Extended to zero inflated Gaussians – metagenomeSeq (Paulson)
- LDA Effect Size (LEfSe) (Segata et. al Genome Biol. 2010)
- Relaxing independence of species – Dirichlet-multinomial (Bayesian – DirichletMultinomial in bioconductor, hmp in CRAN)

Linear Regression

- Use regression to predict dependent variable as a function of explanatory variables
- Linear regression: dependent variable is a linear function of explanatory
  - Simple: one explanatory variable
  - Multivariate: multiple explanatory variables

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

- Assumes response is a normal distribution



S ~ 2.52 + 0.81.pH
Adjusted R-squared: 0.146
F-statistic: 8.52 on 1 and 43 DF, p-value: 0.005569

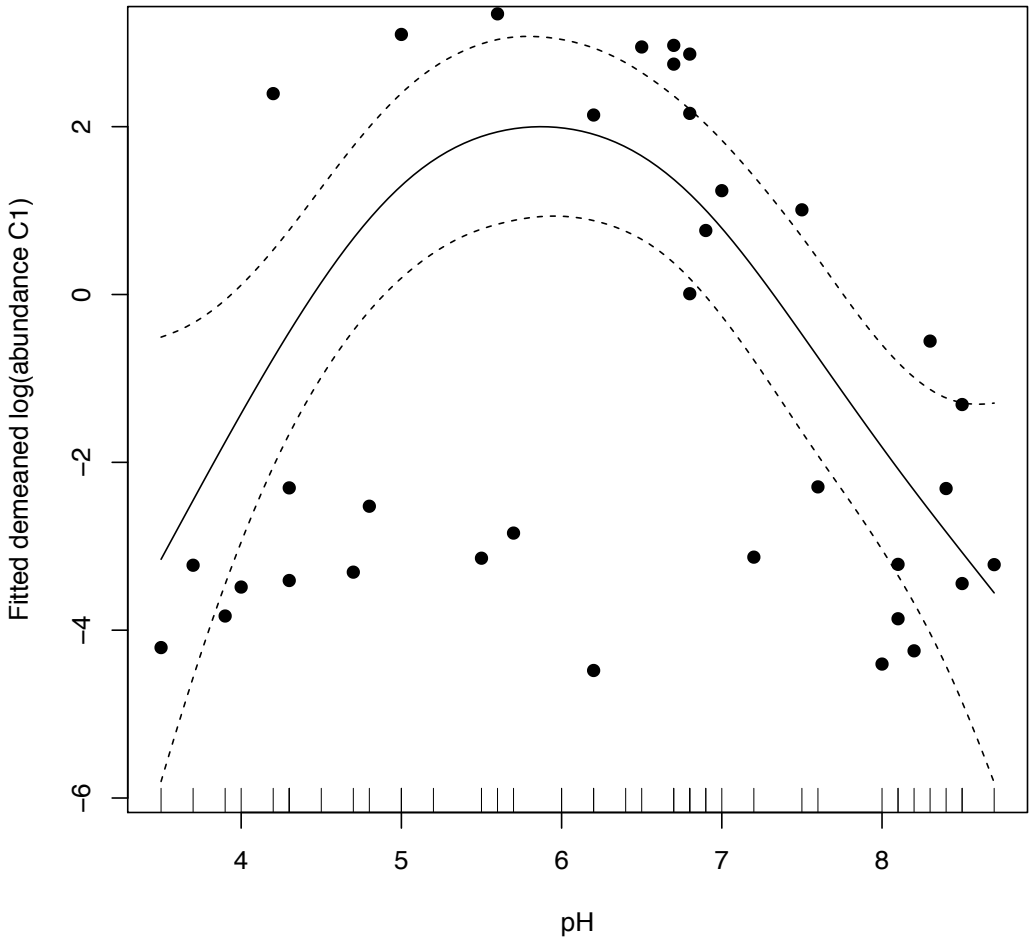
Generalising linear regression

- Generalised linear models (GLMs) use link functions to model non-Gaussian dependent variables
- Generalised additive models (GAMs) replace linear terms with arbitrary functions:

$$y_i = \beta_0 + \beta_1 f_1(x_{i1}) + \ldots + \beta_p f_p(x_{ip}) + \varepsilon_i$$

Modelling individual taxa abundance

- Linear models can be used to regress species abundances against environment variables (Morgan et al. Genome Biology 2012)
- However relationships often non-linear making GAMs more appropriate



Correcting for multiple comparisons

- In either discrete of continous case must correct for multiple comparisons
  - Familywise error rates: controls probability of one false discovery over whole data set: Bonferonni correction $p_i < \alpha / m$
  - False discovery rate: controls proportion of rejections that are false e.g. Benjamini–Hochberg

$$p_k < k\alpha / m$$

Summary

- Multivariate statistics are powerful tools for microbial community analysis
- There is a lot of jargon and a lot of methods
- Key is that you do not need to worry too much try to develop an intuition for what works on your dataset
- Try multiple approaches and if they differ in their predictions attempt to understand why