Ordination Methods – PCA

- *Objectives:*

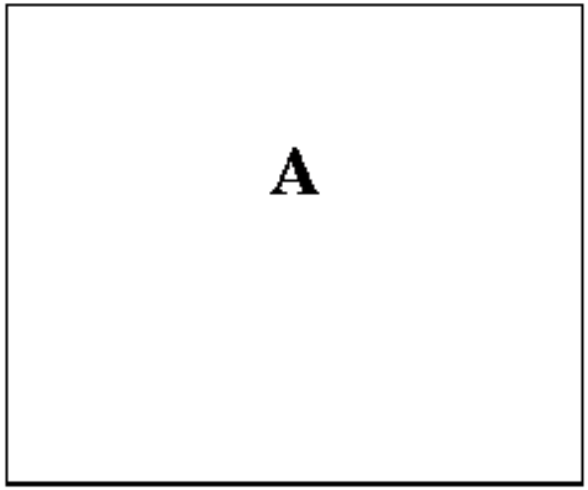
Discuss PCA within context of ordination

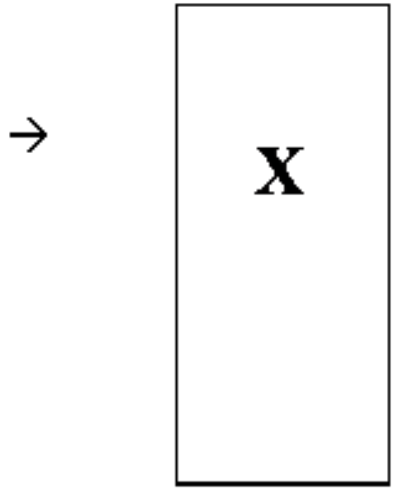
Go over output from ordination methods

Go over PCA set-up and output

Ordination - Objectives

- Arranging items (samples / species) along one or more axes
- Graphical summarization of complex relationships
- Extracting one or more dominant patterns – % variance
- Synthesis (reduction) of large datasets into fewer variables
- These variables are then related to environmental variables





The k axes represent the strongest correlation structure in the data. "Axes" are also called "principal components"

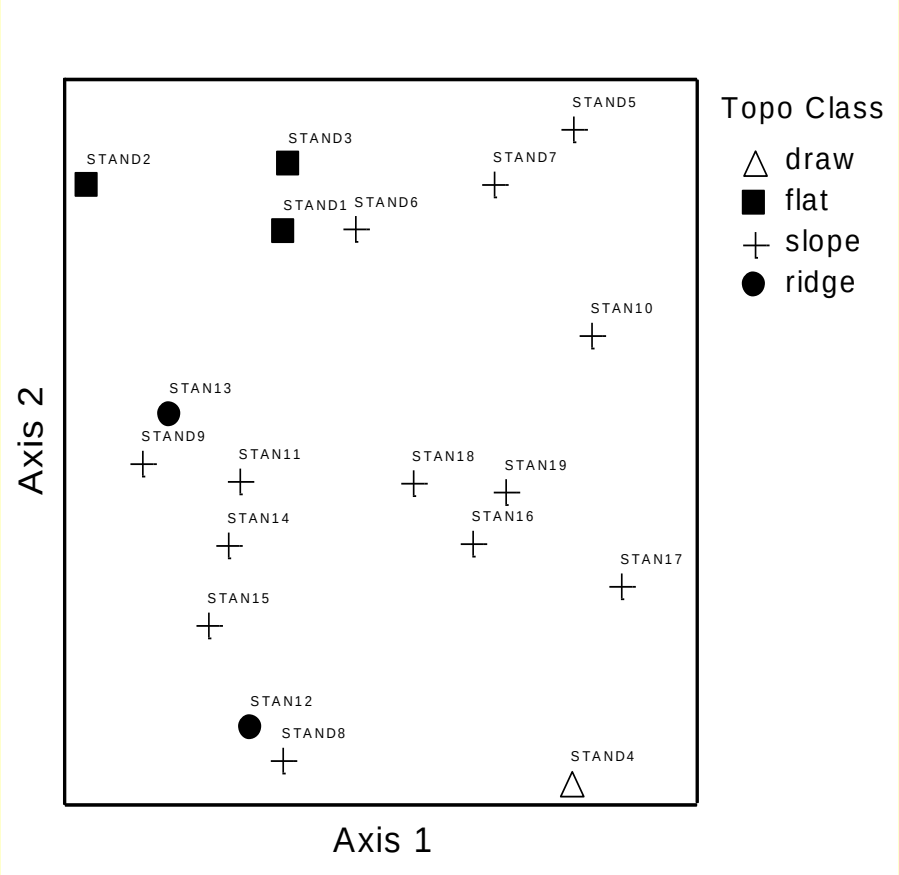
Components are independent from each other

Ordination - Output

- Typically, a 2-dimensional plot of samples / species in terms of synthetic axes (combinations of variables)
- Ideally, the distance between points in ordination space is proportional to underlying distance measures in variable space
- NOT LIKE A REGRESSION

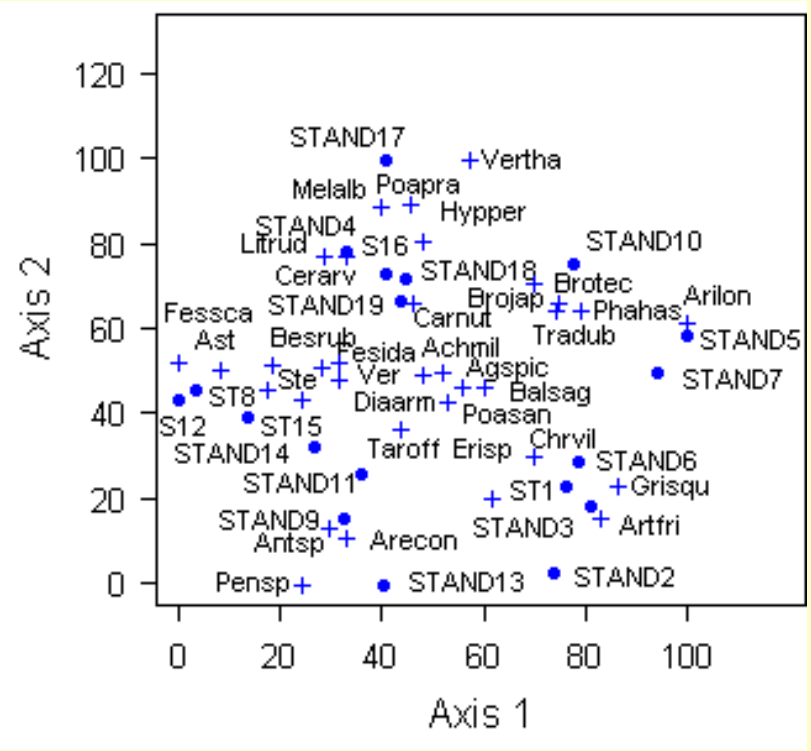
(Axes uncorrelated, by definition)

- Plot samples / species
- If possible, use points for samples, overlays for species
- Also can code samples by habitat types (using a key environmental variable)



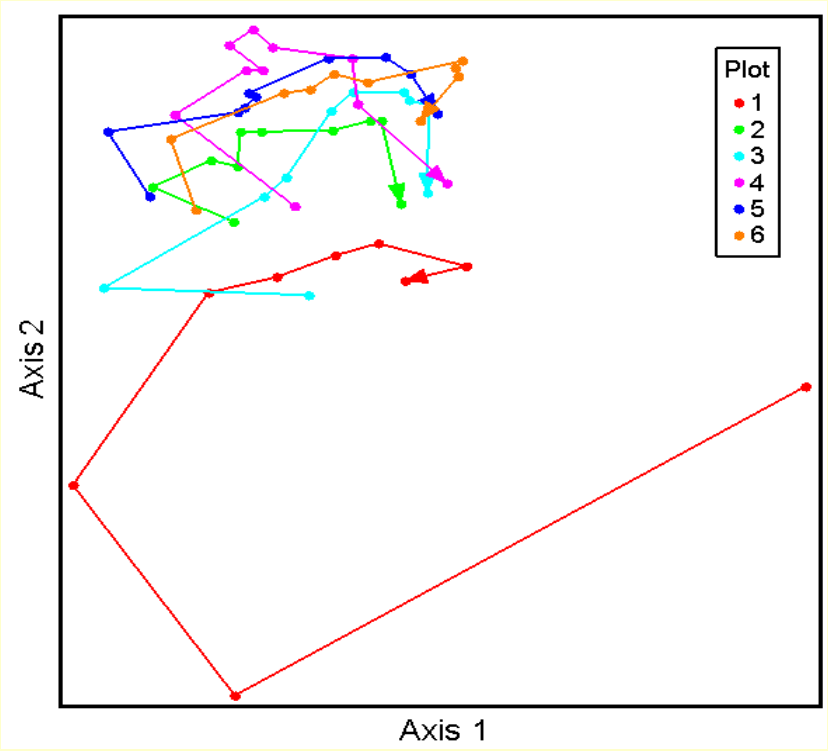
Ordination - Output

- Organization of samples / species :



- Interpretation of temporal change:

successional vectors



Ordination - Output

- How many axes? How many discrete signals in dataset

  Different rules for selecting number of axes

- Significance assigned to axes and contributing variables

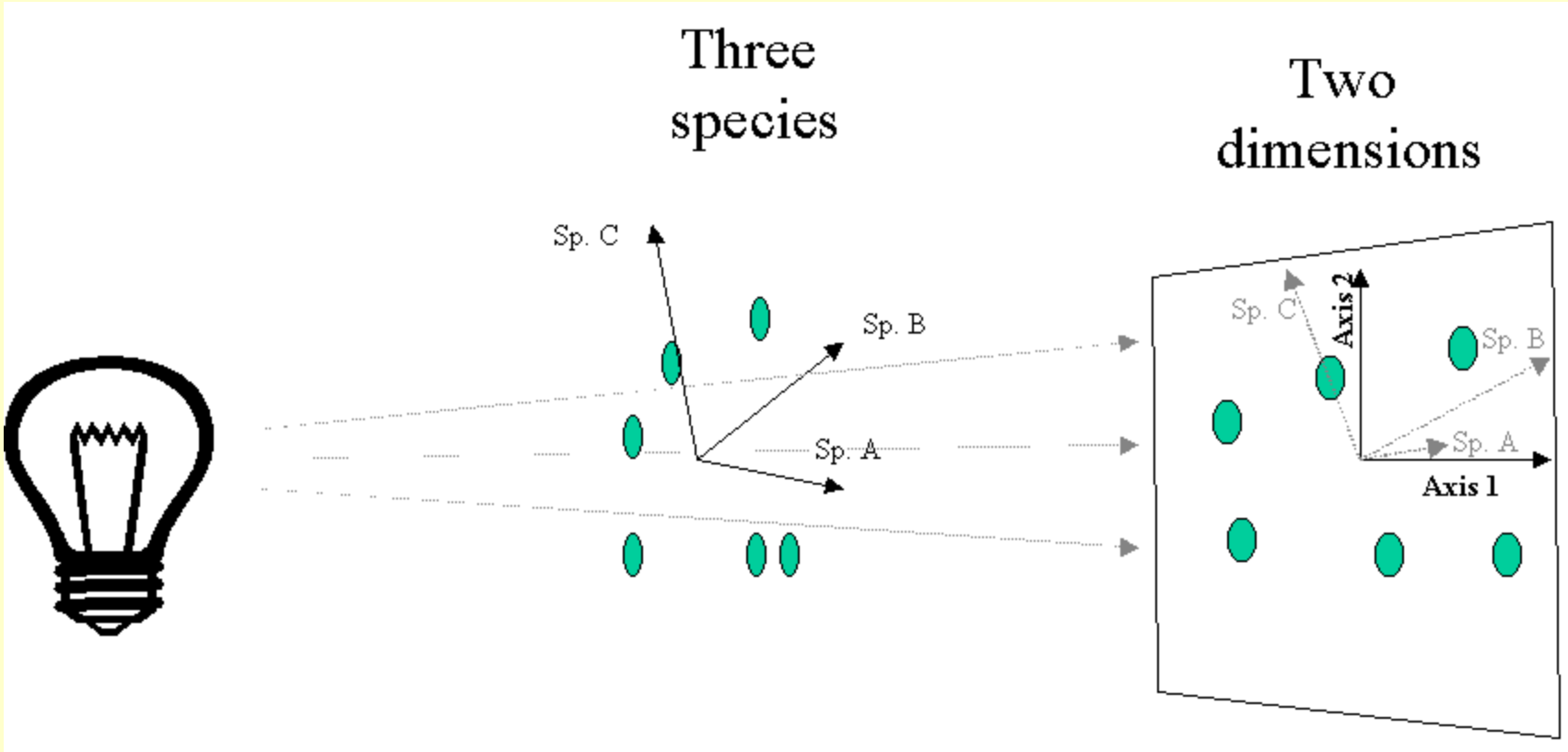
  Yet, most studies select 2 or 3 axes

- Interpretation of results: overlays, correlations with axes

Ordination - Output

- Visualizing your Results: Dimensionality Matters

Easier to visualize 2D or 3D ordinations…
but the PC axes represent gradients.



Ordination - Output

- How many axes? Number of independent signals in dataset

	r^2	
Axis	Increment	Cumulative
1	0.371	0.371
2	0.199	0.570
3	0.146	0.716

Coefficient of determination: % variance represented

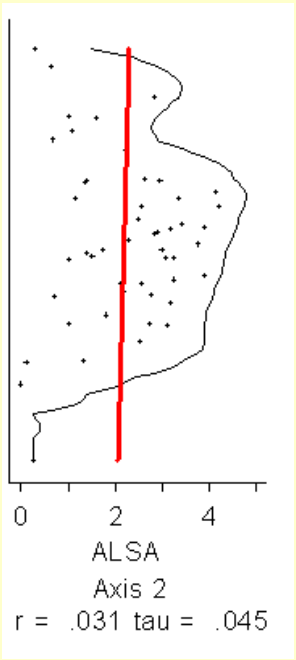
- Rules of thumb: relative vs absolute

  (% variance vs # axes) (PCA eigenvalues)

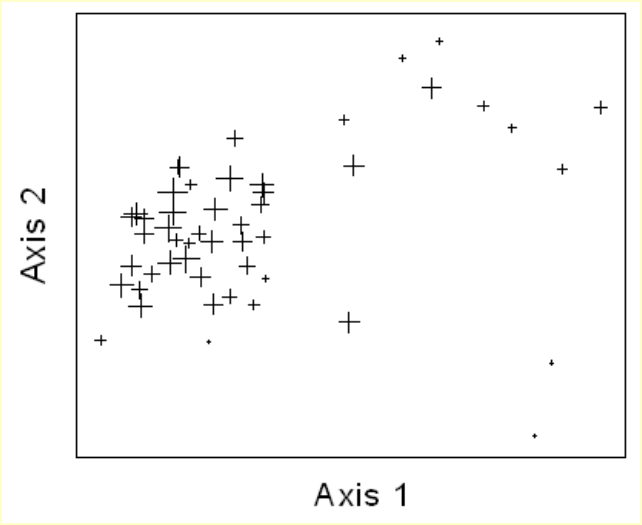
- Yet, most studies select 2 (or 3 axes): Intuitive explanation
- Strength assigned to axes and contributing variables (r)

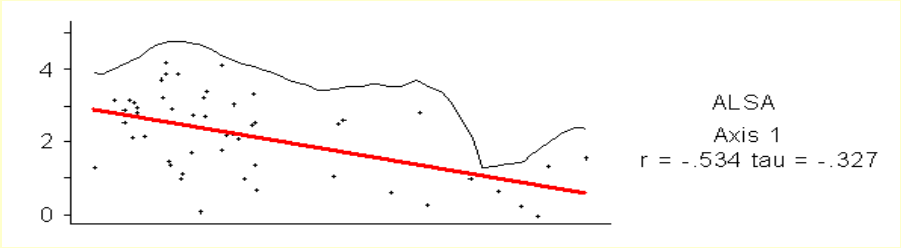
Ordination - Output

- Interpretation of results: overlays, correlations with axes



r = + 0.031
tau = + 0.045





r = - 0.534
tau = - 0.327

Conclusions: Species ALSA negatively correlated with Axis 1. Variance explained = 25%

- Beware when interpreting correlation coefficients:
  - outliers can have strong influence
  - coefficients meaningless if relationships not linear
  - correlations coefficients invalid with binary data

Ordination - Output

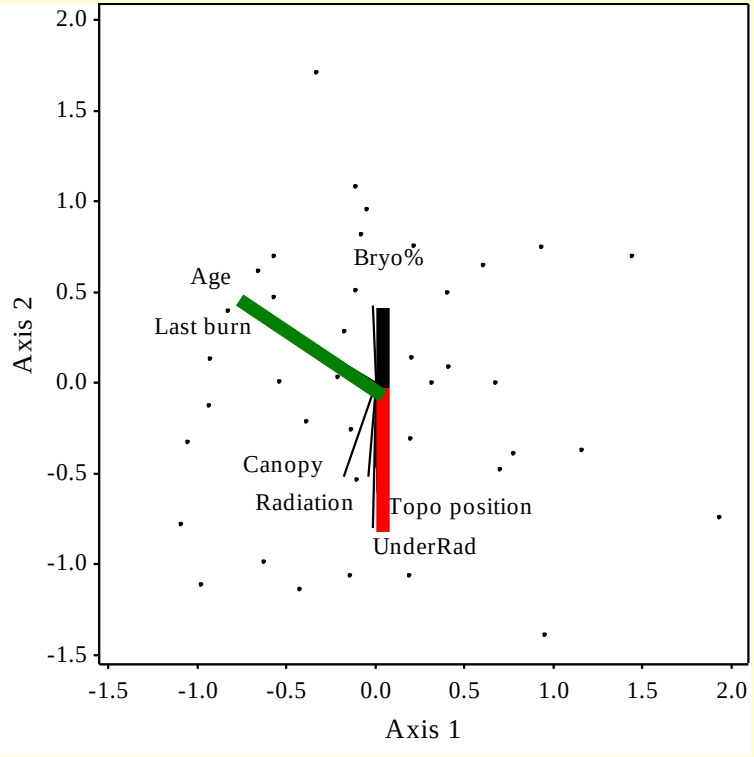
- External Evaluation: Correlations with second matrix

(e.g., how are environmental variables related to axes ?)

correlation coefficients (species & variables / axes)

- Graphical representation of environmental correlates:

angles / lengths indicate strength and direction of each environmental variable's association with ordination axes



Ordination - Output

- Comparisons with null model:

Comparing results observed from *real data* with results from *randomized data* using randomization tests (p values).

Beware

This approach can yield no clear results if strong outliers cause spurious significant patterns with random dataset.

New option in PC-ORD: Test Type-I error rate

(shuffle data and test random patterns for significance)

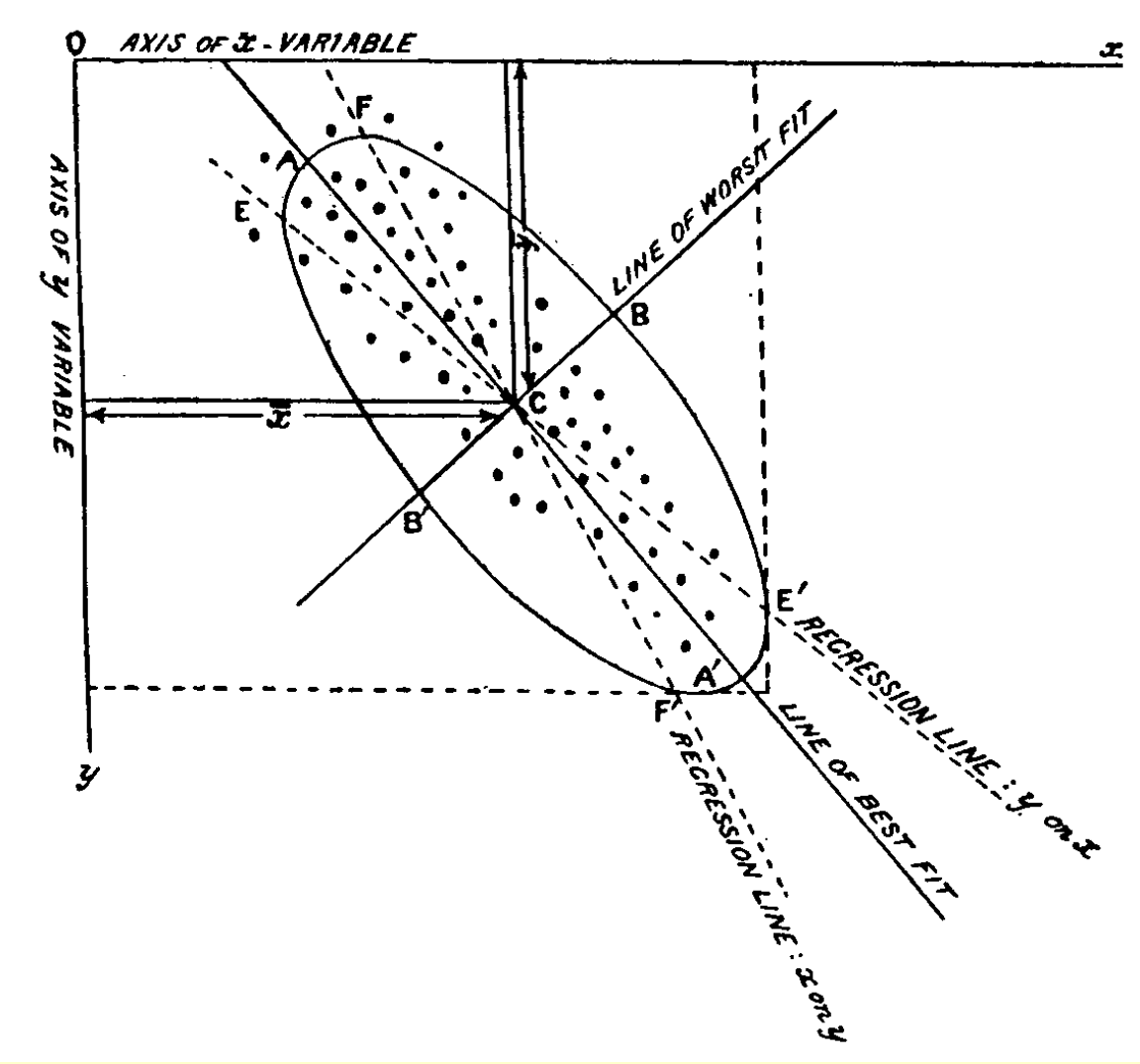
Principal Component Analysis (PCA)

- Approach:

Identify independent axes of variability

(at 90 degree angle)

Linear combinations of multiple variables

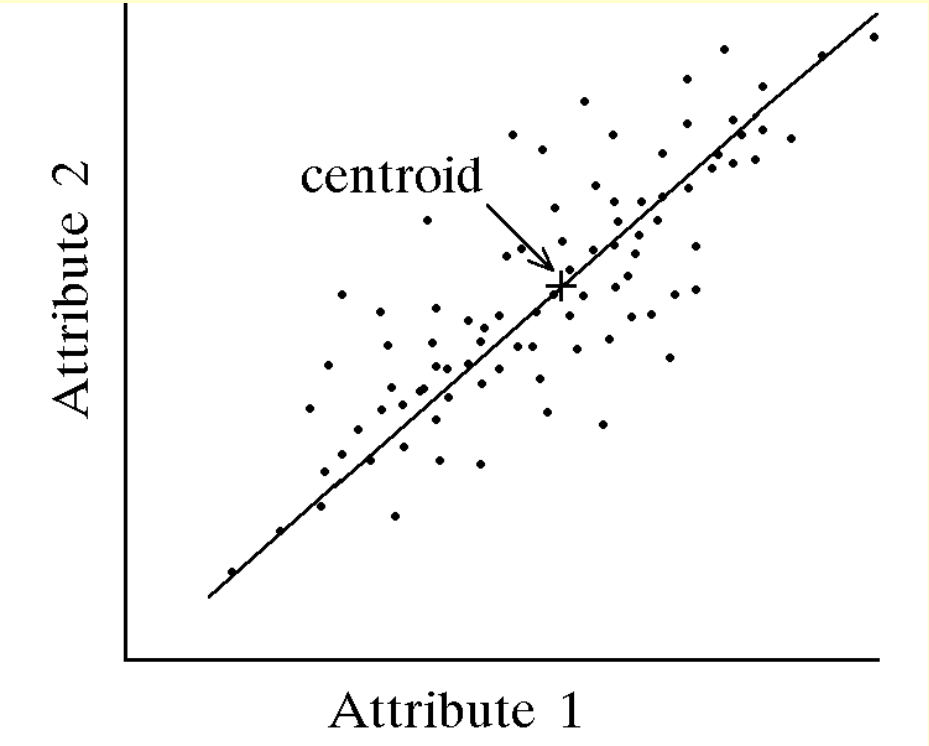


Principal Components Analysis (PCA)

- Using the best-fit straight line to describe a system of points in multiple dimensions using straight lines (Pearson 1901)

$$Y = a_0 + a_1 X_1 + a_2 X_{2+} \ldots$$

- Start with cloud of *n* points in *p*-dimensional space
- Center the axes in the point cloud (centroid)
- Rotate axes to maximize the variance along axes
- As rotation angle changes, the variance changes



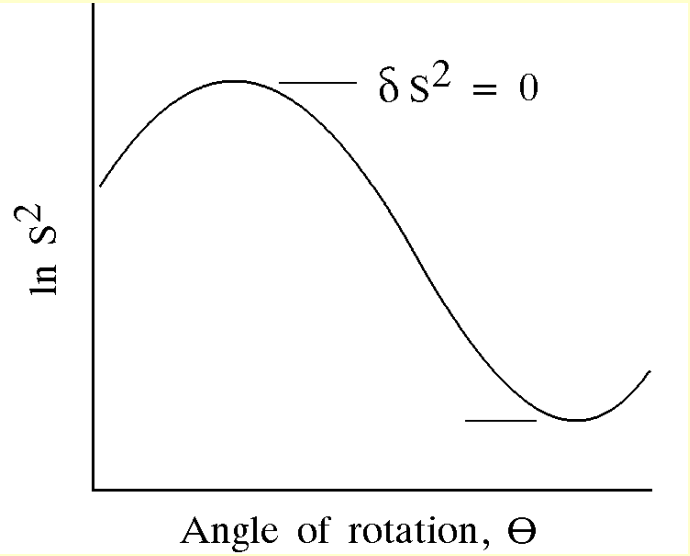
Principal Component Analysis (PCA)

- Approach: At maximum variance, all partial derivatives will be zero (no slope in all dimensions). This means that we found the angle of rotation θ such that:

$$\frac{\delta \ln s^2}{\delta \cos\theta} = 0$$

NOTE: for each component , the lower case delta (δ) indicates a partial derivative.

PCA rotates the point cloud to maximize the variance along axes.



Principal Component Analysis (PCA)

- Approach:

From data matrix **A** containing *n* objects (samples) by *p* variables, calculate **cross-products matrix**:

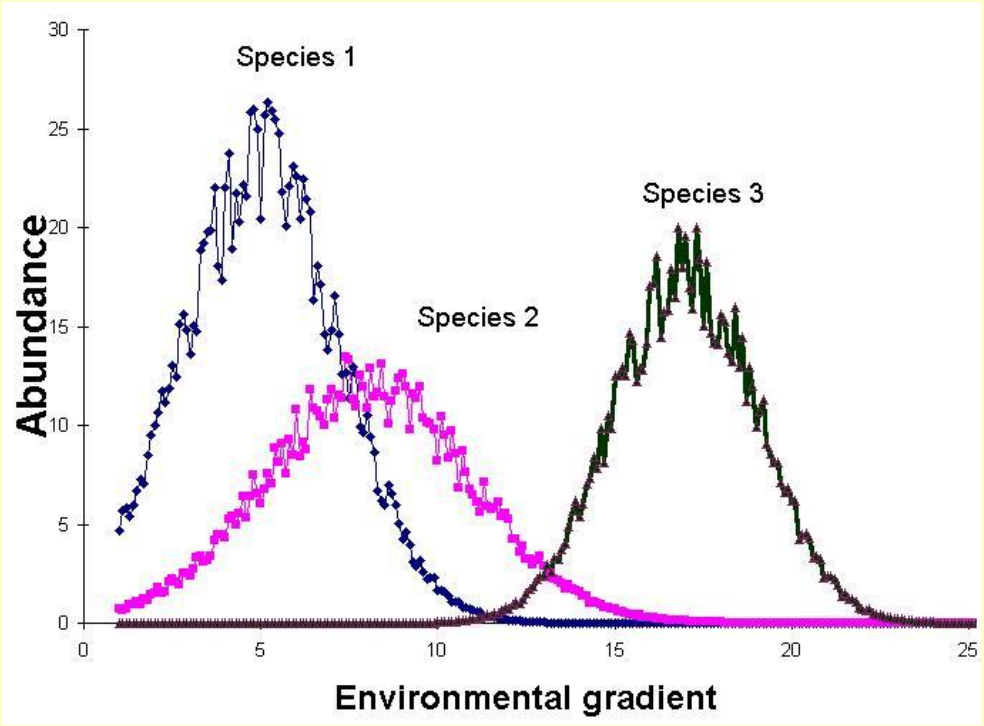
$$\text{where } b_{ij} = a_{ij} - \bar{a}_j$$

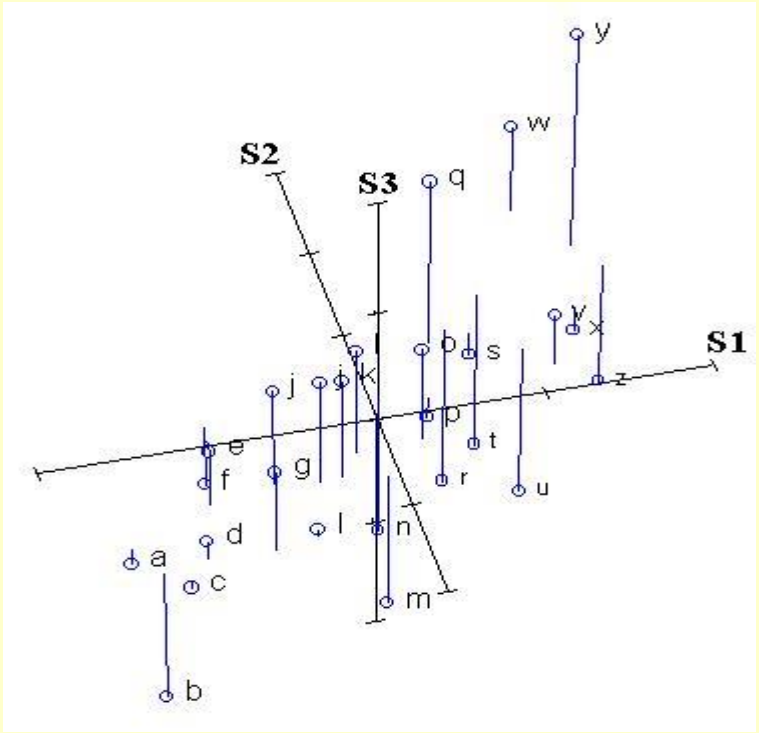
The equation for a **correlation matrix** is the same as above except that each difference is divided by the standard deviation, s_j.

$$\text{where } b_{ij} = \frac{a_{ij} - \bar{a}_j}{s_j}$$

PCA Example

- Samples plotted 1D gradient or in 3D - species space





First, CA rotates cloud of data points, so the maximum variability is visible.

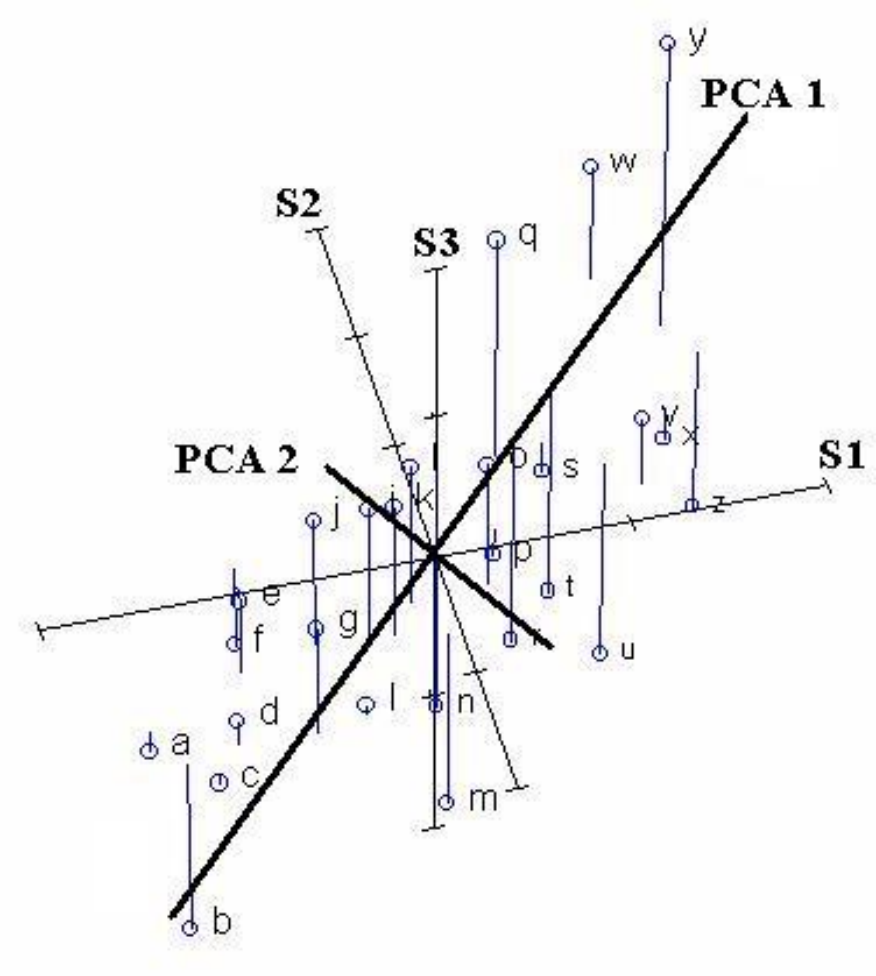
PCA identifies the strongest gradient.

Next stage is to standardize the data by subtracting the mean and dividing by the standard deviation.

Centroid of whole data set is zero.

PCA Example

- Samples plotted in 3D - species space



PCA chooses first axis as that line that goes through centroid, but also minimizes square of the distance of each point to that line.

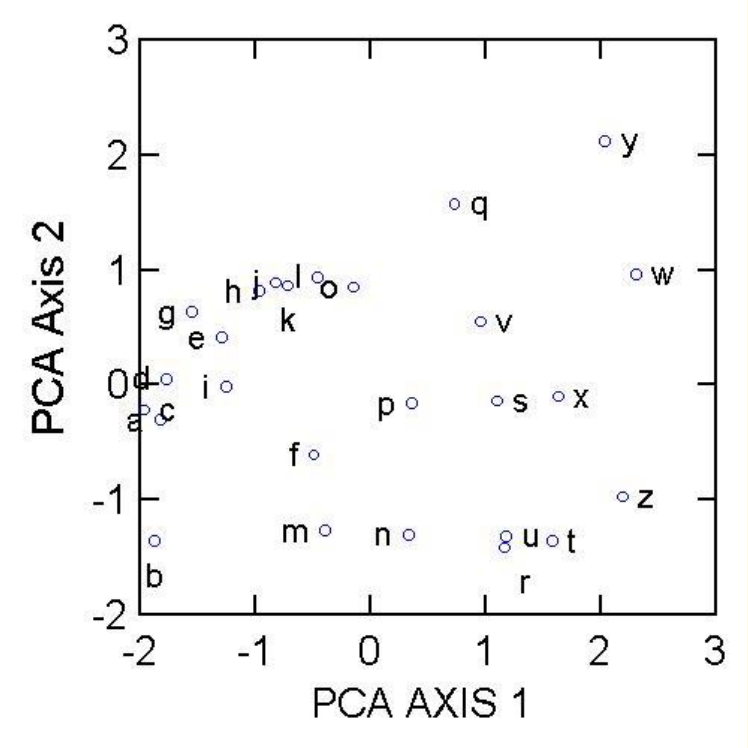
Thus, the line is as close to the data as possible. It goes through the dimension of maximum variation in the data.

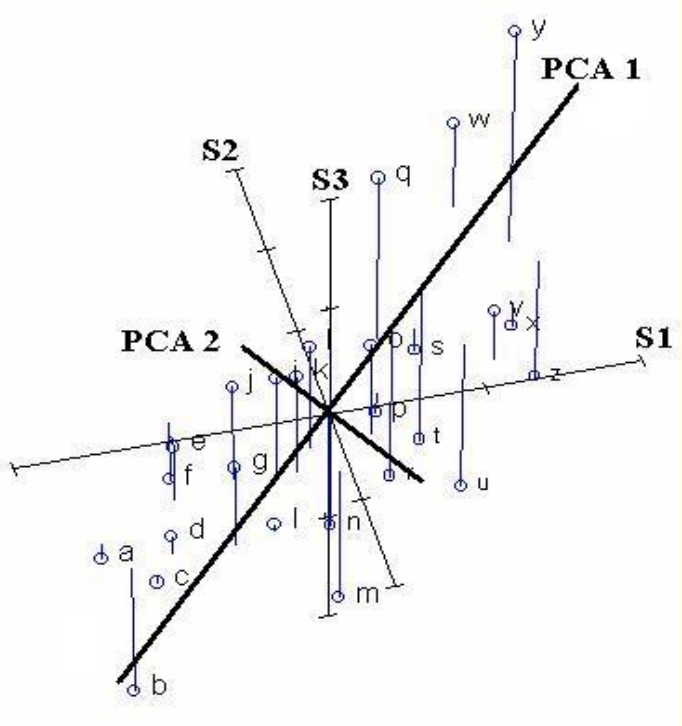
Second PCA axis also goes through centroid, and through the maximum variation in the data, but with constraint:

It is uncorrelated to PCA axis 1.
(i.e., at right angles, or "orthogonal")

PCA Example

- Samples plotted in 2D PCA





PC Axis1: From samples (a, b, c, d, g) to samples (t, w, x, y, z).

PC Axis2: From samples (b, m, n, u, r, t) to samples (o, l, q, w, y).

PCA Example

- Amount of variance explained by PC axes:

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

Interpretation: 100% of variance explained by axes, when # of variables = # of axes. PC1 > PC2 > … PCn

- Loadings of species in the PC axes:

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

Interpretation: The value of a sample along the first PC axis is:
+ 0.9688 times the standardized abundance of species 1 PLUS
+ 0.9701 times the standardized abundance of species 2 PLUS
- 0.1045 times the standardized abundance of species 3.

When to Use PCA

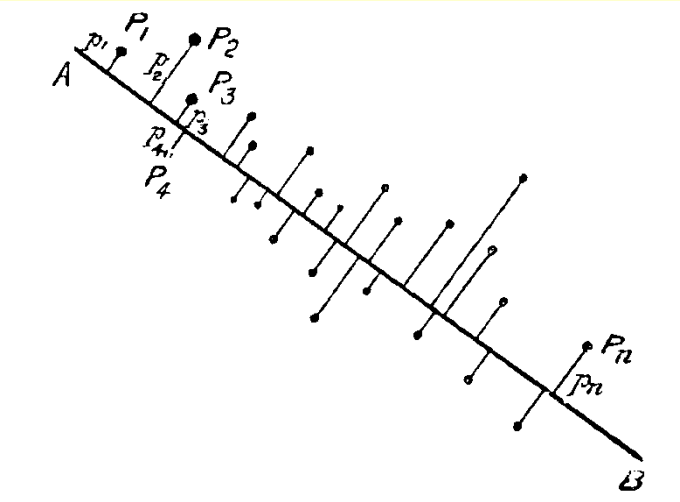
- Normality: Ideal for normal data with approximately linear relationships amongst variables – Rarely for community data
  - Beware of heterogeneous community data
  - Critical to justify the use of this parametric approach

- Sample size: Need robust estimate of correlation structure
  - Stronger patterns require smaller sample sizes
  - Increasing number of variables, strengthens results (Pillar 1999)
  - Rule of thumb for sample size: 5 samples per variable (Tabachnick and Fidell 1989)

PCA Limitations

Critical to meet the assumption of normality

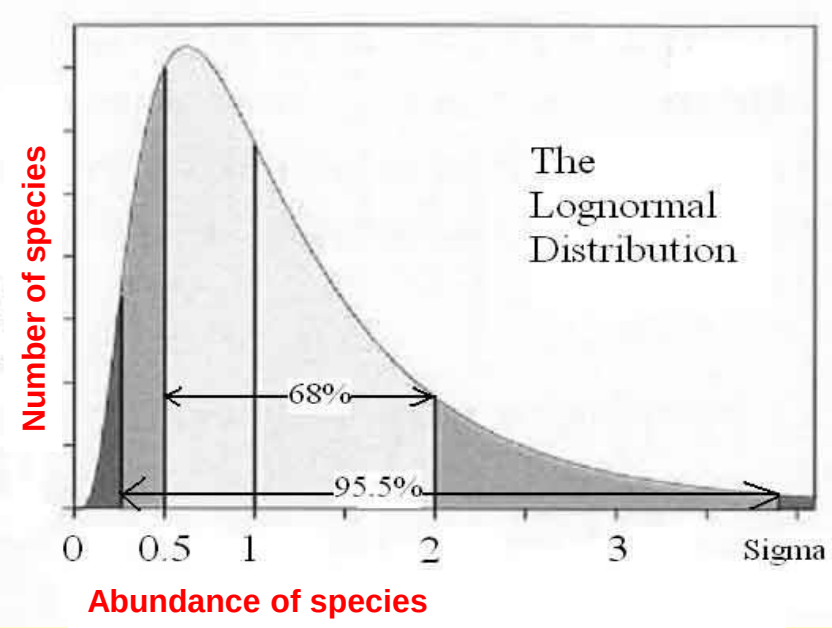
(for linear regression)



Beware of Community Data (species):

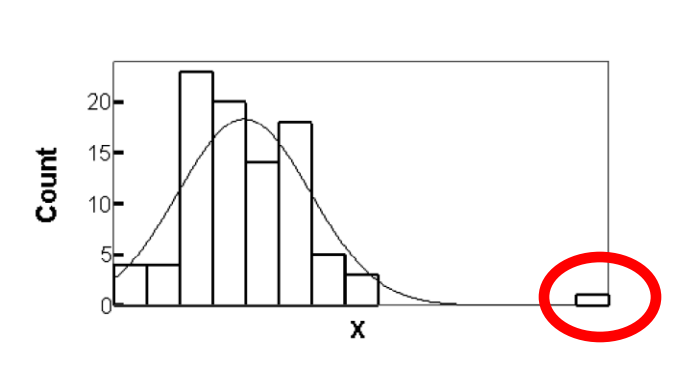
High Abundances

Species Absences

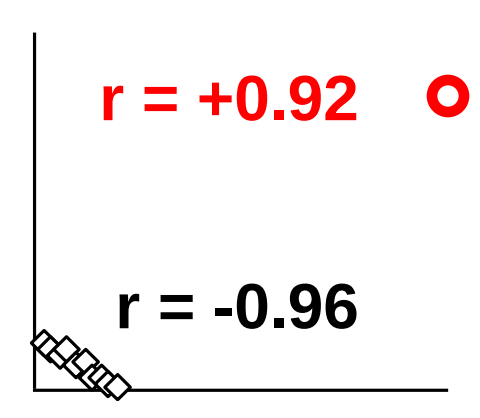


PCA Assumptions – No Outliers

- Use bivariate scatterplots to assess linear relationships



- Beware of outliers – they can change cross-correlations



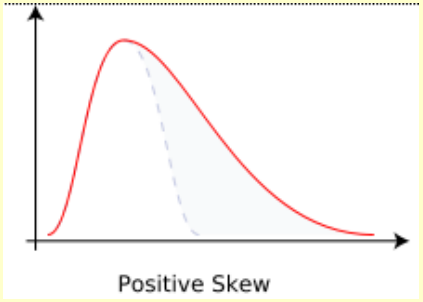
Solutions:

Transform the data

Remove outliers

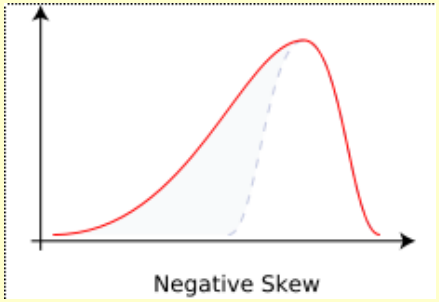
PCA Assumptions – Normality

- Assessed with skewness (asymmetry) / kurtosis (peakiness)

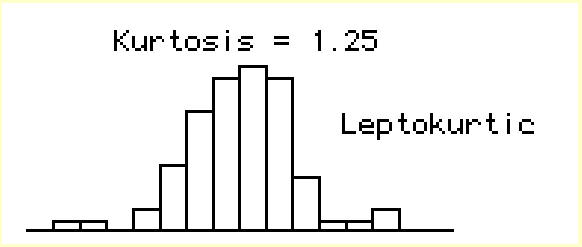


skew= 0

(normal data)

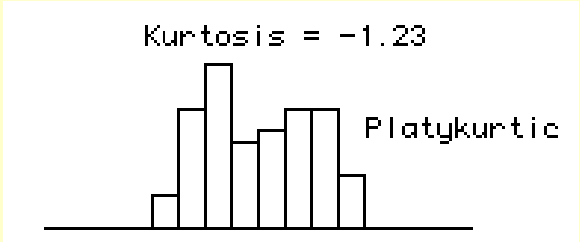


skew > 0 (right tail too long) skew < 0 (left tail too long)



kurtosis = 0

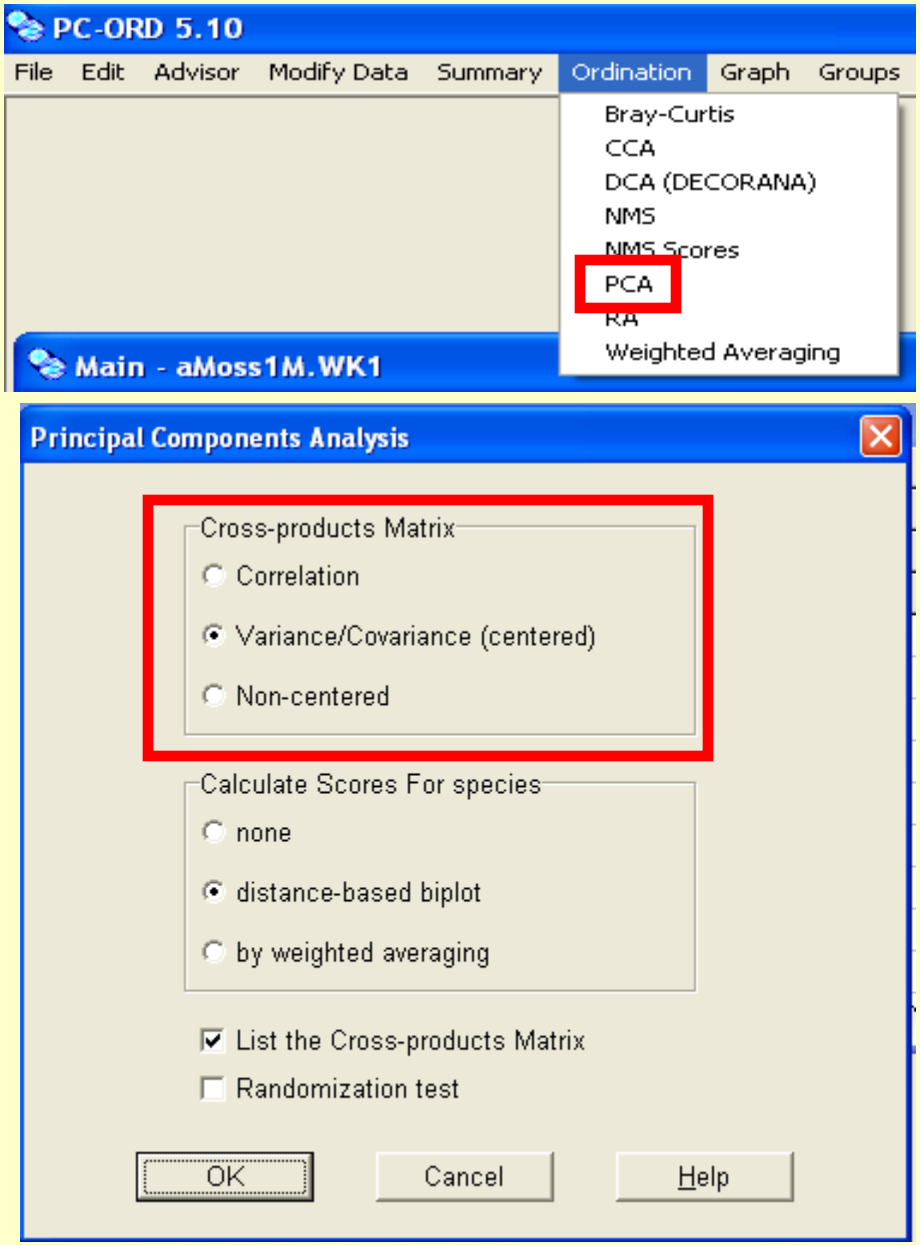
(normal data)



kurtosis > 0 (more peaky) kurtosis < 0 (less peaky)

- Rule of thumb (McCune & Grace 2002): **-1 < Skew < 1**

Principal Components (PCA) – PC-ORD



➢ Setup:

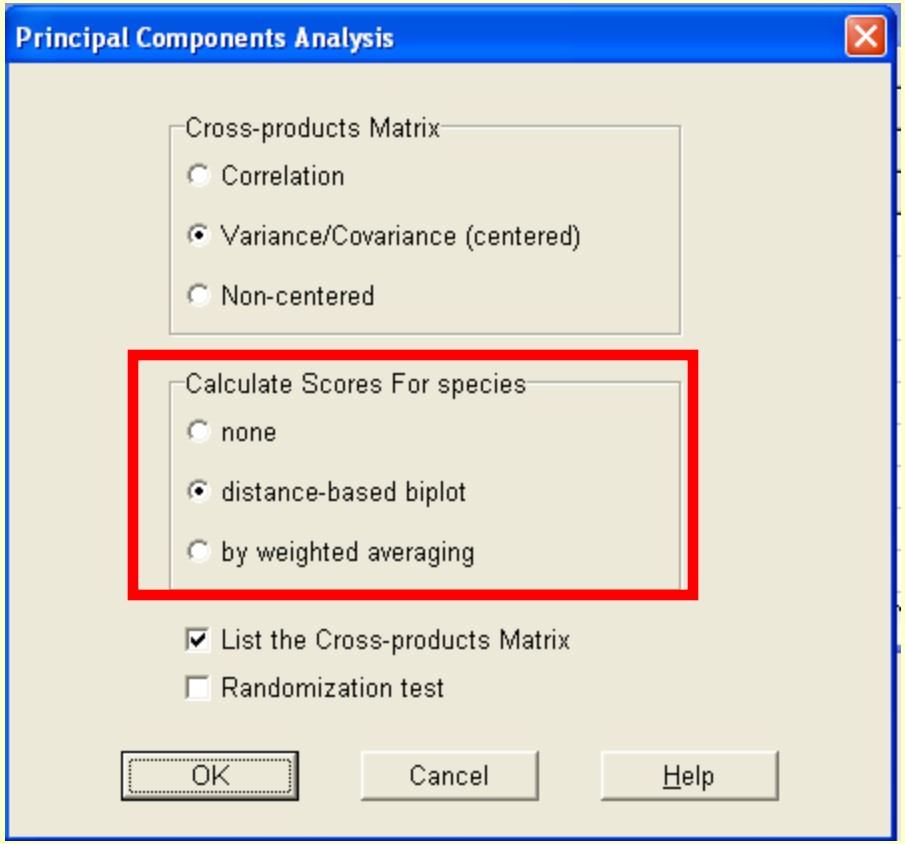
- PCA uses only Euclidean Distances (real metric)
- Matrix can be calculated in three ways:

Correlation: very susceptible to outliers (DO NOT USE)

Variance / Covariance: Less sensitive to outliers (USE)

Non-centered: Experimental (DO NOT USE)

Principal Components (PCA) – PC-ORD



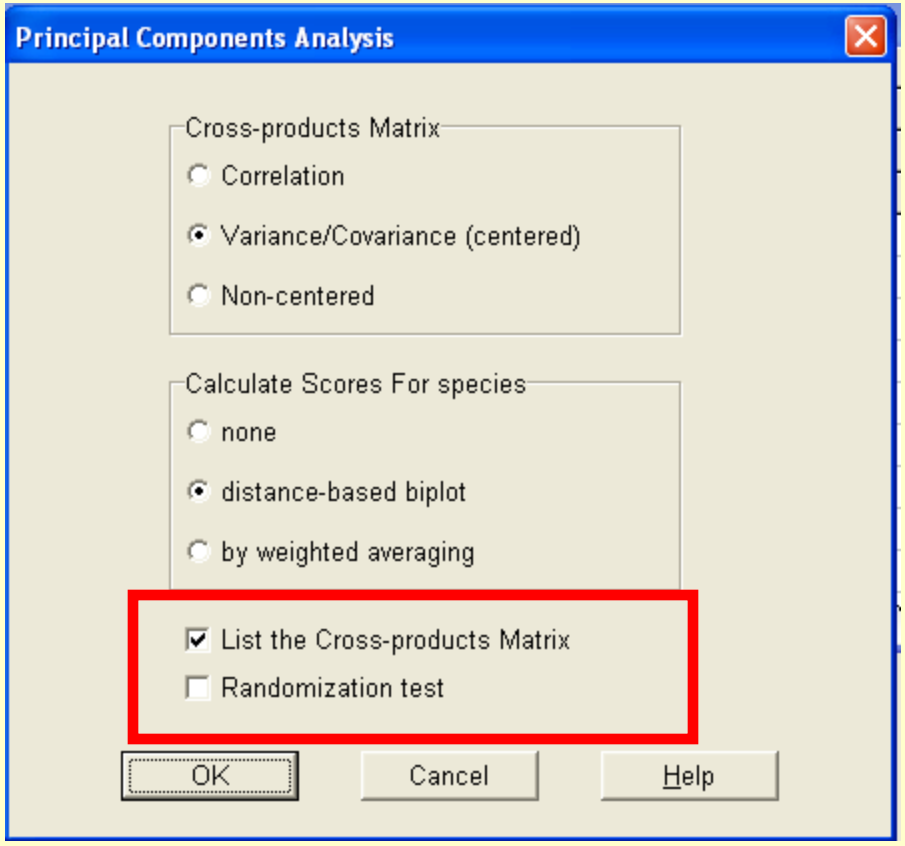
➢ Setup II:

• Scores for species can be calculated in two ways:

Distance-based: Relates species- samples to each axis – represents species as vectors from centroid (Standard) USE

Weighted Average: Species represented as points and outliers (DO NOT USE)

Principal Components (PCA) – PC-ORD



➢ Setup III:

• Output Options

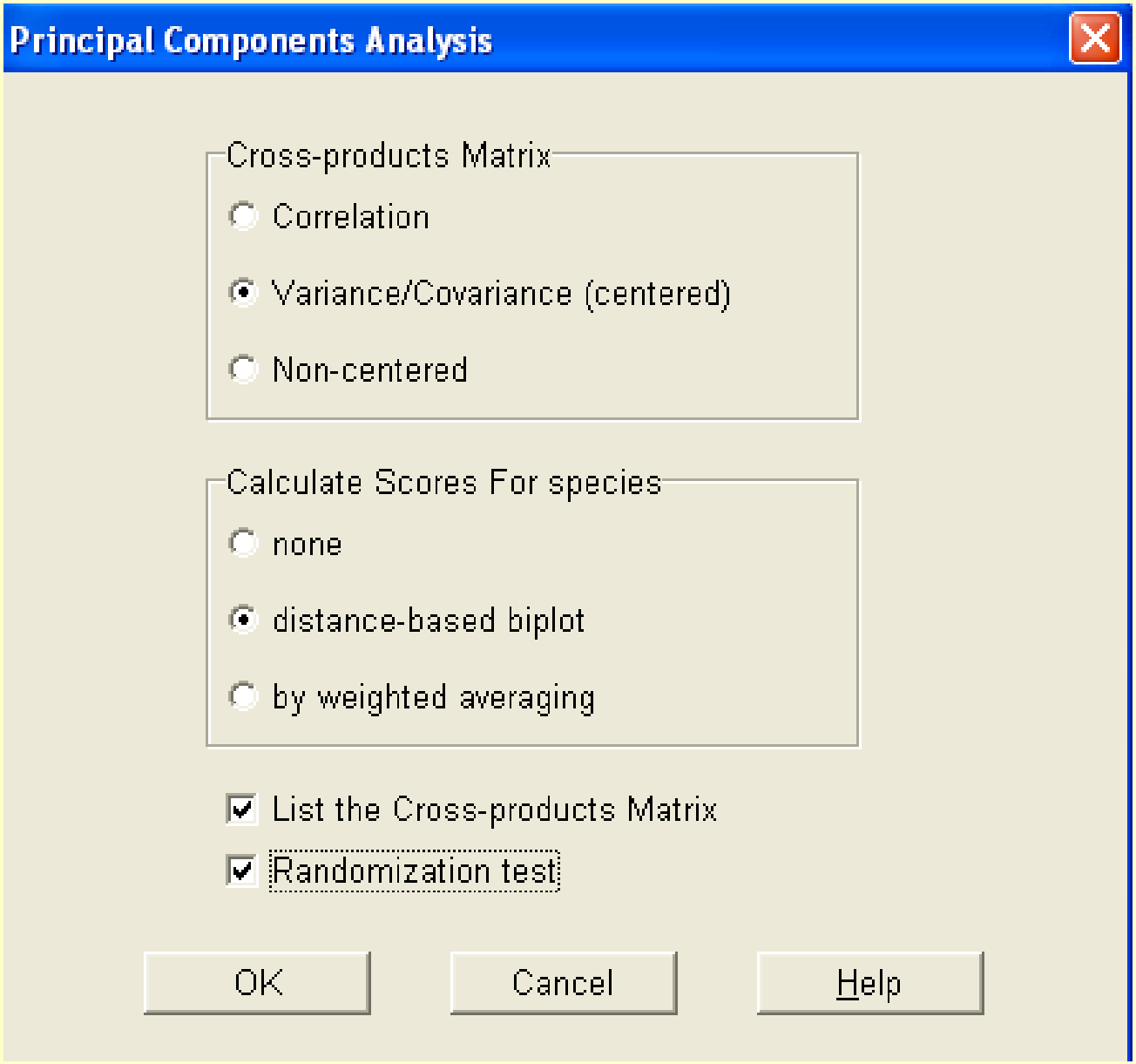
Cross-Product Matrix:
Shows pair-wise distances

(USE)

Randomization tests:
Use bootstrap to assess significance of the results

(USE)

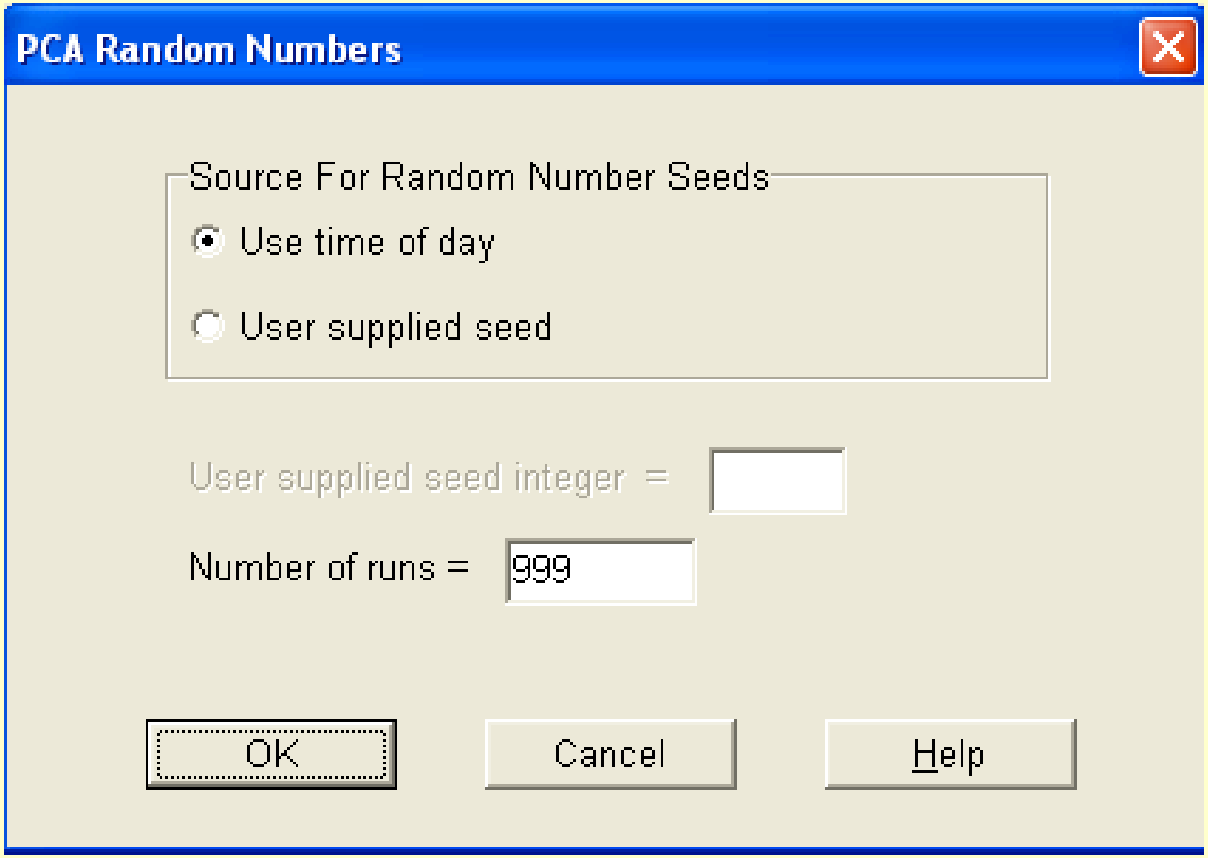
Principal Components (PCA) – PC-ORD



- Variance
- Distance-based
- Cross-product Matrix
- Randomization

Principal Components (PCA) – PC-ORD

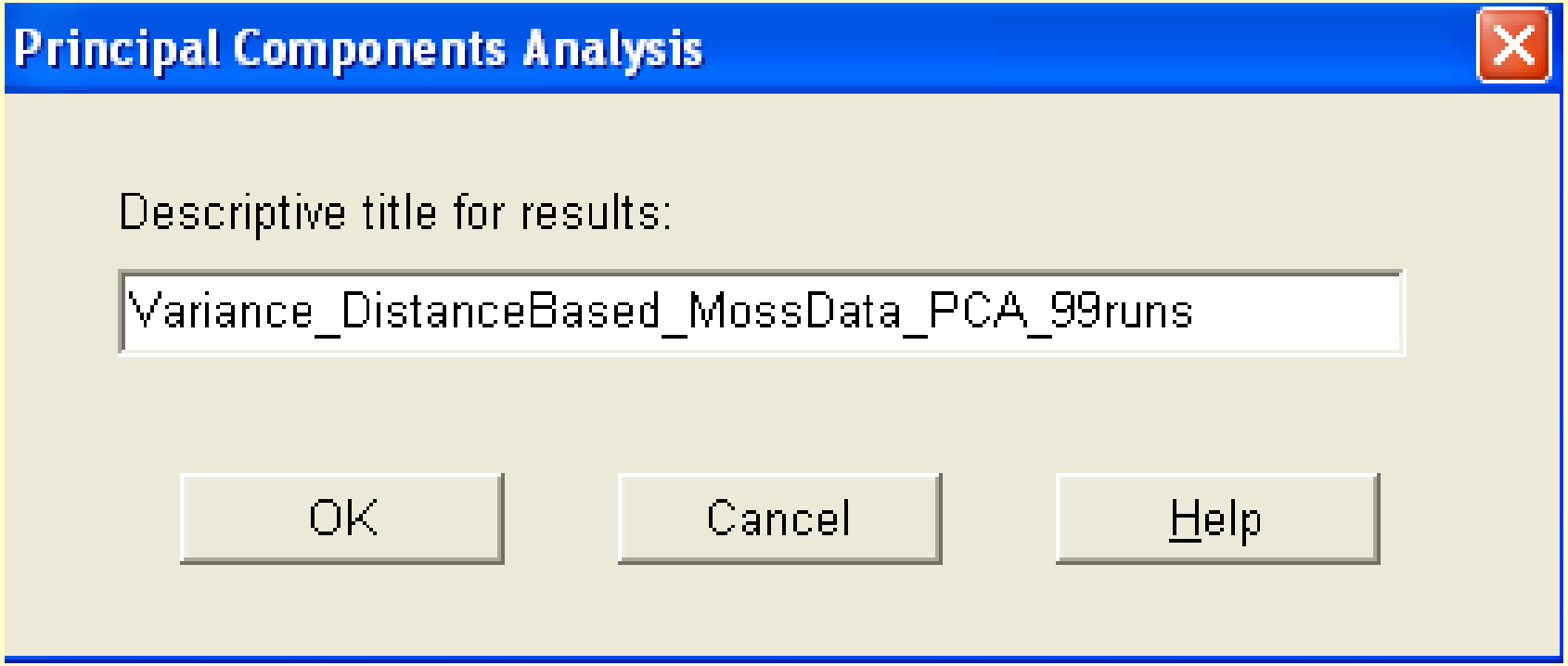
➢Setting up the Randomization Test:



- Seed: makes multiple tests comparable by using the same sequence of random numbers (supply seed)

- Runs: number of permutations used in the test (determines p value of statistic)

Principal Components (PCA) – PC-ORD



Enter descriptive explanation to document the analysis – this label will be added to results

Principal Components (PCA) – PC-ORD

- Results: Covariance matrix – species distances

Result - RESULT.TXT

```
********* PRINCIPAL COMPONENTS ANALYSIS -- stands   in species  space **********
PC-ORD, 5.10


Randomization test requested.          99 runs.
         2 = Seed for random number generator.


Variance_DistanceDabsed_MossData_PCA_99runs

Cross-products matrix is VARIANCE-COVARIANCE centered by species

CROSS-PRODUCTS MATRIX
---------------------
Ancu       0.7206D+03
Cloc      -0.7370D-01  0.5937D-03
Clad       0.2585D+00 -0.1607D-02  0.3663D+00
```

Principal Components (PCA) – PC-ORD

- Results: Eigenvalues - Variance explained (up to 10 axes)

Result - RESULT.TXT

VARIANCE EXTRACTED, FIRST 10 AXES

AXIS	Eigenvalue	% of Variance	Cum.% of Var.	Broken-stick Eigenvalue
1	11630.331	74.521	74.521	1404.362
2	1957.836	12.545	87.066	1092.227
3	914.362	5.859	92.925	936.159
4	580.305	3.718	96.643	832.114
5	292.397	1.874	98.516	754.080
6	123.200	0.789	99.306	691.653
7	68.209	0.437	99.743	639.630
8	19.814	0.127	99.870	595.039
9	9.192	0.059	99.929	556.022
10	4.004	0.026	99.954	521.341

Eigenvalues are proportional to variance explained

Broken-stick eigenvalues are produced by chance

Principal Components (PCA) – Example

- Results: Never explain 100% of variance (axes = variables)

Result - RESULT.TXT

VARIANCE EXTRACTED, FIRST 10 AXES

AXIS	Eigenvalue	% of Variance	Cum.% of Var.	Broken-stick Eigenvalue
1	11630.331	74.521	74.521	1404.362
2	1957.836	12.545	87.066	1092.227
3	914.362	5.859	92.925	936.159
4	580.305	3.718	96.643	832.114
5	292.397	1.874	98.516	754.080
6	123.200	0.789	99.306	691.653
7	68.209	0.437	99.743	639.630
8	19.814	0.127	99.870	595.039
9	9.192	0.059	99.929	556.022
10	4.004	0.026	99.954	521.341

Observed **Variance Explained** Expected

Principal Components (PCA) – PC-ORD

- Results: Species Loadings onto the PC Axes

Result - RESULT.TXT

FIRST 6 EIGENVECTORS, scaled to unit length.
These can be used as coordinates in a distance-based biplot,
where the distances among objects approximate their Euclidean
distances.

	Eigenvector					
species	1	2	3	4	5	6
Ancu	0.0864	0.2616	0.5844	0.5076	0.3160	-0.1714
Cloc	0.0001	-0.0001	-0.0002	-0.0001	-0.0003	0.0002
Clad	-0.0003	0.0027	0.0026	-0.0084	0.0153	0.0133
Clcr	0.0441	-0.0358	-0.0593	-0.0656	0.2207	-0.0049
Deab	0.0025	0.0097	-0.0182	0.0269	-0.0154	0.0048
Difu	-0.0003	0.0006	0.0007	-0.0015	-0.0058	-0.0011
Disc	-0.0003	0.0054	0.0030	-0.0371	0.0557	0.0282
Dita	0.0000	0.0001	0.0001	-0.0004	0.0021	0.0036
Euor	0.0045	0.1564	-0.0550	-0.6312	-0.0283	-0.3405
Frbo	0.0049	-0.0021	-0.0030	-0.0292	0.0252	-0.0553
Frni	-0.0016	0.0229	0.0485	-0.1844	0.0203	-0.4908

- Use the scaled eigenvectors

Principal Components (PCA) – PC-ORD

- Results: Randomization tests

Result - RESULT.TXT

BEGINNING RANDOMIZATIONS

RANDOMIZATION RESULTS

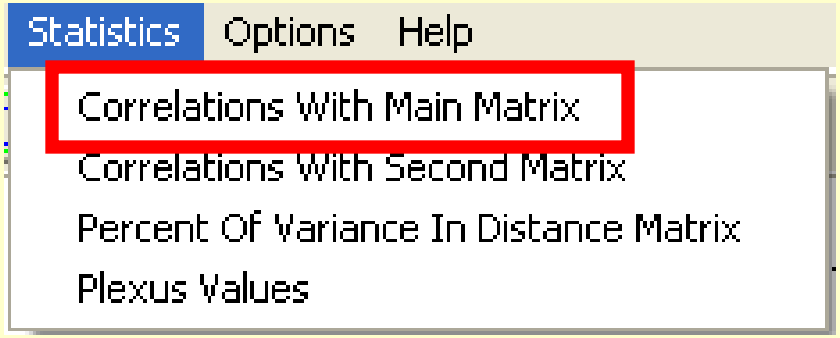
99 = number of randomizations

Axis	Eigenvalue from real data	Eigenvalues from randomizations: Minimum	Average	Maximum	p *
1	11630.	8570.4	9083.7	10419.	0.010000
2	1957.8	2514.7	3968.0	4697.2	1.000000
3	914.36	592.18	977.23	1598.7	0.730000
4	580.30	445.95	679.82	857.07	0.870000
5	292.40	180.90	432.49	586.93	0.930000
6	123.20	80.846	232.88	327.03	0.980000
7	68.209	59.470	116.65	180.76	0.970000
8	19.814	19.194	58.665	99.950	0.990000
9	9.1917	14.382	29.881	52.426	1.000000
10	4.0042	5.8316	10.980	19.344	1.000000

* p-value for an axis is (n+1)/(N+1), where n is the number of randomizations with an eigenvalue for that axis that is equal to or larger than the observed eigenvalue for that axis. N is the total number of randomizations.

Principal Components (PCA) – PC-ORD

- Results: Correlation with Axes



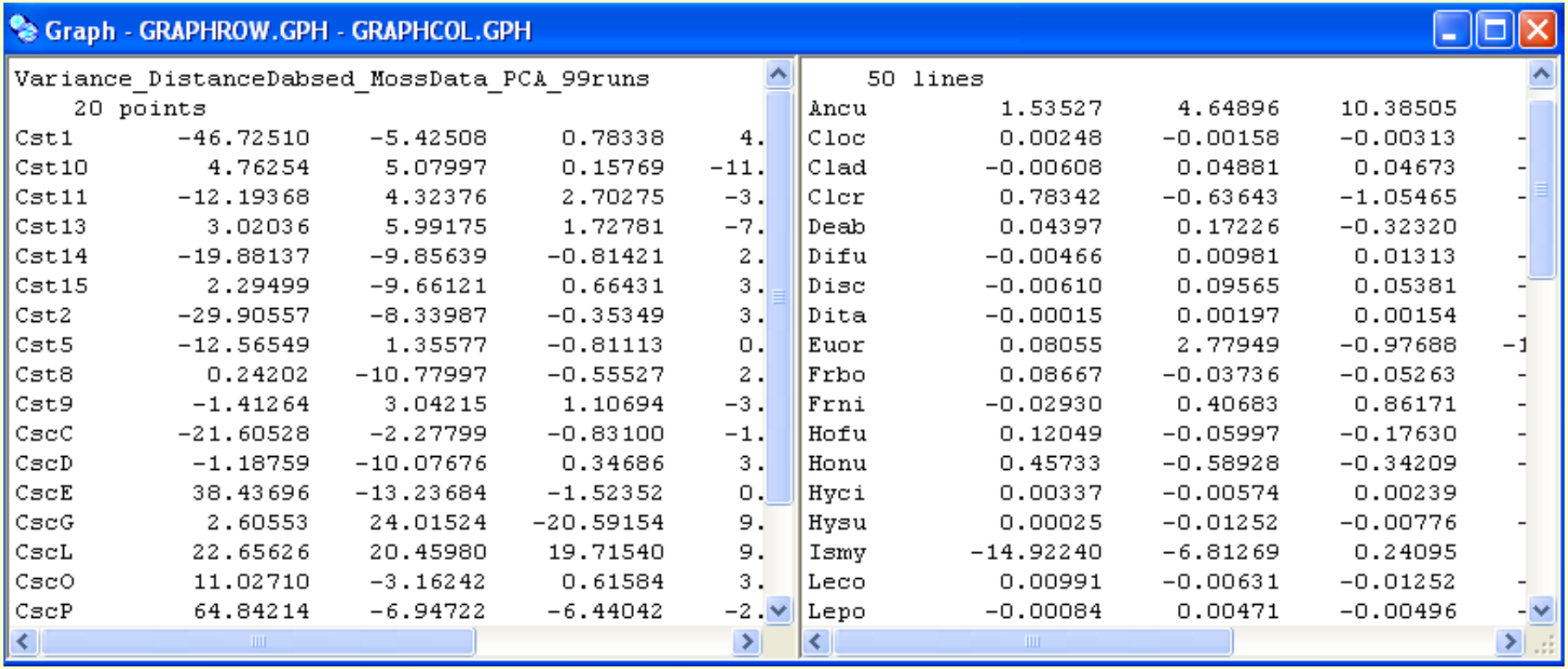
first matrix

Pearson and Kendall Correlations with Ordination Axes N= 20

Axis:	1			2			3		
	r	r-sq	tau	r	r-sq	tau	r	r-sq	tau
Ancu	.347	.120	.531	.431	.186	-.016	.658	.433	.263
Cloc	.617	.381	.316	-.161	.026	-.117	-.219	.048	-.283

Principal Components (PCA) – Example

- Results: Graphs



Graph - GRAPHROW.GPH - GRAPHCOL.GPH

Variance_DistanceDabsed_MossData_PCA_99runs

20 points

Cst1	-46.72510	-5.42508	0.78338	4.
Cst10	4.76254	5.07997	0.15769	-11.
Cst11	-12.19368	4.32376	2.70275	-3.
Cst13	3.02036	5.99175	1.72781	-7.
Cst14	-19.88137	-9.85639	-0.81421	2.
Cst15	2.29499	-9.66121	0.66431	3.
Cst2	-29.90557	-8.33987	-0.35349	3.
Cst5	-12.56549	1.35577	-0.81113	0.
Cst8	0.24202	-10.77997	-0.55527	2.
Cst9	-1.41264	3.04215	1.10694	-3.
CscC	-21.60528	-2.27799	-0.83100	-1.
CscD	-1.18759	-10.07676	0.34686	3.
CscE	38.43696	-13.23684	-1.52352	0.
CscG	2.60553	24.01524	-20.59154	9.
CscL	22.65626	20.45980	19.71540	9.
CscO	11.02710	-3.16242	0.61584	3.
CscP	64.84214	-6.94722	-6.44042	-2.

50 lines

Ancu	1.53527	4.64896	10.38505	
Cloc	0.00248	-0.00158	-0.00313	-
Clad	-0.00608	0.04881	0.04673	-
Clcr	0.78342	-0.63643	-1.05465	-
Deab	0.04397	0.17226	-0.32320	
Difu	-0.00466	0.00981	0.01313	-
Disc	-0.00610	0.09565	0.05381	-
Dita	-0.00015	0.00197	0.00154	-
Euor	0.08055	2.77949	-0.97688	-1
Frbo	0.08667	-0.03736	-0.05263	-
Frni	-0.02930	0.40683	0.86171	-
Hofu	0.12049	-0.05997	-0.17630	-
Honu	0.45733	-0.58928	-0.34209	-
Hyci	0.00337	-0.00574	0.00239	
Hysu	0.00025	-0.01252	-0.00776	-
Ismy	-14.92240	-6.81269	0.24095	
Leco	0.00991	-0.00631	-0.01252	-
Lepo	-0.00084	0.00471	-0.00496	-

Samples: points

Species: vectors

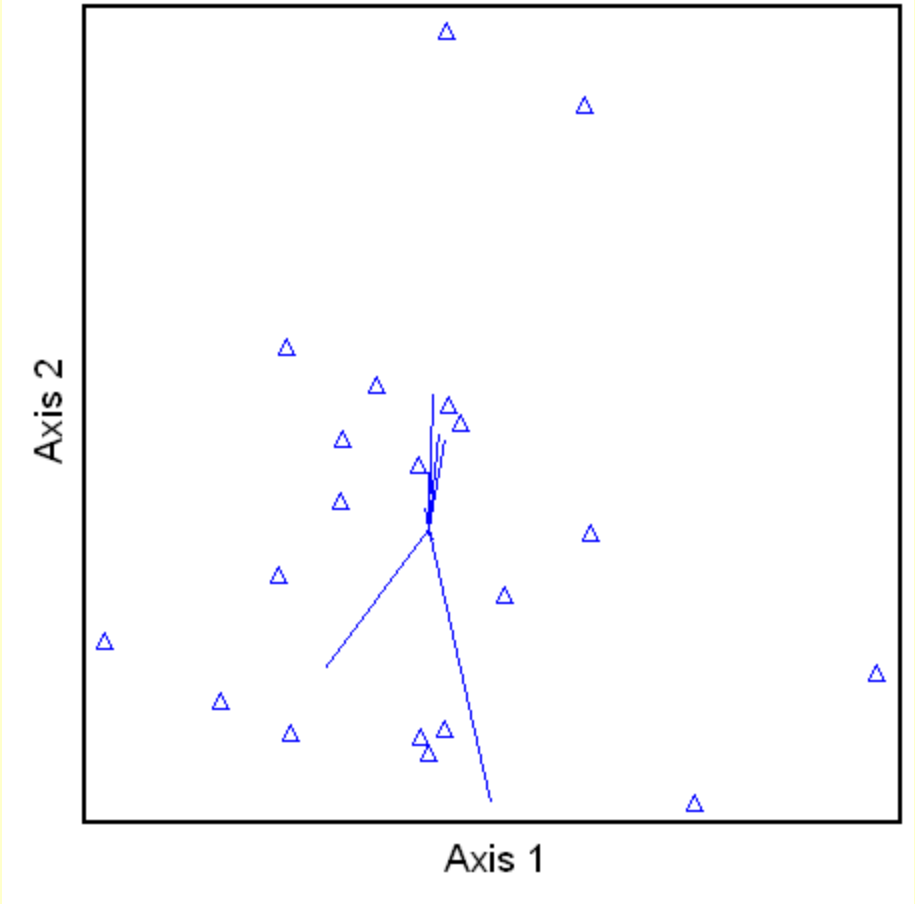
Principal Components (PCA) – PC-ORD

- Results: Graphs

PC ORD recommends displaying species as vectors / samples as points

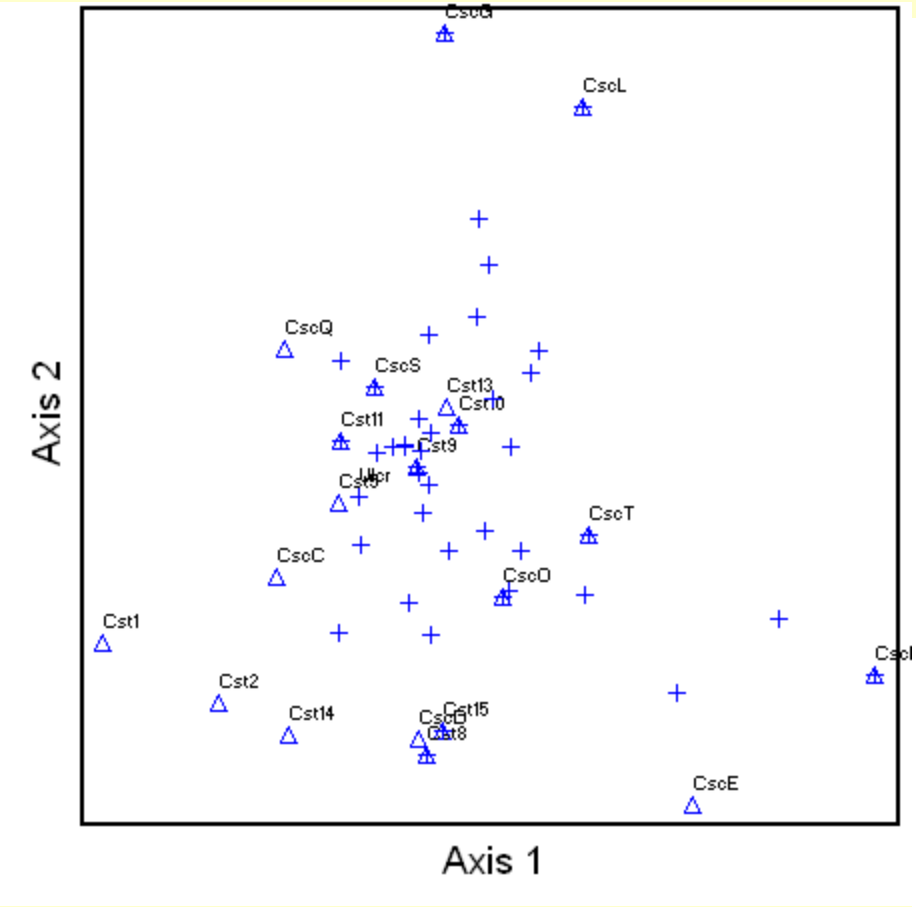
Samples: points

Species: vectors

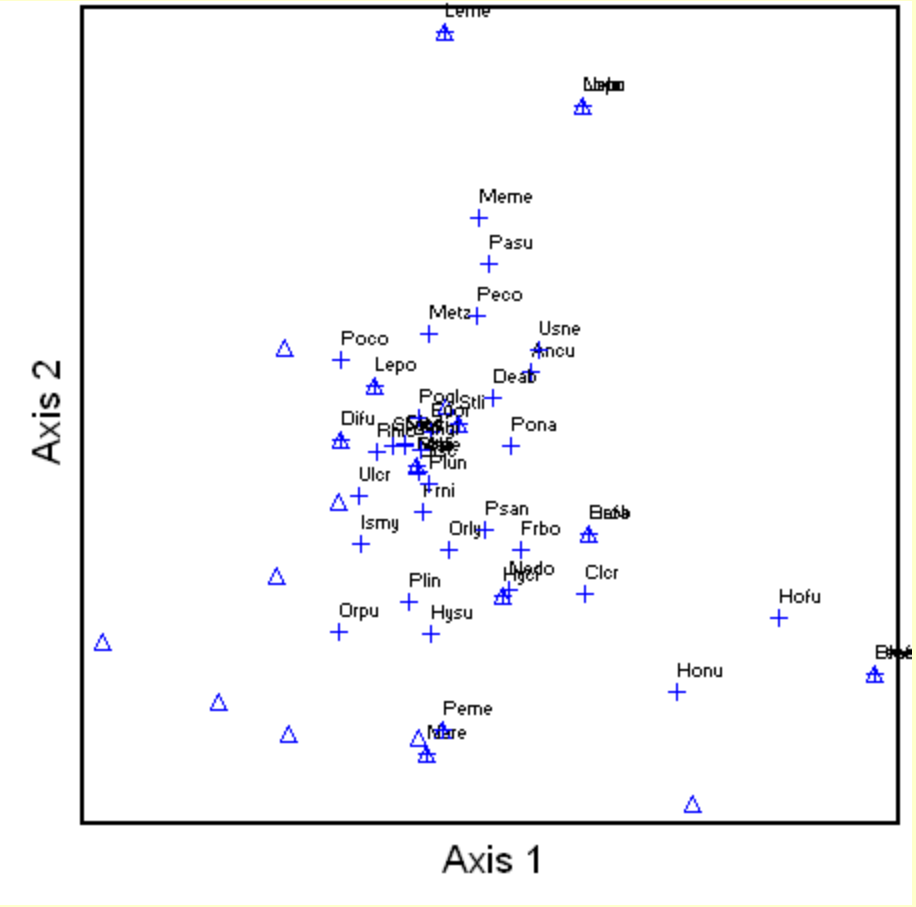


Principal Components (PCA) – PC-ORD

- Results: Graphs



Samples Labeled



Species Labeled

Principal Components (PCA) – PC-ORD

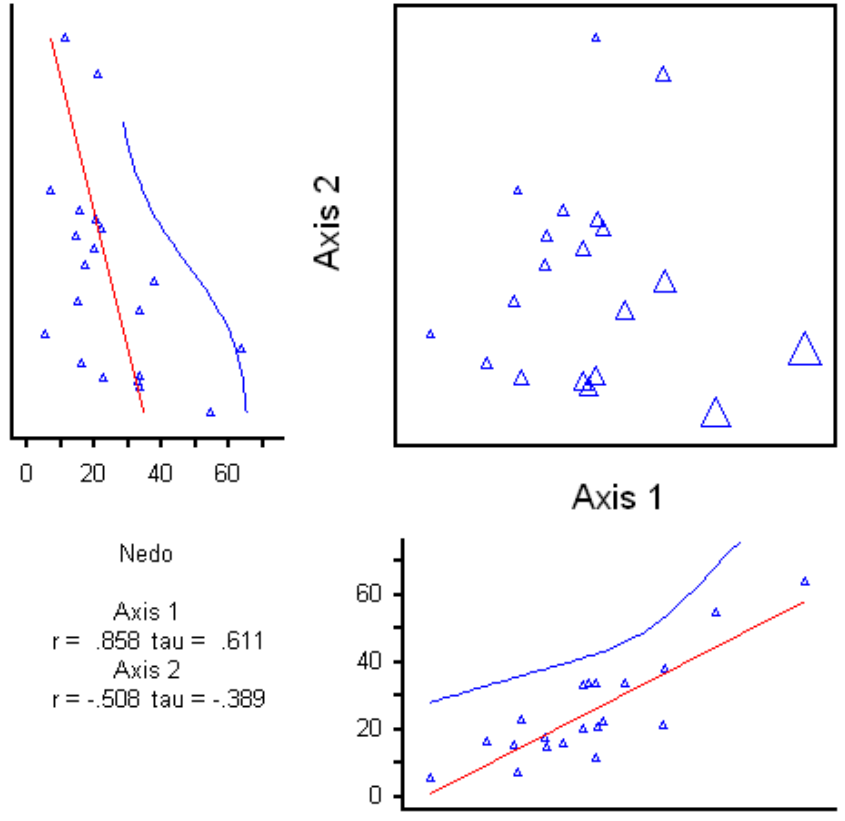
- Results: Species Response Graphs

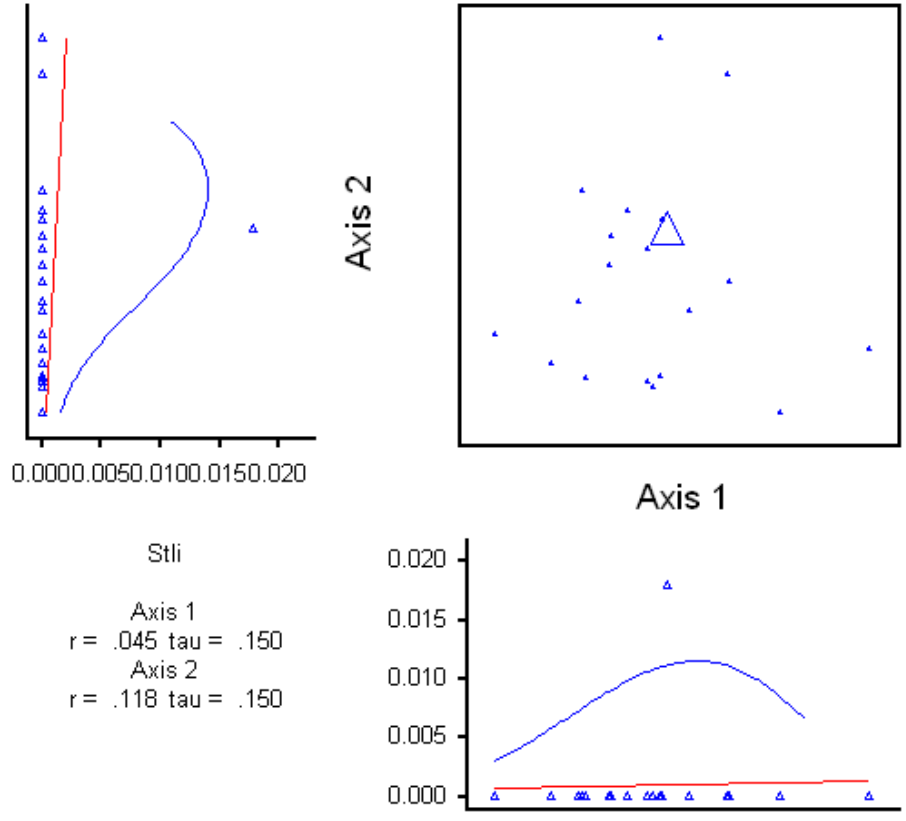
Loadings NEDO

Axis 1: +0.51 Axis 2: -0.74

Loadings STLI

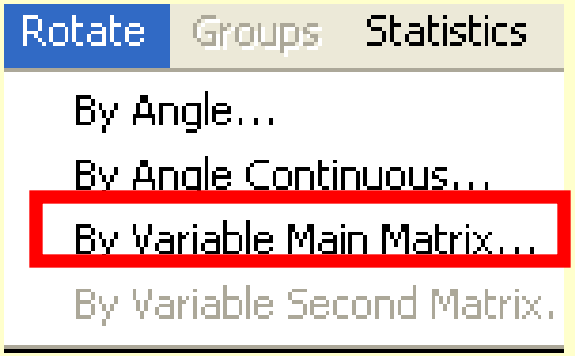
Axis 1: 0.00 Axis 2: 0.00





Principal Components (PCA) – PC-ORD

- Rotation: Highlights certain patterns. Report in results

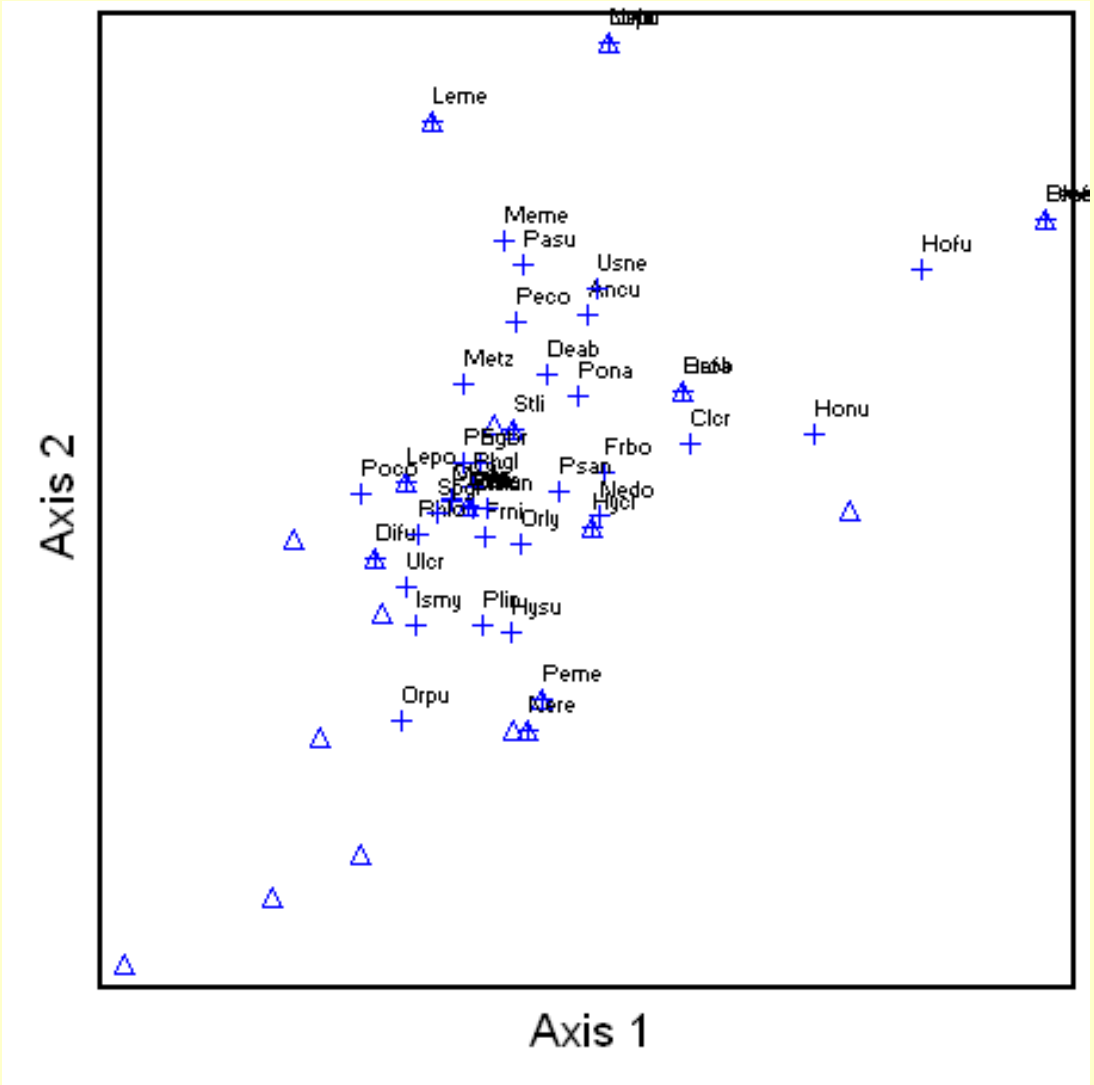


NEDO Axes Correlations

Axis 1: +0.51

Axis 2: -0.74

- Rotation by NEDO Stretch plot along direction of most variation for species



Principal Components (PCA) – Reporting

- What type of cross-correlation matrix you used?

  Use covariance matrix. Use euclidean distance

- If used with community data, justify using this linear model for species data?

  Were assumptions of linearity / normality met?

- How many axes were interpreted, and what proportion of variance was explained by these axes?

  Describe the axes – and the individual / cumulative variance

- Principal eigenvectors - Test of significance?

  Not necessary, but an option using randomization tests

- Rotation of the solution? Use of interpretation aids?

  Explain overlays and correlations of variables with axes