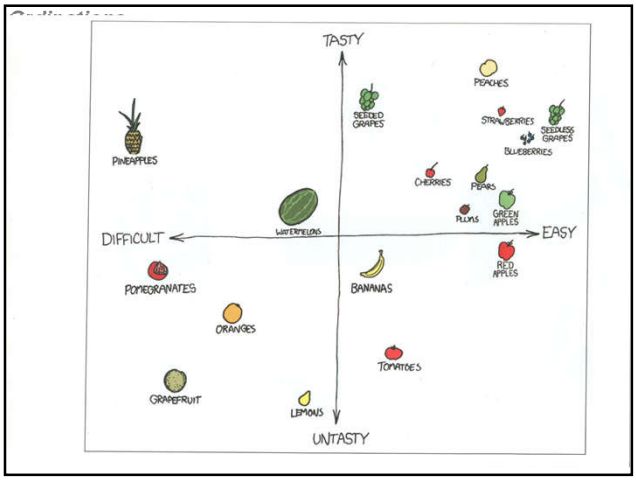
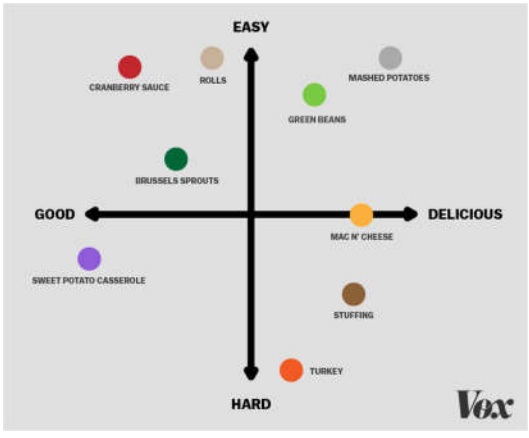
Ordinations

- **Goal** – order or arrange samples along one or two meaningful dimensions (axes or scales)
- Not specifically designed for hypothesis testing
- Similar to goals of clustering, so why not just use clustering?
- **Result**
  - 2 or more dimensional representation of samples
    - Dimensions typically orthogonal
  - Can be used as a data reduction technique



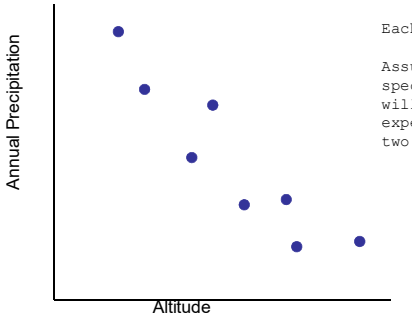


What do the dimensions or axes represent?

- **Ecological assumption** – species are distributed along ecological gradients in meaningful ways.
  - Eg. altitudinal gradient and vegetation
  - Gradients may not (usually not) be known
- **Direct Gradient Analysis** – gradients are known and measured, look for patters of species or samples distributed along gradients.
- **Indirect Gradient Analysis** – Gradients are unknown, goal of analysis is to construct gradients that explain distribution of species or points. Gradients = axes of ordinations

Direct Gradient Analysis

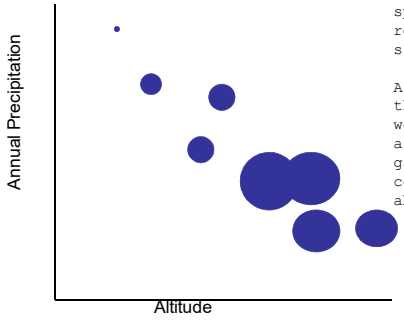
- Doesn't need to be an ordination.



Each point a sample.

Assumption is that species in the samples will respond in expected ways to the two gradients.

Direct Gradient Analysis - Overlay



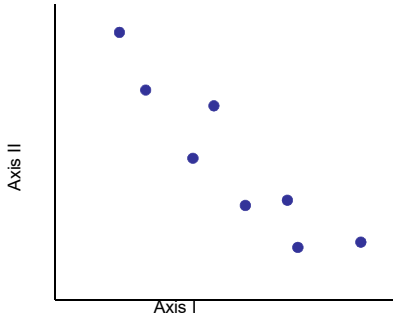
Each point a sample.

Here, the abundance of species 1 is represented by the size of the symbol.

A best fit line through the points would represent an altitude-precipitation gradient that would correlate with species abundance.

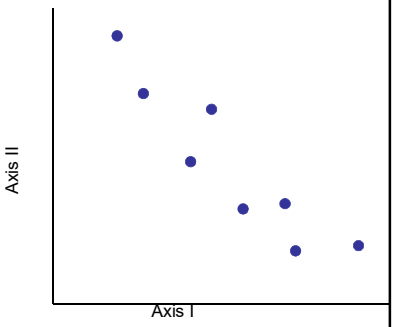
Indirect Gradient Analysis

- Most of the time the gradients are unknown, unmeasured, or the ultimate item of interest.
- Ordinations create axes that explain as much of the variability in the data as possible. These are your artificial gradients.



Indirect Gradient Analysis

- Major assumption – species do not covary randomly, there are patterns and thus hidden gradients.
- As species covariance drops, so does the percent of variance acounted for by the gradients. Less biological meaning to an axis.

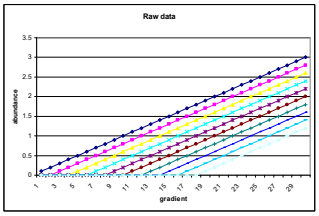


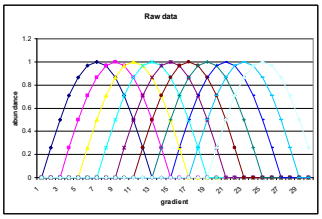
Direct vs. Indirect Gradient Analysis

- Direct gradient analysis will always be biased towards the gradients measured.
- There is no ability to detect the influence of other gradients, they are all “noise”
- Indirect gradient analysis – can look at relationships between constructed gradients and measured environmental variables or individual species abundance.

Ordination Data Assumptions

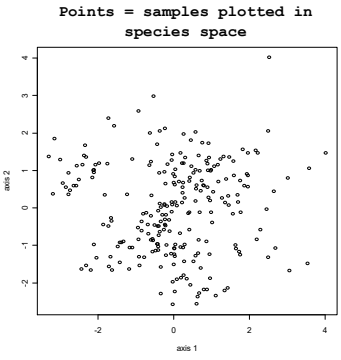
- Species distributions in relation to gradients
  - Monotonic or linear
  - Unimodal
- What to do with rare species
  - Eliminate?
  - Sampling bias (rarefaction)?





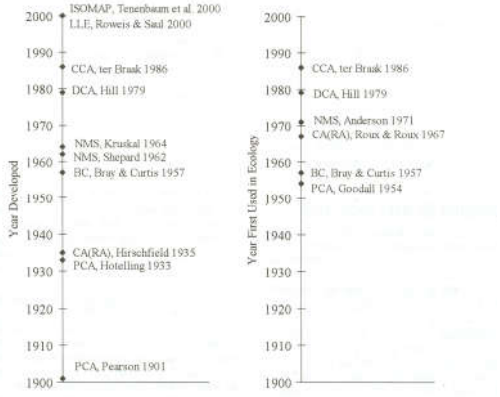
Interpretation of Ordinations

- Axes = gradients
- species will covary in meaningful ways = gradient represents trends in abundance of multiple species
- Thus, points close together in ordination space are at similar points along the gradient, should contain similar communities



History of Ordinations in Ecology

- From: McCune, B. and J. B. Grace. 2002. Analysis of Ecological Communities. MJM Software Design.



Polar Ordination (Bray and Curtis, 1957)

- Very first ordination, can be computed by hand
- Common sense extraction of gradients explaining species distributions
  - Start with triangular distance matrix
  - Select two most dissimilar samples (A and B), they define the ends (poles) of the first gradient (g)
  - Each (X) sample is projected onto that gradient:

$$X_{gi} = \frac{D_{AB}^2 + D_{Ai}^2 - D_{Bi}^2}{2D_{AB}}$$

  - X_{gi} = position of sample i on gradient g
  - Repeat for a second and third axis

Types of Ordinations (unconstrained)

- **Distance based ordination** – analysis works with a similarity (distance) matrix
  - Flexibility in what type of distance metric used
  - **NMDS** – non-metric multidimensional scaling
  - **PCoA** – principal coordinates or metric multidimensional scaling
- **Correlation or variance/covariance based ordination** – works with a variance/covariance or correlation matrix
  - **PCA** – principal coordinates analysis
  - **DCA** – detrended correspondence analysis
  - **CA** – correspondence analysis
- All but NMDS rely on eigenvalues and eigenvectors to construct axes.

Table 9.1 from Legendre and Legendre (1998)

Method	Distance	Variables
PCA	Euclidean	Quantitative, linear relationships assumed, beware of double-zeroes
PCoA	Any	Quantitative, qualitative or mixed
NMDS	Any	Quantitative, qualitative or mixed
CA	X^2	Non-negative, quantitative or binary
FA	Euclidean	Quantitative, linear relationships assumed, beware of double-zeroes

Eigenvectors and Eigenvalues

- Eigen – German for characteristic or proper
- For a given square matrix (A) with x by x dimensions there will be:
  - *x* eigenvalues (λ)
  - x eigenvectors (*v)* each with x values
  - Eigenvalues are the scalars associated with vectors, can be calculated with R function eigen()
- For a given matrix A:
  - $A\,v = \lambda\,v$

For a given matrix A:

$A\ v = \lambda\ v$

Matrix a:

$$\begin{matrix} 2 & 4 & 9 \\ 4 & 15 & 9 \\ 11 & 4 & 3 \end{matrix}$$

Function `eigen()` calculates:

Eigenvalues: 21.12 -7.38 6.25

Eigenvectors:

-0.3664331 0.6375372 -0.4237078
-0.8367287 0.1866362 0.7363839
-0.4069543 -0.7474713 -0.5274566

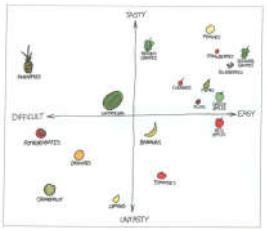
$A\ v = \lambda\ v$

$$\begin{matrix} 2 & 4 & 9 \\ 4 & 15 & 9 \\ 11 & 4 & 3 \end{matrix} * \begin{matrix} -.366 \\ -.837 \\ -.407 \end{matrix} = -7.74\ -17.68\ -8.59$$

$$\begin{matrix} -.366 \\ -.837 \\ -.407 \end{matrix} * 21.12 = -7.74\ -17.68\ -8.59$$

Eigenvectors and Eigenvalues

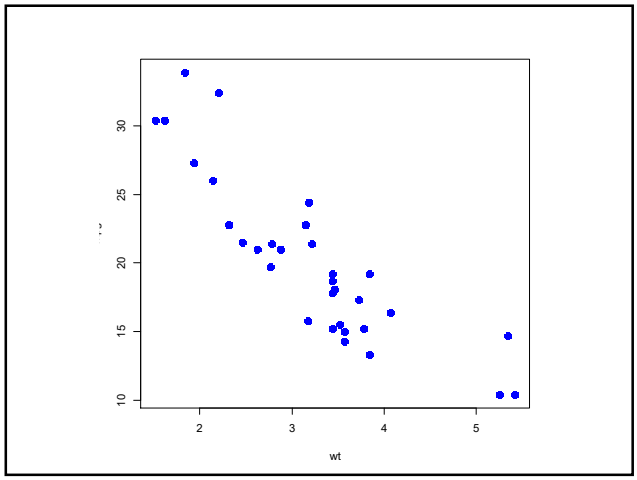
- Eigenvalues can only be calculated on a square matrix, most ecological data are in rectangular matrices.
- Most ordinations begin with summarizing a large rectangular matrix into a square matrix (similarity index, distance measure, correlation, variance/covariance etc.)
- The square matrix then contains information on relationships among samples and among variables.
- Extracting eigenvectors and values is an attempt to summarize these relationships into fewer dimensions.

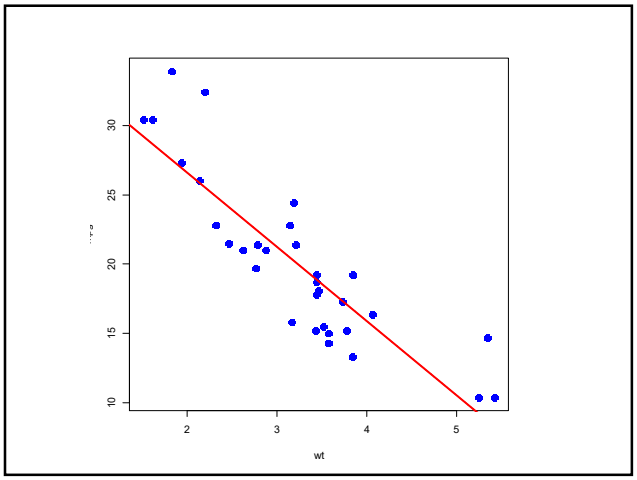


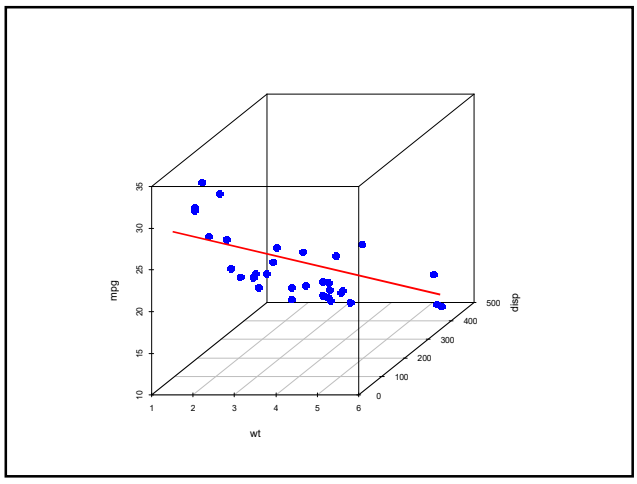
Eigenvectors and Eigenvalues

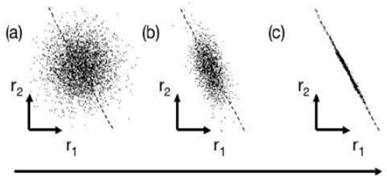
- Once this is done, eigenvalues and vectors can be interpreted as:
- **Eigenvalues** – the proportion of variation explained by each axis. Largest eigenvalue associated with the first, decreases with each additional axis.
- **Eigenvectors** – Vector of coefficients representing the contribution of an element (variable) to each axis. Eigenvectors are orthogonal (perpendicular in multivariate space).

- A square matrix with x dimensions will have x eigenvalues and x eigenvectors of length x
- Eigenvalues and eigenvectors are paired, typically in order of largest to smallest eigenvalue
- The number of eigen pairs (value+vector) is determined by the smaller dimension of the original data matrix (the dimensions of the square correlation, variance/covariance or similarity matrix)









- The variance explained will be determined, in part, with how much structure is in the data.
- Recall that one assumption is that there is meaningful structure in the data (i.e. species will show patterns of covariance among stamples)

Principal Coordinates Analysis (PCoA)

- Also called multidimensional scaling (MDS), not to be confused with non-metric multidimensional scaling (NMDS)
- Both MDS/PCoA are distance-based
- Start with triangular similarity matrix (of any type)
- Axes are formed which maximize the correlation between the similarity matrix and the Euclidian distance among samples in ordination space.
  - Logically aligned with trying to maximize a cophenetic correlation in a clustering framework
- If the distance measure is Euclidean, PCoA is identical to PCA (next week).

Code and assumptions

- Code (very simple)
  - `distance<-vegdist(community, method="bray")`
  - `prin_coord<-cmdscale(distance, k=3, eig=TRUE)`
    - Requires a similarity matrix
    - Specify how many axes to calculate, whether or not to report eigenvalues
- PCoA assumes a **linear** relationship between each species and the environmental gradient.
- If species show unimodal response, not appropriate.

PCoA Results

- Function cmdscale() returns
  - Points – location of all samples in ordination space
  - Eig – eigenvalues for all k axes
    - Decreasing order of strength (~% variance)
  - GOF – two measures of goodness of fit, want to maximize these (scree plot)

Reading and Examples

- Text: Ch 5.1, 5.2 and 5.5
- Ramachandran, S., O. Deshpande, C. Roseman, N. Rosenberg, M. Feldman, and L. Cavalli-Sforza. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences 102(44) 15942-15947.
- Two sample datasets and a script

Assignment

- Use the spaeth.csv data from earlier:
- Eliminate species with 2 or fewer occurrences, then log transform the data
- Calculate Bray-Curtis similarity with the log transformed data
  - Perform PCoA with k=3
    - Report total variance explained, and the percent variance explained by each axis
  - Repeat PCoA with k=4
    - Report total variance explained, and the percent variance explained by each axis
- Plot the first two axes of PCoA with different symbols or colors for creek.
- What species have the highest correlation with axis 1 and 2? How does that help interpret your plot?