Procrustes Distance

- metaMDS uses a Procrustean analysis instead of stress values to assess differences in configurations for each iteration
  - Procrustes – son of Poseidon, had an iron bed that all visitors had to sleep in. He would stretch or cut limbs so they all fit.
  - Procrustes analysis – how much “cutting” and “stretching” is necessary to get one configuration (ordination) to match another.

Procrustes Statistic

- Any two configurations can be compared and their differences quantified with a Procrustes statistic – sum of squares of the residuals.

- How different are these two configurations?
Stress

- Stress is calculated from the fit (residuals) of the original distance matrix and Euclidian distance in ordination space.
- The Non-metric fit is
  \[ 1 - S^2 \]
  \[ 1 - (\text{nmds}\text{stress}^2) \]
  \[ \text{[1]} \ 0.974788 \]
- The linear fit (R^2) is not the same as stress, but is analogous.
- Using R^2 as a measure of % variance explained is not precise because of some things nmds does to improve fit (noshare option etc.).

Each point here represents a pair of sites.
- Stress is calculated from the residuals.
- Thus, you can figure out how much each site is contributing to stress.
- This is the goodness of fit measure.

\[ \sum \text{goodness}^2 = \text{Stress}^2 \]

NMDS with weighted averages for 9 most abundant species

- Warnings...
  - Zero distance among samples
  - No species scores (ordiplot)
- Despite the lower stress, the K=2 NMDS is better if you were to present a 2D plot. The first two dimensions of K=2 is better than first two of K=4.
Correspondence Analysis (reciprocal averaging)

- Recall some of the properties (potential weaknesses) with PCA:
  - Assumes variables are linearly related with each other and/or gradients
  - Samples are ordinated in variable space
  - Results in “horseshoe effect” where ends of ordination axes are distorted (shared 0 seen as a similarity)
- Correspondence analysis allows for non-linear unimodal relationships
- Both samples and variables handled similarly, axes do not explicitly represent species-space

Table 9.1 from Legendre and Legendre (1998)

<table>
<thead>
<tr>
<th>Method</th>
<th>Distance</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>Euclidean</td>
<td>Quantitative, linear relationships assumed, beware of double-zeros</td>
</tr>
<tr>
<td>PCoA</td>
<td>Any</td>
<td>Quantitative, qualitative or mixed</td>
</tr>
<tr>
<td>NMDS</td>
<td>Any</td>
<td>Quantitative, qualitative or mixed</td>
</tr>
<tr>
<td>CA</td>
<td>$\chi^2$</td>
<td>Non-negative, quantitative or binary</td>
</tr>
<tr>
<td>FA</td>
<td>Euclidean</td>
<td>Quantitative, linear relationships assumed, beware of double-zeros</td>
</tr>
</tbody>
</table>

History of Ordinations in Ecology

Correspondence Analysis

- Based on traditional Chi-Square approach to measure correspondence between rows and columns.

<table>
<thead>
<tr>
<th>Site</th>
<th>sp1</th>
<th>sp2</th>
<th>sp3</th>
<th>sp4</th>
</tr>
</thead>
<tbody>
<tr>
<td>sam1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>sam2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>sam3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>sam4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>sam5</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Row Total</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

- As rows and column deviate (more independent), Chi Sqr values (and inertia) grows.

<table>
<thead>
<tr>
<th>Site</th>
<th>sp1</th>
<th>sp1</th>
<th>sp2</th>
<th>sp2</th>
<th>sp3</th>
<th>sp3</th>
<th>sp4</th>
<th>sp4</th>
<th>sp5</th>
<th>sp5</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>sam1</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12-21^2</td>
</tr>
<tr>
<td>sam2</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>3.462503</td>
</tr>
<tr>
<td>sam3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>27.81347</td>
</tr>
<tr>
<td>sam4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>0.590862</td>
</tr>
<tr>
<td>sam5</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>16.11917</td>
</tr>
</tbody>
</table>

- In this simple dataset, rows and columns are not independent. Thus, the contents of each cell are predictable based on row and column totals and the grand total.

Correspondence Analysis

- As rows and column deviate (more independent), Chi Sqr values and inertia grow.

<table>
<thead>
<tr>
<th>Site</th>
<th>sp1</th>
<th>sp1</th>
<th>sp2</th>
<th>sp2</th>
<th>sp3</th>
<th>sp3</th>
<th>sp4</th>
<th>sp4</th>
<th>sp5</th>
<th>sp5</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>sam1</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12-21^2</td>
</tr>
<tr>
<td>sam2</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>3.462503</td>
</tr>
<tr>
<td>sam3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>27.81347</td>
</tr>
<tr>
<td>sam4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>0.590862</td>
</tr>
<tr>
<td>sam5</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>16.11917</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Site</th>
<th>sp1</th>
<th>sp1</th>
<th>sp2</th>
<th>sp2</th>
<th>sp3</th>
<th>sp3</th>
<th>sp4</th>
<th>sp4</th>
<th>sp5</th>
<th>sp5</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>sam1</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12-21^2</td>
</tr>
<tr>
<td>sam2</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>3.462503</td>
</tr>
<tr>
<td>sam3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>27.81347</td>
</tr>
<tr>
<td>sam4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>0.590862</td>
</tr>
<tr>
<td>sam5</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>16.11917</td>
</tr>
</tbody>
</table>

- In this simple dataset, rows and columns are not independent. Thus, the contents of each cell are predictable based on row and column totals and the grand total.

Inertia = Chi Sqr / Grand Sum

- This matrix describes all the variability in the dataset not explainable by row or column profiles (totals).

<table>
<thead>
<tr>
<th>Site</th>
<th>sp1</th>
<th>sp1</th>
<th>sp2</th>
<th>sp2</th>
<th>sp3</th>
<th>sp3</th>
<th>sp4</th>
<th>sp4</th>
<th>sp5</th>
<th>sp5</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>sam1</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12-21^2</td>
</tr>
<tr>
<td>sam2</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>3.462503</td>
</tr>
<tr>
<td>sam3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>27.81347</td>
</tr>
<tr>
<td>sam4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>0.590862</td>
</tr>
<tr>
<td>sam5</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>16.11917</td>
</tr>
</tbody>
</table>

- Total variance the analysis will attempt to explain.

Correspondence Analysis

- CA does an eigenvalue decomposition to summarize this variability in fewer axes (components).

- Species and sites that contribute most to the inertia have the largest magnitude CA1 scores.

- Scores are centered and scaled to be directly comparable.
Correspondence Analysis

- Unlike NMDS
  - Not iterative, no local minima problem
  - First axis always most informative
  - Number of axes produced is set by the dimensionality of the data (n-1), not a user option
  - Not distance based, data transformations typically more important
  - Row and column sums must be >0. If your dataset includes negatives, think carefully about whether CA is appropriate.
  - No missing data allowed (like PCA)
  - Data expected to be frequency-based (contingency table)
  - Ordinates both samples and variables directly
  - Also the bases of the most popular direct gradient analysis (CCA)

CA Code

- Correspondence Analysis is run as an unconstrained canonical correspondence analysis (CCA)
- CCA function with no environmental matrix
- Code
  - `ca <- cca(community)`
- Options
  - All options for this function apply to CCA
  - Transformations to raw data are often used before
    - Log or eliminating rare species
  - Scale option – scale species data to unit variance
  - Na.action – how to handle missing data
  - Function `downweight(community, fraction=x)`

Correspondence Analysis

- Output
  - Row and column sums, total Chi Square
  - Species and sample scores that can be plotted in the same space. Interpretation is similar to sample scores and species weighted averages in NMDS.
  - # axes = n-1 for whichever dimension of the data matrix is lower (samples or species).
  - Eigenvalues – relative importance of each axis, interpreted as the percentage of total inertia explained.
**CA Example**

- The "arch effect" here is unwanted.
- The ends of the axes are also compressed.
- Detrending (detrended correspondence analysis, DCA) deals with the arch by:
  - 5 segment smoothing of 1st axis. Divide into segments (weights of 1, 2, 3, 2, 1), center each at 0.
  - Rescaling of axis into "standard deviation" units of species turnover.
- Assumptions
  - Same as for CA
  - DCA is not really an analysis. It is a post hoc modification of a CA.
  - "vague bag of tricks"

**DCA**

- DCA (function `decorana`) – detrending and rescaling of CA
- Only first 4 axes are adjusted, the rest are discarded
- Units on axes are in SD of species turnover, beta diversity in samples measured in length of 1st axis
- Code
  ```R
  decorana(community)
  ```
- Options
  - Downweight rare species (uses downweight function)
  - Number of rescaling iterations
  - Number of detrending segments to use
  - Whether or not to detrend at all (avoid detrending = regular CA)
DCA Output
- For first 4 axes only:
  - Eigenvalues
  - Variable scores
  - Site scores
  - “Decorana Values” – values estimated before detrending…interpretation unclear

DCA Example
- Arch effect removed
- First axis a good representation of the original gradient
- Species evenly distributed along first axis
- First axis length = 6, indicates complete turnover in variables (species)
- Second axis length ~ 2

CA and DCA
- First axis is usually fine (not very different from CA axis 1), second is often problematic.
- Problems – adjustments made during detrending are arbitrary.
Assignment

- Reading – section 5.4

Assignment

- Use the Galapagos Island microsatellite dataset in the ade4 package.
  
  ```r
  library(ade4)
  data(ggtortoises)
  ```

- Data:
  
  ```r
  ggtortoises$tab
  ```

- Subspecies:
  
  ```r
  ggtortoises$pop$subspe
  ```

- Eliminate alleles with less than 2 occurrences

- Do a correspondence analysis
  
  - Which markers are most closely associated with the first two axes?
  
  - Which populations are most distinct (describe how populations are related based on this dataset)?

- Plot populations with a few of the most important markers.

---

### Allel frequency for 10 msat loci

Load the vegan library last – package ade4 has a cca function that works differently