# Unconstrained ordination (PCA, CA, and NMDS)

## Multivariate Stats Seminar, UCSB 2021

**Robert Fitch**

*Based on:*
*UMass Landscape Ecology Lab*
*McGarigal et al 2000*
*Wikipedia*

# Overview

- **Ordination-** simplify complex data by organizing samples along a gradient of interrelated variables, *account for major patterns and minimize information loss*

- **Unconstrained**- determines patterns without relationships to variables outside of the data set. Analyzing a single set of data, *does not determine mechanism.*

- This is a family of related technique's, method used needs to be critically evaluated for use and desired outcome

# Principal Components Analysis

- Example data sets:
  - plant community composition w/ environmental factors
  - Abundance of animals w/ plant structure data
  - Animal community data w/ landscape features
  - Species data and niche parameters

# Principal Components Analysis

- Important Characteristics:
  - Continuous gradients
  - Single set of variables (i.e., no inherent dep. Or indep. variables
  - Emphasizes variation among samples
  - Reduces dimensionality from many variables to few composites
  - Summarizes redundancy and reduce noise
  - Either all categorical or all continuous
  - Full rank (no NA's, and more rows than columns preferred)

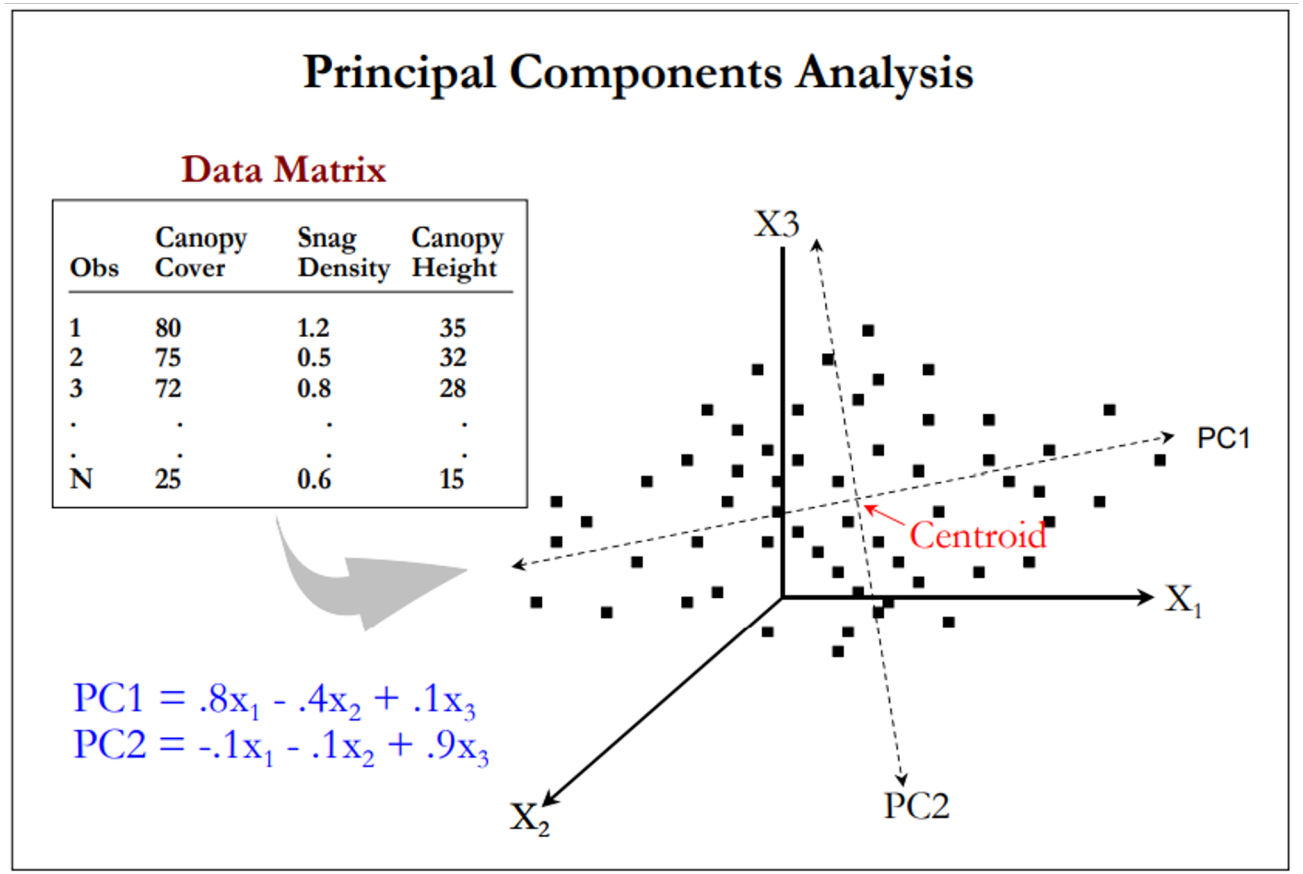
# Principal Components Analysis- Overview

- Condensing information into its principle components
- Creating new axes of composite variables
  - Weighted linear combinations -> maximize variation
- Ecological meaning (the importance of each component is defined by the samples)
- Generates fully uncorrelated variables that can be used in multiple regression, and MANANOVA

# Principal Components Analysis- How it's Done

Deriving the PCs

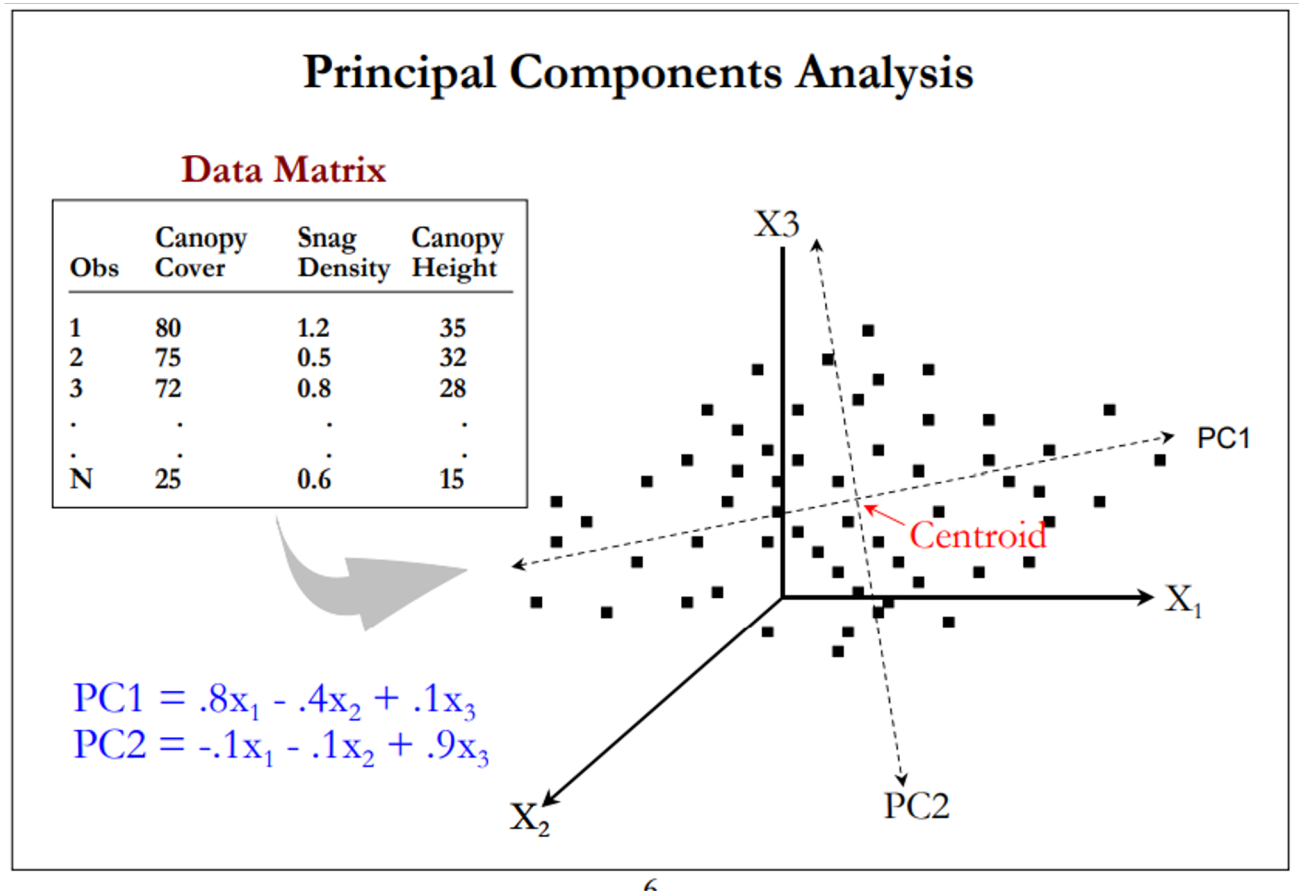
- 1st PC is drawn through the longest portion of the data cloud
- 2nd is drawn orthogonally
- Each subsequent PC…



# Principal Components Analysis- How it's Done

Deriving the PCs

1. Matrix
2. Eigen values
3. Eigen vectors



# Principal Components Analysis- How it's Done

Deriving the PCs

1. Matrix

- Two options- correlation matrix or the covariance
  - Correlation is almost always most appropriate (equal weight to variances)
  - Will give different results

# Principal Components Analysis- How it's Done

Deriving the PCs

2. Eigen values

What is happening?

-Taking a matrix and conducting a linear transformation into a vector

# Principal Components Analysis- How it's Done

Deriving the PCs

2. Eigen values

Now consider the linear transformation of $n$-dimensional vectors defined by an $n$ by $n$ matrix $A$,

$$A\mathbf{v} = \mathbf{w},$$

or

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

where, for each row,

$$w_i = A_{i1}v_1 + A_{i2}v_2 + \cdots + A_{in}v_n = \sum_{j=1}^{n} A_{ij}v_j.$$

Correlation matrix

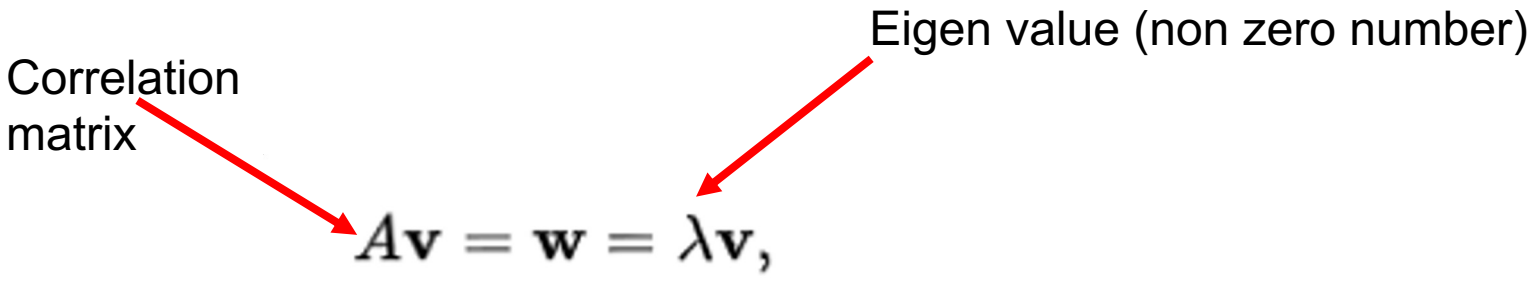
Eigen value (non zero, real number)

$$A\mathbf{v} = \mathbf{w} = \lambda\mathbf{v},$$

# Principal Components Analysis- How it's Done

Deriving the PCs

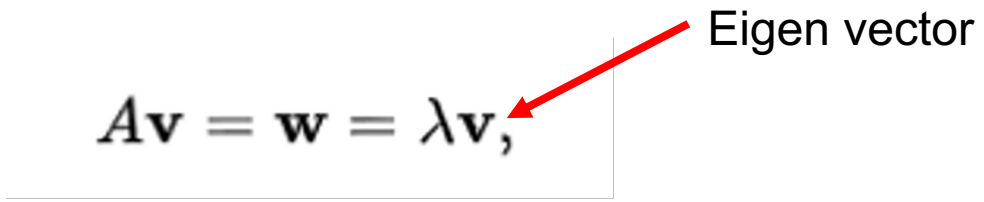
2. Eigen values
   - Scale factor, the magnitude



# Principal Components Analysis- How it's Done

Deriving the PCs

3. Eigen vector
   - linear transformation of *A*
   - **How are the Eigen values calculated?**

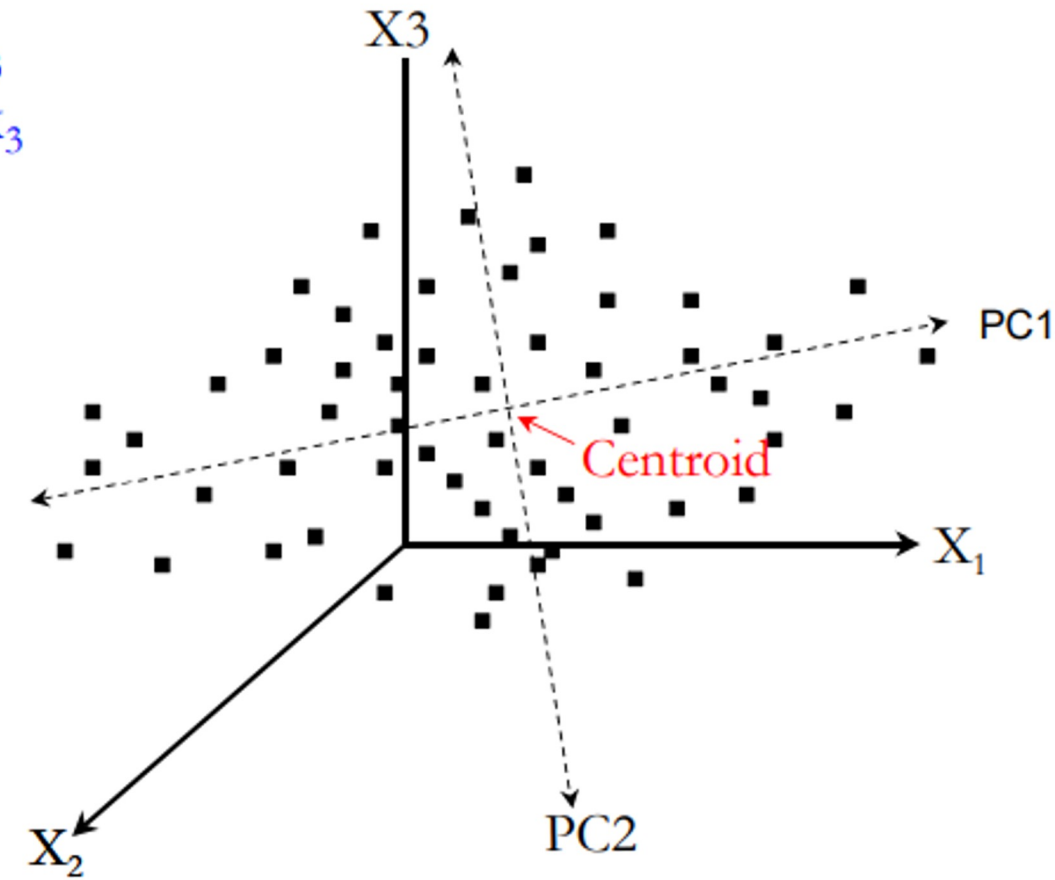



Characteristic Equation: $\left|R - \lambda_i I\right| v_i = 0$

Where: $\lambda_i$ = eigenvalue corresponding to the $i^{th}$ PC
$v_i$ = eigenvector associated with the $i^{th}$ PC

# Eigenvectors:

$PC1 = .8x_1 - .4x_2 + .1x_3$
$PC2 = -.1x_1 - .1x_2 + .9x_3$
...



# Principal Components Analysis- How it's Done

Deriving the PCs

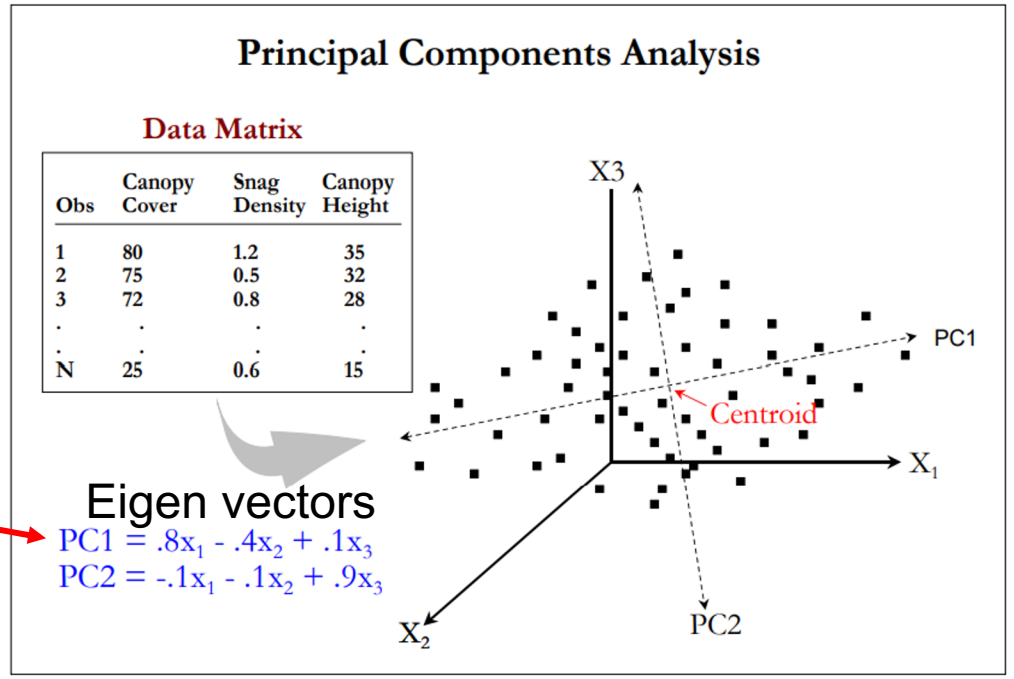
1. Matrix
   - Correlation matrix
2. Eigen values
   - Variance explained by the PC (larger values have more explanatory power
3. Eigen vectors
   - The importance of original variables in the PC, the weights of the variables in the linear equations

# Principal Components Analysis- How it's Done

Deriving the PCs

**PC Scores?**

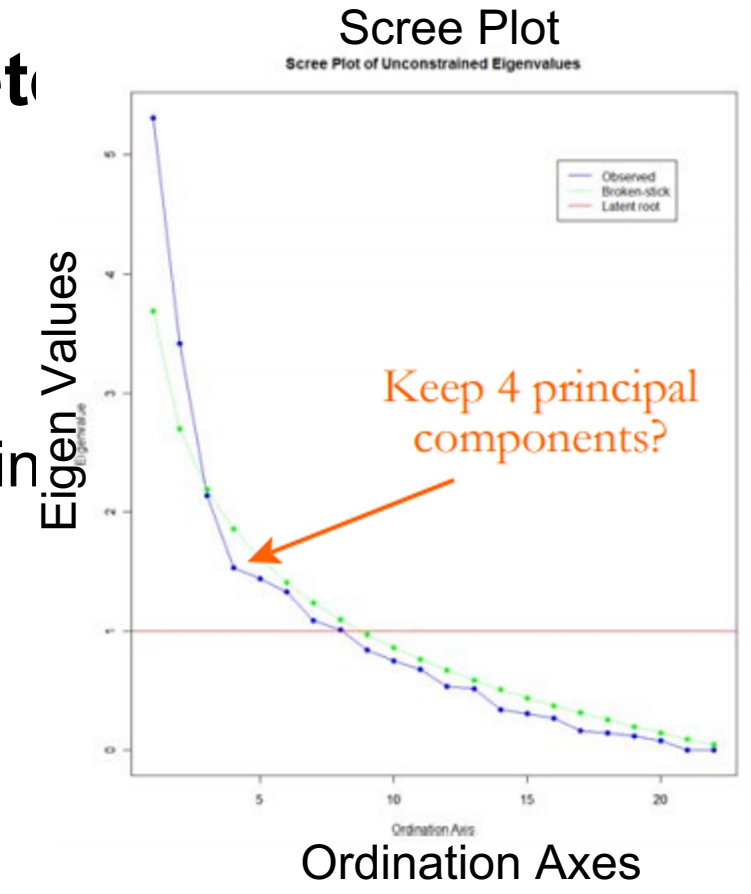
- Location of the samples on the PC
  - New data for subsequent analyses!
  - Sum of



# Principal Components Analysis- Assessing the Importance of PCs

**How many PCs should be kept and interpret**

1. Latent Root Criterion
   - Drop any with Eigen values <1
2. Scree Plot/Broken Stick
   - Keep PC before becoming linear after sharp declin
3. Percent Variance Criterion
   - Want first 3 PCs to explain 70% of variation
4. Significance Test
   - Jackknife, bootstrapping



# Principal Components Analysis- Interpretation

**Structure Coefficients**

1. Principal Component Structure (also 'loadings'):

$$s_{ij} = v_{i(j)}\sqrt{\lambda_i}$$

$S_{ij}$ = correlation between the $i^{th}$ PC and the $j^{th}$ variable

$v_{i(j)}$ = eigenvector element of the $j^{th}$ variable in the $i^{th}$ PC *derived from correlation matrix*

$\lambda_i$ = $i^{th}$ eigenvalue ($i^{th}$ PC)

- Component "loadings"- correlation of variables to PCs
  - As they approach 1, carry similar data as variables
- The square of the loading is the %of total variance accounted for by that component

# Principal Components Analysis- Interpretation

**Structure Coefficients**

- Rules of Thumb "significance":
  - Depends on sample size…

Significance of Structure Correlations:

- $s_{ij} > \pm 0.30$ significant, when N > 50
- $s_{ij} > \pm 0.26$ significant, when N = 100
- $s_{ij} > \pm 0.18$ significant, when N = 200
- $s_{ij} > \pm 0.15$ significant, when N = 300

- Can use jackknifing and bootstrapping for testing

# PCA: Interpreting the Principal Components

## Interpreting Structure Correlations:

- Highlight the highest 'significant' loading for each variable (red)
- Highlight other significant loadings (blue)

Structure Correlations:

| | PC1 | PC2 | PC3 |
|---|---|---|---|
| FORB | 0.39 | 0.473 | -0.383 |
| GRASS | | | -0.324 |
| FERN | -0.516 | -0.601 | |
| SALAL | -0.447 | -0.477 | -0.444 |
| GRAPE | -0.451 | | 0.406 |
| SALMON | 0.625 | | -0.33 |
| CURRANT | -0.461 | 0.38 | |
| HUCKLE | -0.644 | | |
| THIMBLE | | | |
| DEVIL | -0.374 | | 0.57 |
| VINE | | -0.789 | |
| ELDER | | | |
| HAZEL | | -0.492 | 0.486 |
| PLUM | -0.48 | -0.561 | -0.534 |
| OCEAN | | | |
| ALDER | 0.786 | | |
| MAPLE | 0.303 | -0.389 | 0.414 |
| FIR | -0.575 | -0.618 | |
| HEMLOCK | -0.373 | 0.483 | -0.311 |
| CEDAR | -0.591 | 0.458 | |
| HARDWD | 0.815 | | |
| CONIFER | -0.884 | | |

PC2 Gradient

Vine
Fir
Fern
Hazel
Plum/Salal

Hemlock
Cedar
Forb

"Seral Stage"

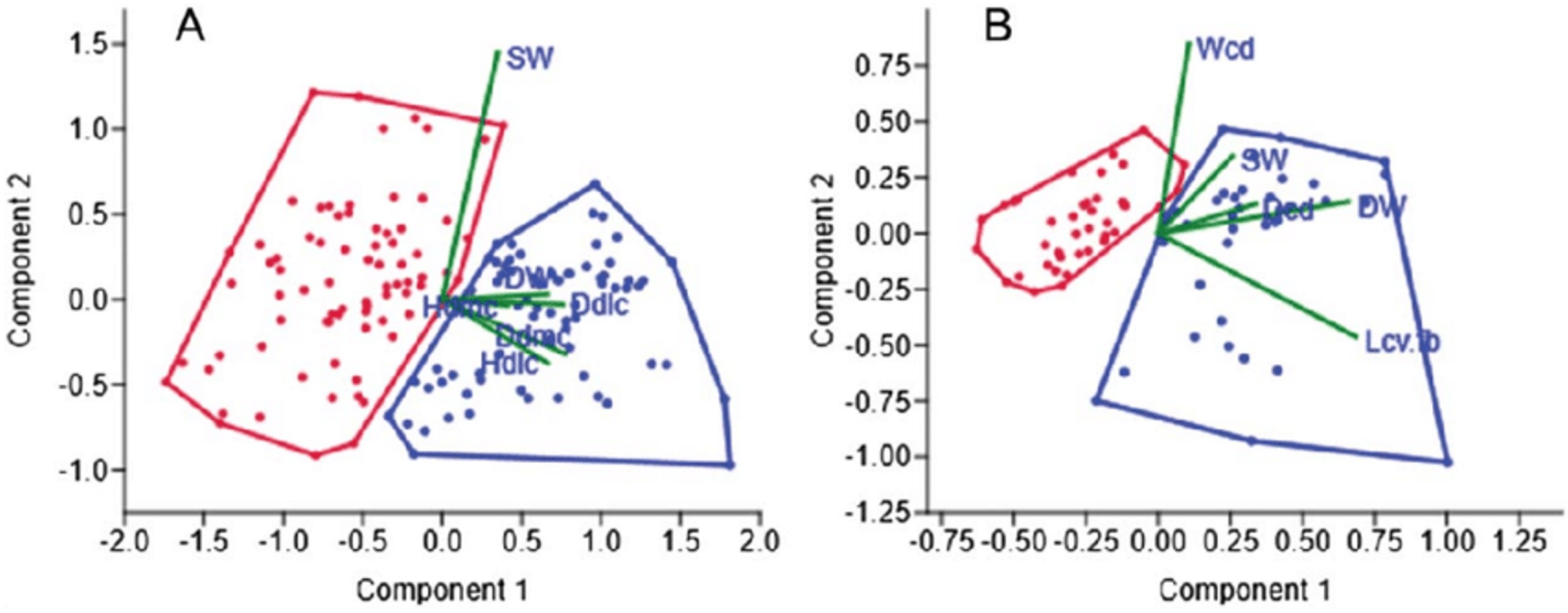
# Principal Components Analysis- Interpretation

**Communality**

$$c_j = \sum_{i=1}^{P} s_{ij}^2$$

- Highly recommended to use
- The proportion of the variable's variance explained by PC
  - 0 < c < 1
  - "how well the original variables are accounted for"

# Principal Components Analysis- Interpretation- Graphing PC Scores



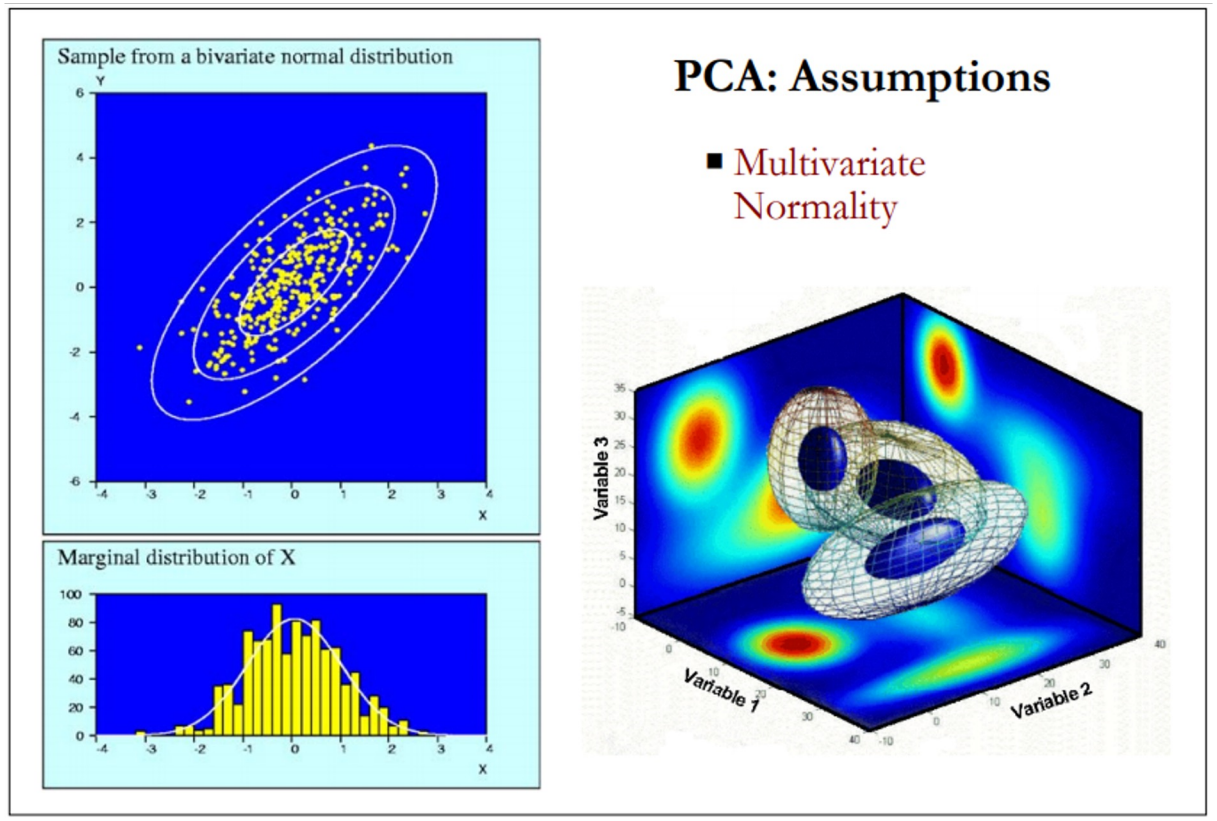
# Principal Components Analysis- Assumptions

1. Multivariate normality
2. Indep. And random samples
3. Linearity
4. Outliers*
5. Sample Size*

Your goal?- Descriptive vs inferential?

# Principal Components Analysis- Assumptions

1. Multivariate normality
   - Data are hyper-ellipsoid and the density of the data is anchored in a centroid
   - Diagnostics?- QQ plots, boxplots, histograms
     - Each variable **and** the PC scores
     - Univariate normal does ensure MVN
     - Checking for skewness, kurtosis, normal dist. For each variable



# Principal Components Analysis- Assumptions

2. Indep. And random samples (addressed by conducting an appropriate study design, random stratified sampling)

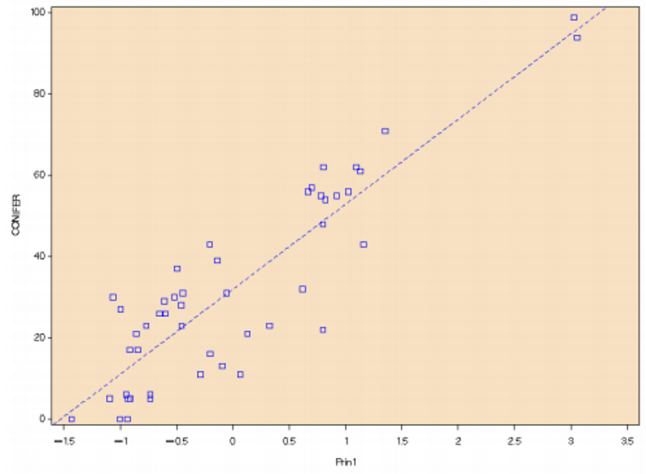
# Principal Components Analysis- Assumptions

3. Linearity
   - The variables change linearly across the gradients
   - Scatter plots of variables vs PC scores and PC scores for each PC

**PCA: Assumptions**

Linearity – Diagnostics:

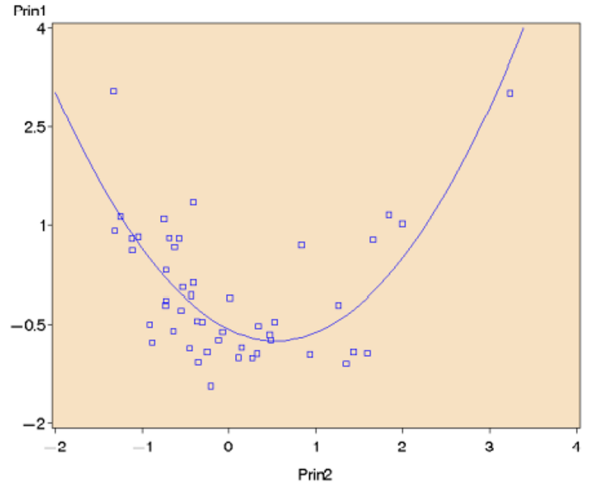
(C) Scatter plots of *variables* vs. principal component (PC) scores.




**PCA: Assumptions**

Linearity – Diagnostics:

(B) Scatter plots of *principal component* (PC) scores.



# Principal Components Analysis- Assumptions

4. Outliers*
   - Can have large effects on analysis, and should be removed >> **data screening**

# Principal Components Analysis- Assumptions

5. Sample Size*

- General Rules of Thumb:
    1. More sample than variables
    2. Should always be large enough to adequately describe your study/questions
    3. Specific calculation: N > 3 * P
    4. Eliminate unimportant variables

# PCA and Limitations

- Long/large gradients can result in patterns being distorted
- Best used in narrow range
- Could miss information if dismissing PCs
- Redundant variables increase variance explained but could be totally ignorant of the data
- Not recommended for community data

# PCA Terminology Summary

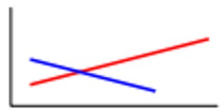
**Principal Components Analysis**

Review:

| | |
|---|---|
| ■ Eigenvalues... | Variances of PC's. |
| ■ Eigenvectors... | Variable weights in PC linear combinations. |
| ■ Structure coefficients (loadings)... | Correlations between original variables and PC's. |
| ■ Principal component scores... | Location of samples on PC's. |
| ■ Final communalities... | % of variance in original variables explained by retained PCs. |

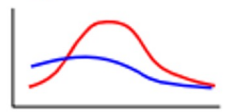
# Other Unconstrained Ordination Techniques

Linear



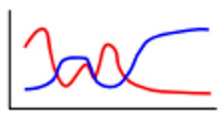
- Factor analysis (FA)
- Multidimensional scaling (MDS/PCO)
- **ML**-Unconstrained linear ordination (ULO)

Quadratic



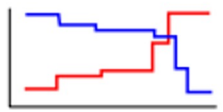
- Correspondence analysis (CA & DCA)
- **ML**-Unconstrained quadratic ordination (UQO)

Smooth



- **ML**-Unconstrained additive ordination (UAO)

Nonlinear



- Nonmetric multidimensional scaling (NMDS)

# Correspondence Analysis (CA)

Overview:

- Simultaneous ordination of samples to maximize correlation between samples and species
  - Still uses Eigen values/vectors
  - PCs are a linear combination of variables
  - Except uses **two sets** one calculated from the species and one for the sample (not Euclidean, uses Chi-square distances- inertia)

# Correspondence Analysis (CA)

Overview:

- Estimates the optima if X has a uniform and symmetric distribution.
- Unbiased estimate

$$w_k = \frac{\sum_{i=1}^{n} y_{ik} x_i}{\sum_{i=1}^{n} y_{ik}}$$

$y_{ik}$ = abundance of species $k$ at site $i$.

$x_i$ = value of the environmental variable at site $i$.

**Correspondence Analysis (CA)**

*Eigenanalysis Algorithm*

Find Eigenvalues: $|S - \lambda I| = 0$

Where: S = covariance matrix (as below)
λ = vector of eigenvalue solutions
I = identity matrix

$S$ is covariance matrix as in PCA except that cross-products are weighted by reciprocals of square roots of sample and species totals.

$$A = \begin{array}{c} \text{Obs} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \text{Species} \\ Y_1 \quad Y_2 \quad Y_3 \end{array} \begin{bmatrix} a_{ij} \end{bmatrix}_{(nxp)}$$

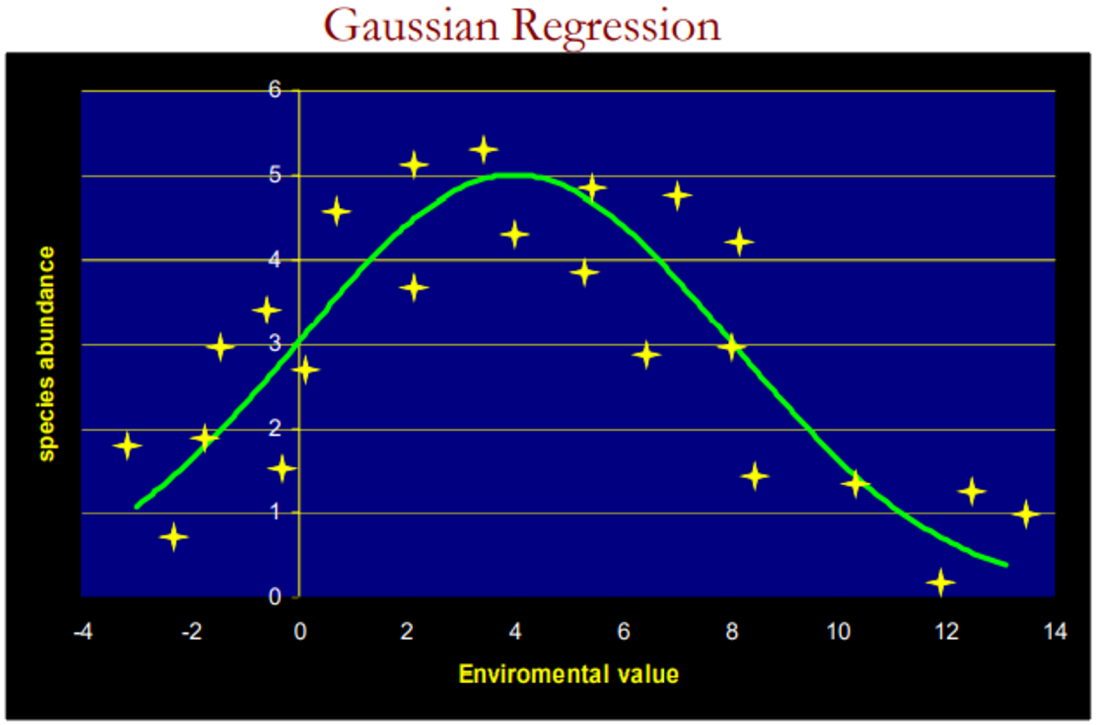
$$B = \begin{array}{c} \text{Obs} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \text{Species} \\ Y_1 \quad Y_2 \quad Y_3 \end{array} \begin{bmatrix} b_{ij} = \frac{a_{ij}}{\sqrt{a_{i+} \cdot a_{+j}}} \end{bmatrix}_{(nxp)}$$

$$S_{(nxn)} = BB' = V^{1/2} AWA' V^{1/2}$$

# Correspondence Analysis (CA)

Assumptions:

- Species are unimodal, optima are equal, have equal tolerances, and are evenly spaced



# Correspondence Analysis (CA)

**Limitations:**

- Species' responses must be unimodal and similarly scaled…
- Extreme values are compressed towards center (compression)
- Ordination is done sequentially appearing as an arch (no ecological basis) – arching effect
  - Can resolve issue of compression and arch by detrending CA, however methods are criticized…
- Chi square distance is heavily criticized for community data

# NMDS

**Non-metric multidimensional scaling**

- Eigen analysis but of a dissimilarity matrix
- No linear assumptions (uses rank order)

# NMDS

**Algorithm**

- Searches for best positions of k dimensions (axes) in order to minimize stress
  - Stress = departure from monotonicity (only increasing/decreasing, preserving order) i.e., representation of original data
  - Euclidean distances

# NMDS

**Algorithm**

1. Calculates dissimilarity matrix (Bray-Curtis)
2. Assigns random configuration (of axes and samples in p-space)
3. Calculate Euclidean distances, rank elements
4. Calculate stress (goodness-of-fit)
5. Reconfigure to minimize stress until stable solution is reached

# NMDS

**Checking validity**

1. # of dimensions
2. Fit based R2
3. Influential samples
4. Stable solution

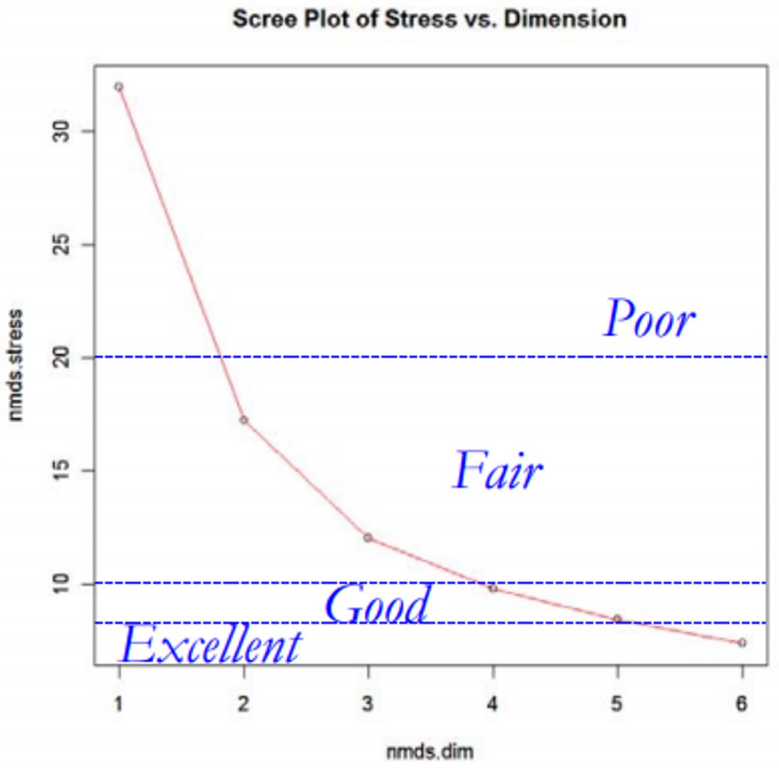
# NMDS

## 1. Choosing #of Dimensions

- Changes solutions! (not unique)

- Low stress (in %) = good fit of ordination configuration to original dissimilarities.

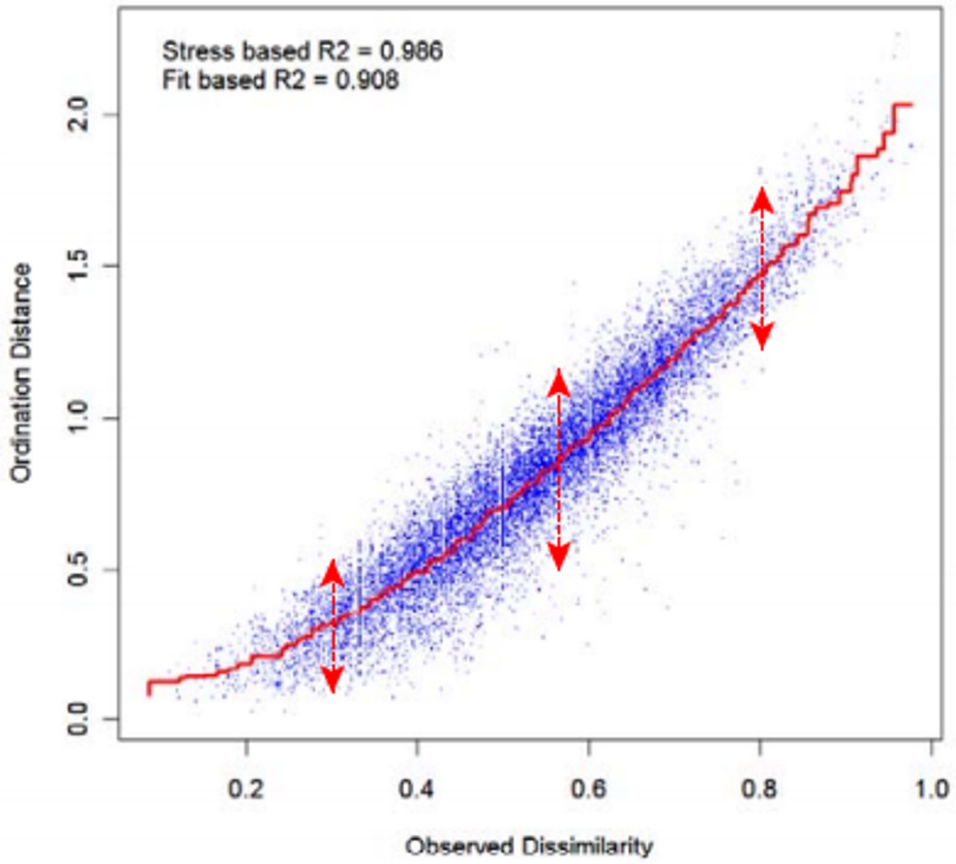
| | |
|---|---|
| <5% | Excellent representation with no risk of misinterpretation |
| 5-10% | Good ordination with no real risk of false inferences |
| 10-20% | Fair ordination; need cautious interpretation |
| >20% | Unreliable |



# NMDS

## 2. Fit Based R2

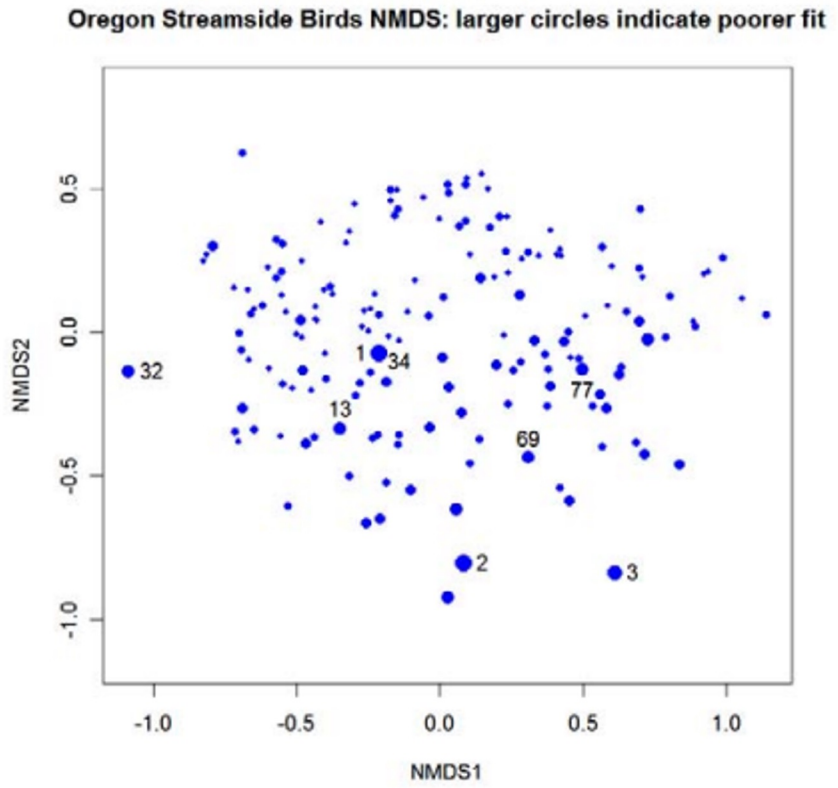
- Should be a staircase



# NMDS

## 3. Influential samples

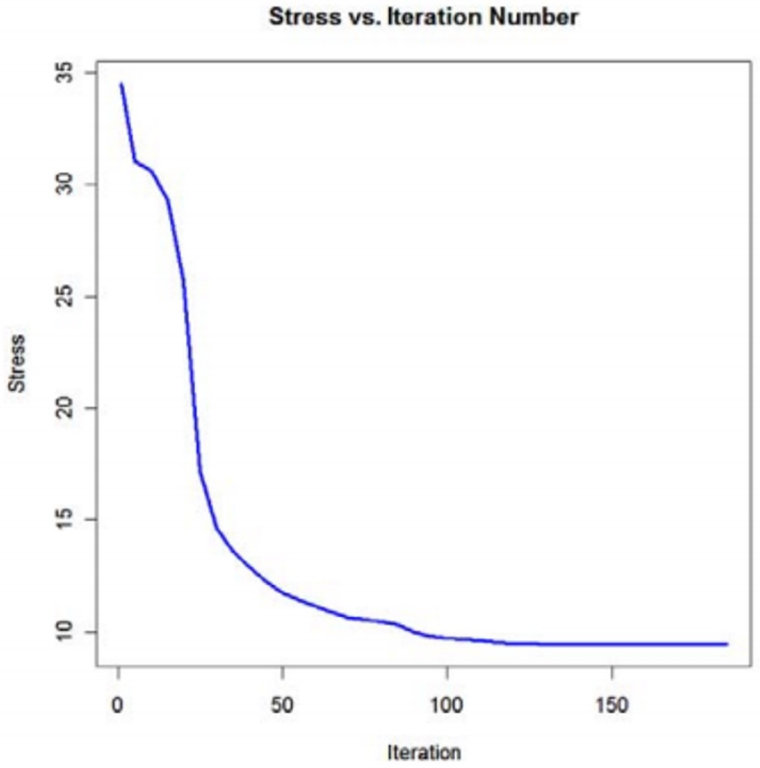
- Final stress is sensitive to high leverage points with particularly poor fit.
- Goodness-of-fit statistic for each point (i.e., its stress) can reveal influential samples.



# NMDS

## 4. Stable Solution

- For a stable solution, the iterative search to minimize stress should decline steeply and then stabilze.
- Unstable solution can be caused by overfitting the data (using a dim too high given N), or where 2 or more local minima are equal competitors for the global minimum.



# NMDS

- **Strengths**
  - No assumptions of linearity or distribution of species or responses
  - Unbiased and faithful to original data structure
  - Most suitable of community data
- **Weaknesses**
  - Not based on an ecological model
  - Sensitive to misspecification
  - Axes do not contain inherent information

## Indirect (Unconstrained) Ordination Techniques

*Comparison of Ordinations*

| Technique | When to use |
|---|---|
| PCA | Not for community data; with short gradients; when *linear model* is appropriate |
| CA | For community data when *unimodal* assumption met well (i.e., long gradients) |
| DCA | As above when arch and compression effect is severe |
| MDS | When Euclidean metric and linear mapping desirable but PCA assumptions not met (e.g., binary data) |
| NMDS | When other models not suitable |