BIO 682
Multivariate Statistics
Spring 2008

Steve Shuster
http://www4.nau.edu/shustercourses/BIO682/index.htm

Lecture 11

Properties of Community Data

Gauch 1982, Causton 1988, Jongman 1995

a. Qualitative: presence / absence, color, geomorphic settings, etc.

b. Quantitative: abundance, biomass, density, cover

c. Semi-quantitative / scale data (Braun-Blanquet, Domin, Log; but see Maarel 1979, Jager and Looman 1995)

Community Data Modifications

1. Transformations (Sokal and Rohlf 1995 pp. 409 – 422)
2. Relativizations by species and samples (Faith et al. 1987)
3. Taxonomic levels (Dauvin 1984, Gee et al. 1985, Warwick 1988, Warwick et al., 1990, Sommerfield and Clarke 1995)

Elements of Diversity

Magurran 1988, Brower et al. 1998, Buzas and Hayek 1998, Legendre and Legendre 1998

1. Evenness / Dominance (Simpson 1949)
2. Shannon's index (Shannon 1948)
3. K-dominance plots (Platt et al. 1984)

Species Richness

(Palmer 1990, 1991)

Species richness is the number of species in a given area. It is represented in equation form as S.

$$S = \Sigma\, s_i$$

Where s_i = the number of individuals in the i-th species.

Species ***richness*** is most often used in conservation studies to determine the sensitivity of ecosystems and their resident species.

The actual number of species calculated alone is largely an arbitrary number.

Shannon-Weiner Index of Diversity

$$\mathbf{H} = -\sum_{i=1}^{k} p_i \log p_i$$

where: k = the number of categories;
p_i = proportion of sample in category i,
$p_i = (f_i/n_i)$
f_i = number of cases in category i
n_i = sample size of category i
N = total cases

An Easier Method Is,

$$H = [N \log N - \Sigma f_i \log f_i]/N$$

where: f_i = number of cases in category i
N = total cases

This eliminates necessity for calculating p_i

An Estimate of Evenness

$$J = H / H_{max}$$

where H_{max} is the maximum diversity possible.

a. Perhaps a better estimator because the magnitude of H is affected by:
   1. the number of categories
   2. the distribution of data.

Other Problems

1. The SW index is *un-standardized*.
   a. Makes its use somewhat suspect unless standardized across analyses.

2. Confidence limits are not clearly defined.
   a. Therefore, it is difficult to make comparisons across situations.

The S-W Equation

$$H = \frac{n \log n - \sum f_i \log f_i}{n}$$

Note that here n = N.

Nests in Different Locations

$$H = \frac{n \log n - \sum f_i \log f_i}{n}$$

Sample 1	
Vines	5
Eaves	5
Branches	5
Cavities	5

$$[20 \log 20 - (5 \log 5 + 5 \log 5 + 5 \log 5 + 5 \log 5)]/20$$
$$= [26.0206 - (3.4949 + 3.4949 + 3.4949 + 3.4949)]/20$$
$$= 12.0410/20 = 0.602$$
$$H_{max} = \log 4 = 0.602$$
$$J = \frac{0.602}{0.602} = 1.000$$

Nests in Different Locations

$$H = \frac{n \log n - \sum f_i \log f_i}{n}$$

Sample 2	
Vines	1
Eaves	1
Branches	1
Cavities	17

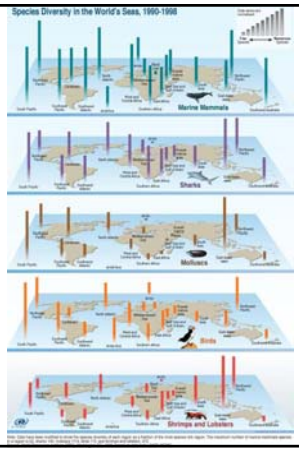
$$[20 \log 20 - (1 \log 1 + 1 \log 1 + 1 \log 1 + 17 \log 17)]/20$$
$$= [26.0206 - (0 + 0 + 0 + 20.9176)]/20$$
$$= 5.1030/20 = 0.255$$
$$H_{max} = \log 4 = 0.602$$
$$J = \frac{0.255}{0.602} = 0.424$$

Nests in Different Locations

$$H = \frac{n \log n - \sum f_i \log f_i}{n}$$

Sample 3	
Vines	2
Eaves	2
Branches	2
Cavities	34

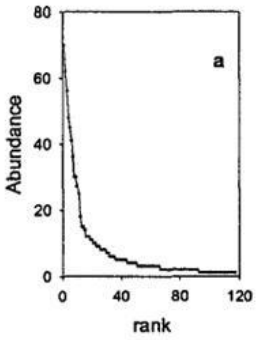
$$[40 \log 40 - (2 \log 2 + 2 \log 2 + 2 \log 2 + 34 \log 34)]/40$$
$$= [64.0824 - (0.6021 + 0.6021 + 0.6021 + 52.0703)]/40$$
$$= 10.2058/40 = 0.255$$
$$H_{max} = \log 4 \quad 0.602$$
$$J = \frac{0.255}{0.602} = 0.424$$



Rank/abundance, ABC or k-dominance plots

1. A rank/abundance plot (or Whittaker plot).
   a. Is used to visualize *species abundance distributions*
   b. The number of individuals of each species are sorted in descending order,
   c. The proportion of the total number of individuals for each species is plotted on the log scale against the species rank.

The rank/abundance plot



Its shape can provide an indication of dominance or evenness,

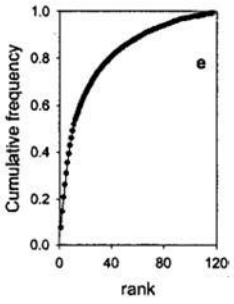
Steep plots signify assemblages with high dominance.

Shallower slopes indicate higher evenness.

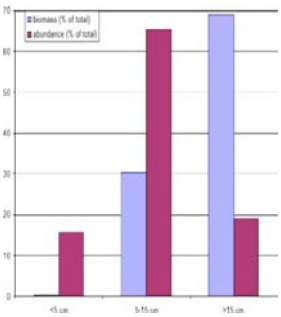
K-dominance plots

These display the cumulative proportion abundance against the log species rank.

Here, more elevated curves represent less diverse assemblages.



ABC plots



Are an adaption of k-dominance curves

Two measures of abundance are plotted:

1.Number of individuals

2.Biomass data.

This plot is useful for exploring the level of disturbance affecting the assemblage (abundance/biomass comparison).

Ecological Dissimilarity

A. Properties of useful ecological distance measures

B. Euclidean and non-Euclidean measures

C. Qualitative distance measures (Gauch 1982, van Tongeren 1995)

1. Jaccard's index - a statistic used for comparing the similarity and diversity of sample sets.

Jaccard's Index

is calculated by dividing the number of species found in both of two samples (j) by the number found in only one sample or the other (r) and then multiplying by 100. This gives a percentage of faunal similarity:

$$Jaccard's\ Index = \frac{j}{r} x\ 100$$

Sorensen's Quotient of Similarity (Q/S)

computes the percentage similarity between two samples:

$$Q/S = \frac{2j}{(a+b)}\ x\ 100$$

where *a* is the total number of species in sample #1, *b* is the number of species in Sample #2, and *j* is the number of species common to both samples.

Ecological Dissimilarity

D. Quantitative dissimilarity measures (Faith et al. 1987, Legendre and Legendre 1998)

1. Manhattan / City block
2. Euclidean
3. Bray – Curtis / Quantitative Sorensen
4. Chord / Relative Euclidean
5. Chi-Squared

The Problem

- Data consisting of measurements or counts from multiple variables collected from the same individual, group, community, population.
- e.g., Arthropod communities consist of many species
  - Need to rescale it to a few dimensions (traits)
- Multivariate statistics offer some options:
  - Principal Components Analysis (PCA)
  - Canonical Discriminant Analysis (CDA)
  - Non-metric Multi-Dimensional Scaling (NMDS)

Principal Components Analysis (PCA)

- A mathematical procedure that transforms a larger number of correlated variables into a smaller number of uncorrelated variables called ***principal components***.

Principal Components Analysis (PCA)

- The first principal component identifies an axis in multivariate space that accounts for the maximum variability in the data.
- In morphometric analyses, PC1 is usually associated with the "size" of the object under investigation, whereas PC2 is usually associated with object "shape;"

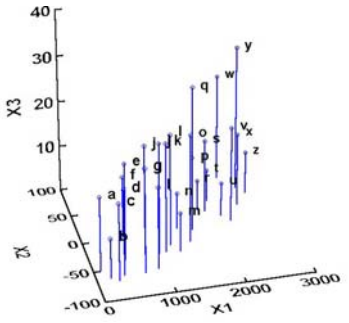
Principal Components Analysis (PCA)

- For community analysis these 2 axes can be used to define a "centroid" of community phenotypes (the average of all points; usually represented as a cluster in 2D space).
- Each succeeding component accounts for the remaining variability in the data.

Principal Components Analysis (PCA)

- PCA takes your cloud of data points, and ***rotates*** it (in multidimensional space) such that the maximum variability is visible.
- Another way of saying this is that it identifies your most important ***gradients***.

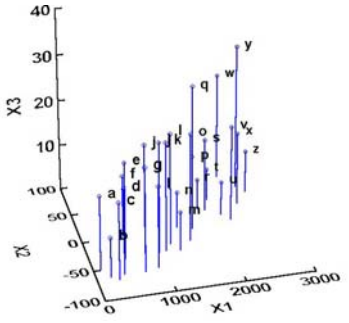
Hypothetical Example of 3 Species Abundances



Here, X1 and X2 are related to each other (note the covariance),

But X3 is not clearly related to either X1 or X2 (note the elevational variation in X3).

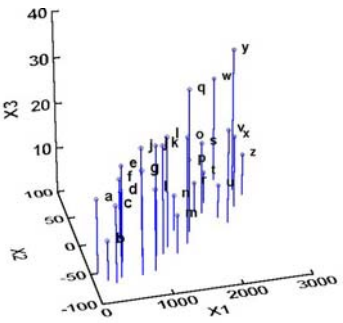
Hypothetical Example of 3 Species Abundances



Question: is there a ***hidden gradient*** along which our samples vary with respect to species composition?

Rotating the Data

First, PCA standardizes the data by subtracting the mean of each axis from each point $(X - x_i)$, $(Y - y_i)$ and $(Z - z_i)$.



For Example,

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

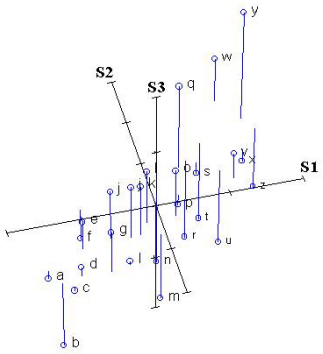
X=1.81; Y=1.91

Rotating the Data

The makes the centroid of the whole data set ***equal to zero***.

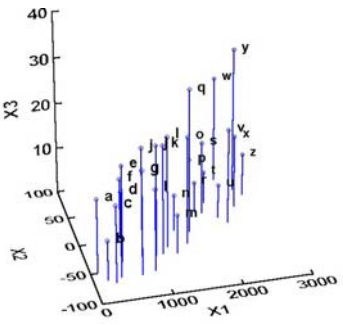
The standardized axes can be labeled, S1, S2, and S3.

Note: The ***relative location*** of points remains the same:



Rotating the Data

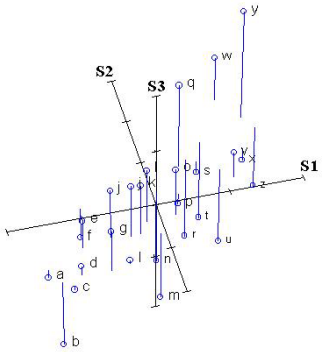
You can then either divide by the standard deviation - or not.



Which Denominator?

A PCA ***without*** dividing by the standard deviation is an eigenanalysis of the ***covariance*** matrix.

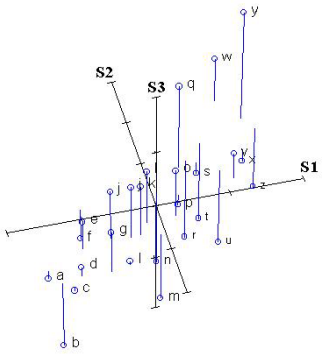
A PCA in which you ***do*** divide by the standard deviation is an eigenanalysis of the ***correlation*** matrix.



When to Standardize?

If we want relative abundances to matter, a covariance matrix might be better.

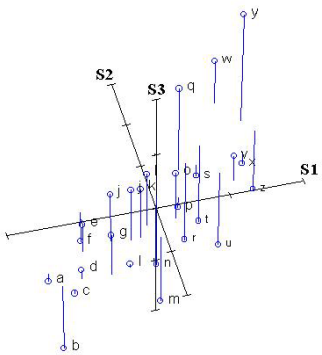
By standardizing, we are giving all species the same variation, i.e. a standard deviation of 1; this is better if units differ.



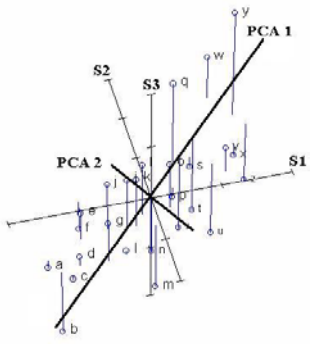
Eigenanalysis?

Eigenanalysis is the process of finding the eigenvectors and eigenvalues of a multivariate data set.

Eigenvectors identify the line that defines the relationship between pairs of variables; ***Eigenvalues*** scale the vector to define its length and direction.



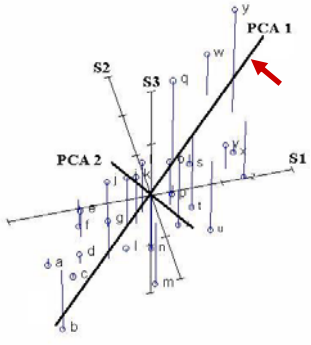
The “Principal Components”



A gradient is visible : from the lower left front to the upper right back.

An underlying gradient exists along which species 1 and species 2 both increase (In the language of Gauch (1982), species 1 and 2 both contain some "redundant" information.

PCA 1



PCA1 is the line that goes through the centroid, and minimizes the square of the distance of each point to that line.

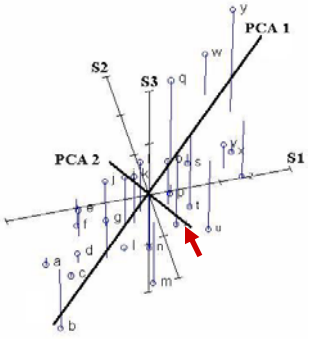
This line is as close to all of the data as possible.

Or, the line goes through the maximum variation in the data.

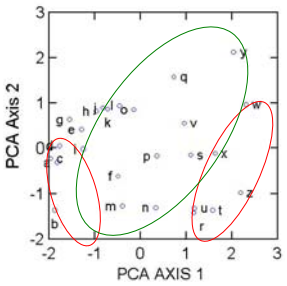
PCA 2

The second PCA axis also goes through the centroid and transects th maximum variation in the data,

BUT, it must be completely uncorrelated (i.e. at right angles, or "orthogonal") to PCA axis 1.



A Bivariate Plot of the Result



What is the underlying biology behind such a gradient?

PCA, and any other indirect gradient analysis, ***cannot answer this question***.

This is where the biological interpretation comes in.

Which Axes to Include?

PCA Axis 1: 63%
PCA Axis 2: 33%
PCA Axis 3: 4%

Every axis has an eigenvalue (also called latent root) associated with it, and they are ranked from the highest to the lowest.

The smaller the eigenvalue, the less important the axis.

Which Species Contribute Most?

Look at the component loadings (or "factor loadings"): the highest values in each PC identify the most important species for that PC.

Species	PCA 1	PCA 2	PCA 3
S1	0.9688	0.0664	-0.2387
S2	0.9701	0.0408	0.2391
S3	-0.1045	0.9945	0.0061

S1 and S2 load most heavily in PCA1; these are the species that contribute most to this axis

S3 loads most heavily in PCA1; this species contributes most to this axis

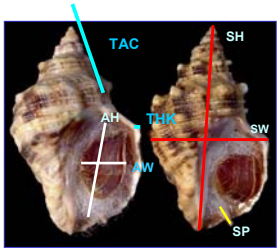
None of the species appear to load heavily in PCA3

Shell Characteristics Measured

R.M.Cota 2008

7 measurements:

- Shell Height (SH)
- Shell Width (SW)
- Aperture Height (AH)
- Aperture Width (AW)
- Apertural spine length (SP)
- Top of aperture to crown (TAC)
- Thickness of aperture (THK)



Individuals Measured

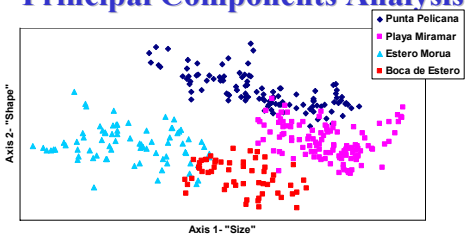
Punta Pelicana-111

Playa Miramar- 118

Estero Morua- 77

Boca de Estero-56

Principal Components Analysis



Punta Pelicana (N=111), Playa Miramar (N=118), Estero Morua (N=77), Boca de Estero (N=56). (total N=362)

PCA axis one (size) explained the greatest amount of variation data with 93.9%. The second axis (shape) accounts for 3.6% of the variation in morphology of *Acanthina angelica* (Fig 4).

The Multi-Response Permutation Procedure resulted in an A value=0.48 and a p value<0.00000001.

BIO 682 - Quantitative Biology
Multiple Regression Homework
Due 13 April 2009

This is a short assignment designed to help you become familiar with using multivariate data sets. You are required to analyze the data below, print out the output (or put it into "journal" format), and then explain in a page or less (i.e., you need not go into excruciating detail) what you can say about the data using multiple regression. You can either hand in the hard copy or send it to me by email.

The data below give the body fat, triceps skinfold thickness, thigh circumference and midarm circumference for twenty healthy females aged 20 to 34 (reference: http://jamsb.austms.org.au/staff/dunn/Datasets/applications/health/bodyfat.html). The body fat estimate was obtained by an expensive and cumbersome procedure requiring the immersion of the person in water. Researchers would like to find a regression model with some or all of these predictor variables that could provide reliable predictions of the amount of body fat a person has, since the measurements needed for the predictor variables are easy to obtain. Your job is to find out which variables are most important in predicting this relationship.

For this assignment, you will need to import the data into a JMP data table and then use JMP's stepwise regression platform to become familiar with this procedure. JMP provides a detailed description of this procedure in its HELP section. I suggest you read this section, then use it to work your way through your analysis of the body fat data set.

Fat	Triceps	Thigh	Midarm
11.9	19.5	43.1	29.1
22.8	24.7	49.8	28.2
18.7	30.7	51.9	37
20.1	29.8	54.3	31.1
12.9	19.1	42.2	30.9
21.7	25.6	53.9	23.7
27.1	31.4	58.5	27.6
25.4	27.9	52.1	30.6
21.3	22.1	49.9	23.2
19.3	25.5	53.5	24.8
25.4	31.1	56.6	30
27.2	30.4	56.7	28.3
11.7	18.7	46.5	23
17.8	19.7	44.2	28.6
12.8	14.6	42.7	21.3
23.9	29.5	54.4	30.1
22.6	27.7	55.3	25.7
25.4	30.2	58.6	24.6
14.8	22.7	48.2	27.1
21.1	25.2	51	27.5