PCoA – Where are the species?

- Because PCoA is based on a distance matrix, the analysis never “sees” any species data.

We can add species by going back and calculating weighted averages.

Species Weighted Averages

From the original data matrix (species abundance): for a given species, average the product of that species abundance and axis scores for all sites. Repeat for each species and axis combination.

Weighted average for *N. texanus* on axis 1 = -0.0969
Calculating Weighted Averages

- Function `wascores(x, y)`
  - `x`: data matrix
  - `y`: sample scores ($points part of PCoA object)

- Returns weighted average scores that can be plotted (example in the PCoA answer posted)

- Note that you could also use `wascores()` to plot environmental variables…anything for which you have data for each sample.

Principal Components Analysis (PCA)

- First and most basic eigenvalue based ordination
- Works with the original dataset, not a distance matrix
- Eigenvalue decomposition of correlation or covariance matrix
  - Relationship among variables and gradients is important
  - Relationship to variables is not lost (no need for wascores)
- Works best with linear relationships among variables, thus, some data are not appropriate (species) and others require transformation
Principal Components Analysis (PCA)

Two dimensional regression - best fit line through points describes the relationship, explains some proportion of variance.

PCA is similar but in multivariate space. Best fit line in multiple dimensions = component. By definition passes through centroid.

Axis 1

Axis 2

Both vectors pass through the centroid of the data and are rotated to minimize residuals (maximize variance explained).

Relative spatial positioning among points unchanged.

Axis 1 longer = more variance explained.

PCA Functions in R

- Three functions in R:
  - `prcomp()` Singular value decomposition of variance/covariance matrix
    - Variance calculated using N-1
    - By default, data are not scaled
  - `princomp()` Eigen analysis of wither the correlation of variance/covariance matrix
    - Variance calculated differently (N)
    - No option for scaling data (need to do this before)
    - Option `cor` specifies if you use correlation or covariance matrix
  - `rda()` Vegan package function that will be used for constrained ordinations
    - Similar to `prcomp()`, species and site scores are rescaled according to Legendre, P. and Legendre, L. (1998) Numerical Ecology. 2nd English ed. Elsevier
Shared Zeroes and Outliers

Shared zero data among species will lead to a false high positive relationship.

A single large outlier can define the linear relationship among two variables.

PCA restrictions

- Relationships among species and gradients are important.
- Thus, variables should be normally distributed and linearly related to gradients.
- Shared zero data is going to be problematic.

Data Assumptions

- Often helpful to look at scatter plots of data first.
- Common transformations – eliminate rare species, log transform, scale data etc.
Principle Components Analysis (PCA)

- **Code**
  - prin_comp <- rda(community, scale=TRUE)

- **Options**
  - Scale = rescale data automatically to have unit variance (~ 1SD per variable)

- **Output**
  - Eigenvalues and eigenvectors
  - Summary of % variance accounted for ("inertia" = % variance)
    - If scale=TRUE then you have 1 unit of inertia per variable
    - If scale=FALSE then variance contributed is not equal across variables
  - Scaling component
  - Species scores
  - Sample scores

---

**Biplot – species and samples in ordination space**

- What to present:
  - Ordination –
  - Sample Scores –
  - Loadings –
  - Percent variance accounted for –

---

**How many axes, which to use?**

- **By definition**
  - # axes = # of variables (species in this case)
  - all axes account for 100% of variance
  - First axis accounts for most variance, decreases from there.

- **How many to use?**
  - Descriptive (2D plot, diminishing returns)
  - A priori % variance accounted for
  - Analytical (non-random axis, Monte Carlo determination)
How many axes to use?

- Function `screeplot()` will plot the variation explained by axis.
- Option `bstick=TRUE` will include the broken stick distribution of variance explained. Broken stick – each axis gets an equal amount of the remaining variation.
- Recommendation is to ignore axes where variance explained is less than broken stick.

**Prin Comp**

![Graph showing variation explained by axes](image)

**Climate change and hockey sticks**

**Global-scale temperature patterns and climate forcing over the past six centuries**

Spatially resolved global reconstructions of annual surface temperature patterns over the past six centuries are based on the multivariate calibration of safely distributed high-resolution proxy climate indicators. Time-dependent correlations of the reconstructions with time-series indices representing changes in greenhouse-gas concentrations, solar irradiance, and volcanic aerosols suggest that each of these factors has contributed to the climate variability of the past 600 years, with greenhouse gases emerging as the dominant forcing during the twentieth century. Northern Hemisphere mean annual temperatures for three of the past eight years are warmer than any other year since (at least) 1600.

Data used

- Multiple tree ring datasets
- Not directly comparable (differences in local climate, species ecology etc.)
- Data were normalized

**Tree Ring Dataset**

![Image of tree ring dataset](image)

- The most variable tree ring datasets were most closely related to early axes (a function of how PCA works).
- “Criticism” – normalization procedure over emphasized the more variable datasets, the analysis was “fixed” to emphasize recent (anthropogenic) change.
- However – underlying pattern among all datasets is similar, changing the normalization procedure does not change the analysis in any meaningful way.
Assignment

- Section 5.3 in the text
- Papers
  - FYI papers
- Example dataset and script

Assignment

- Environmental data describing stream habitat
  - Drainage
  - Variables
    - Cumulative drainage area
    - pH
    - DC – dissolved oxygen
    - Temp
    - TDS – total dissolved solids
    - Depth
    - Flow
    - Substrate – mean on a 1-6 scale where 1=silt and 6=boulder
    - Cover
    - Vegetation
    - Width
    - Canopy cover
    - Turbidity
- Do a PCA using the rda() function
  - How much variance is explained? How do the variables relate to the axes?
  - Plot the ordination with colors for different drainages.
  - What two variables are best linked to axes 1 and 2?
  - How do the drainages differ?