

## Correspondence Analysis

### Kinds of data.

- counts or types or taxa in assemblages (the sort of counts that it makes sense to compute percentages from).
- assemblages from stratified deposits
- spatially scatter quadrats
- both
- presence/absence or occurrence data

### Typical applications

- frequency seriation (examples to follow)
- occurrence seriation (e.g. medieval deeds)
- ordination
  - seeing similarity among observations in a low-dimensional space
  - seeing similarity among variables in low-dimensional space
- seriation ... plus: "time is not the only dimension"

## Correspondence Analysis

1. Eigenanalysis of a funky covariance matrix (a form of PCA).
  - Inertia (vs. variance).
  - Chi-squared distances (vs. Euclidean distances).
2. "Reciprocal averaging" or indirect gradient analysis
  - A connection to MCD's and seriation
  - Approximation to ML estimation under the Gaussian response model.
  - The arch effect (vs. PCA horseshoe).
3. Dretrended CA: a problem, not a solution.
4. Examples.
  - Woodland assemblages from Georgia\*
  - San Marcos Pueblo

\*thanks to Karen Smith.

## CA as a form of PCA.

### 1. Transform the counts.

$$X = \begin{matrix} & 5 & 10 & 20 \\ 5 & 10 & 40 & 5 \\ 40 & 5 & 0 & \\ 50 & 1 & 0 & \end{matrix}$$

Some data – 3 types and 4 assemblages

$$x_{++} = 186$$

Sum of the counts

$$P = X / x_{++}$$

Counts / sum of the counts

$$P = \begin{matrix} 0.0268817 & 0.0537634 & 0.1075269 & \\ 0.0537634 & 0.2150538 & 0.0268817 & \\ 0.2150538 & 0.0268817 & 0 & \\ 0.2688172 & 0.0053763 & 0 & \end{matrix}$$

### 1. Transform the counts (continued).

$$[p_{+j}] = 0.5645161 \ 0.3010753 \ 0.1344086 \quad \text{The column masses} \sim \text{marginals}$$

$$[p_{i+}] = \begin{matrix} 0.188172 \\ 0.2956989 \\ 0.2419355 \\ 0.2741935 \end{matrix} \quad \text{The row masses}$$

$$P = \begin{matrix} 0.0268817 & 0.0537634 & 0.1075269 & \\ 0.0537634 & 0.2150538 & 0.0268817 & \\ 0.2150538 & 0.0268817 & 0 & \\ 0.2688172 & 0.0053763 & 0 & \end{matrix}$$

$$Q = q_{ij} = \left[ \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \right] \quad \text{Look familiar??}$$

### 1. Transform the counts (continued).

$$Q = q_{ij} = \left[ \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \right]$$

$$[p_{+j}] = 0.5645161 \ 0.3010753 \ 0.1344086$$

$$P = p_{ij} = \begin{matrix} 0.0268817 & 0.0537634 & 0.1075269 & \\ 0.0537634 & 0.2150538 & 0.0268817 & \\ 0.2150538 & 0.0268817 & 0 & \\ 0.2688172 & 0.0053763 & 0 & \end{matrix} \quad [p_{i+}] = \begin{matrix} 0.188172 \\ 0.2956989 \\ 0.2419355 \\ 0.2741935 \end{matrix}$$

$$Q = \begin{matrix} -0.243445 & -0.012144 & 0.517089 & \\ -0.276976 & 0.422375 & -0.06452 & \\ 0.2123518 & -0.170288 & -0.180328 & \\ 0.2898373 & -0.268608 & -0.191974 & \end{matrix}$$

### 1. Transform the counts (continued).

$$q_{ij} = \left[ \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \right] \quad \text{CA transformation}$$

$$\chi_{ij} = \frac{(O_{ij} - E_{ij})}{\sqrt{E_{ij}}} = \sqrt{x_{ij}} \left[ \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \right] \quad \text{Chi-squared table entries. Residuals from expected under hypothesis of independence: "chi-squared residuals"}$$

Still looks familiar...??

1. Transform the counts (continued).

$$q_{ij} = \left[ \frac{p_{ij} - p_{i+} p_{+j}}{\sqrt{p_{i+} p_{+j}}} \right]$$

CA transformation

$$\chi_{ij} = \frac{(O_{ij} - E_{ij})}{\sqrt{E_{ij}}} = \sqrt{x_{++}} \left[ \frac{p_{ij} - p_{i+} p_{+j}}{\sqrt{p_{i+} p_{+j}}} \right]$$

Chi-squared table entries. Residuals from expected under null hypothesis of no-association.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Z-scores!

2. Measuring the amount of variation in the transformed counts.

$$Q = q_{ij} = \left[ \frac{p_{ij} - p_{i+} p_{+j}}{\sqrt{p_{i+} p_{+j}}} \right]$$

How much variation is there in the (transformed) data?

$$I = \sum_{i=1}^r \sum_{j=1}^c q_{ij}^2 = 0.8856929$$

Total variation -- "Inertia"

$$\chi_{ij} = \frac{(O_{ij} - E_{ij})}{\sqrt{E_{ij}}} = \sqrt{x_{++}} \left[ \frac{p_{ij} - p_{i+} p_{+j}}{\sqrt{p_{i+} p_{+j}}} \right]$$

How much variation is there in the chi-square residuals?

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \chi_{ij}^2$$

Total variation -- "Total Chi-Square"

$$\chi^2 = x_{++} \sum_{i=1}^r \sum_{j=1}^c q_{ij}^2$$

$$= (186 \times 0.8856929) = 164.73887$$

Total Chi-Square vs. Inertia

2. Measuring the amount of variation (continued).

$$Q = q_{ij} = \left[ \frac{p_{ij} - p_{i+} p_{+j}}{\sqrt{p_{i+} p_{+j}}} \right]$$

$$Q = \begin{bmatrix} -0.243445 & -0.012144 & 0.517089 \\ -0.276976 & 0.422375 & -0.06452 \\ 0.2123518 & -0.170288 & -0.180328 \\ 0.2898373 & -0.268608 & -0.191974 \end{bmatrix}$$

$$I_i = \sum_{j=1}^c q_{ij}^2 \quad \text{Inertia of the } i\text{'th row.}$$

$$I_j = \sum_{i=1}^r q_{ij}^2 \quad \text{Inertia of the } j\text{'th column}$$

3. Get the Sum of Squares and Cross Products Matrix.

$$Q = q_{ij} = \left[ \frac{p_{ij} - p_{i+} p_{+j}}{\sqrt{p_{i+} p_{+j}}} \right]$$

$$Q = \begin{bmatrix} -0.243445 & -0.012144 & 0.517089 \\ -0.276976 & 0.422375 & -0.06452 \\ 0.2123518 & -0.170288 & -0.180328 \\ 0.2898373 & -0.268608 & -0.191974 \end{bmatrix}$$

transpose

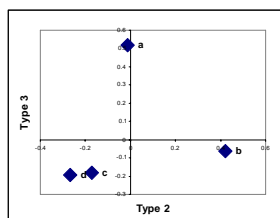
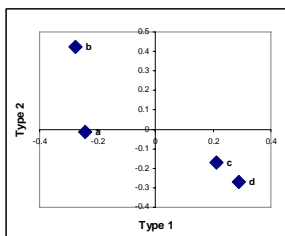
$$Q' = \begin{bmatrix} -0.243445 & -0.276976 & 0.2123518 & 0.2898373 \\ -0.012144 & 0.422375 & -0.170288 & -0.268608 \\ 0.517089 & -0.06452 & -0.180328 & -0.191974 \end{bmatrix}$$

$$C = Q'Q = \begin{bmatrix} 0.2650804 & -0.228045 & -0.201946 \\ -0.228045 & 0.2796964 & 0.0487422 \\ -0.201946 & 0.0487422 & 0.3409161 \end{bmatrix}$$

-A SSCP matrix.  
- "Variances" on the diagonal.  
- Trace(C) = ??  
- "Covariances" elsewhere.

3. Get the SSCP Matrix (continued).

$$C = Q'Q = \begin{bmatrix} 0.2650804 & -0.228045 & -0.201946 \\ -0.228045 & 0.2796964 & 0.0487422 \\ -0.201946 & 0.0487422 & 0.3409161 \end{bmatrix}$$



3. Compute the eigenvalues and eigenvectors of Q.

-Find the line through the points that maximizes the variance or inertia along the line of the perpendicular projections from each row point to the line. (a.k.a. Dimension 1, Axis 1.)  
-Find a second line, orthogonal to the first, that maximizes inertia (Dimension 2).

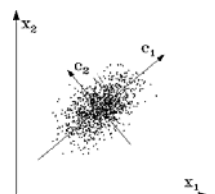
The eigenvalues: one for each axis.

$$\lambda = \begin{matrix} 1 & 0.6200765 \\ 2 & 0.2656164 \\ 3 & -6.96E-18 \end{matrix}$$

-Why only 2 that are non-zero?  
-What is their sum equal to?

The eigenvectors: one vector for each axis.

$$U = \begin{matrix} t1 & u1 & u2 & u3 \text{ (junk)} \\ t2 & -0.649239 & 0.1182083 & 0.7513429 \\ t3 & 0.5151065 & -0.658476 & 0.5487033 \\ t3 & 0.5596022 & 0.7432609 & 0.3666178 \end{matrix}$$



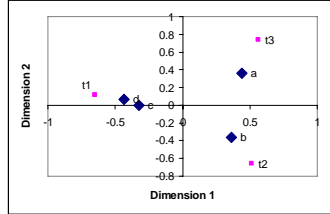
Coefficients defining the first line (Dimension 1) on the basis of the row points (the rows of the q matrix).

4. Use the eigenvectors of C to get the scores of the row points on each dimension.

$QU = S$

$$\begin{matrix} A \\ B \\ C \\ D \end{matrix} \begin{pmatrix} t1 & t2 & t3 \\ -0.243445 & -0.012144 & 0.517089 \\ -0.276976 & 0.422375 & -0.06452 \\ 0.2123518 & -0.170288 & -0.180328 \\ 0.2898373 & -0.268608 & -0.191974 \end{pmatrix} \times \begin{matrix} t1 \\ t2 \\ t3 \end{matrix} \begin{pmatrix} u1 & u2 & u3 \text{ (junk)} \\ -0.649239 & 0.1182083 & 0.7513429 \\ 0.5151065 & -0.658476 & 0.5487033 \\ 0.5596022 & 0.7432609 & 0.3666178 \end{pmatrix} =$$

$$\begin{matrix} A \\ B \\ C \\ D \end{matrix} \begin{pmatrix} \text{Dim 1} & \text{Dim 2} \\ 0.4411626 & 0.3635513 & 5.836E-17 \\ 0.3612862 & -0.35882 & 3.216E-17 \\ -0.326495 & 0.0032014 & -4.7E-18 \\ -0.433964 & 0.0684464 & 2.126E-17 \end{pmatrix}$$



Check the sum of squares!

5. Rescale the eigenvectors – so they AND the scores are cooler.

$$[p_{.j}] = \begin{matrix} t1 & t2 & t3 \\ 0.5645161 & 0.3010753 & 0.1344086 \end{matrix} \quad \text{Start with the column masses}$$

$$\frac{1}{\sqrt{p_{.j}}} = \begin{matrix} t1 & t2 & t3 \\ 1.3309503 & 1.8224787 & 2.7276363 \end{matrix} \quad \text{Get the reciprocals of the square roots.}$$

$$D \frac{1}{\sqrt{p_{.j}}} U = V$$

$$\begin{pmatrix} 1.3309503 & 0 & 0 \\ 0 & 1.8224787 & 0 \\ 0 & 0 & 2.7276363 \end{pmatrix} \times \begin{matrix} t1 \\ t2 \\ t3 \end{matrix} \begin{pmatrix} u1 & u2 & u3 \text{ (junk)} \\ -0.649239 & 0.1182083 & 0.7513429 \\ 0.5151065 & -0.658476 & 0.5487033 \\ 0.5596022 & 0.7432609 & 0.3666178 \end{pmatrix} =$$

$$\begin{matrix} v1 & v2 & v3 \text{ (junk)} \\ 1.3309503 & 0 & 0 \\ 0 & 1.8224787 & 0 \\ 0 & 0 & 2.7276363 \end{matrix}$$

$$\begin{matrix} t1 \\ t2 \\ t3 \end{matrix} \begin{pmatrix} -0.864104 & 0.1573294 & 1 \\ 0.9387707 & -1.200058 & 1 \\ 1.5263913 & 2.0273455 & 1 \end{pmatrix} \quad \text{Check the results! What has happened?}$$

5. Rescale the eigenvectors – so they AND the scores are cooler (continued).

$$D \frac{1}{p_{.j}} P V = F$$

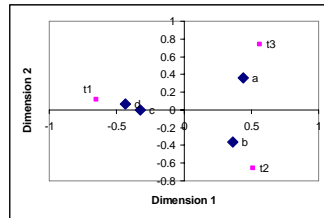
Multiply the row profiles (relative frequencies) by the rescaled coefficients to get the scores.

$$\begin{matrix} A \\ B \\ C \\ D \end{matrix} \begin{pmatrix} t1 & t2 & t3 \\ 0.1428571 & 0.2857143 & 0.5714286 \\ 0.1818182 & 0.7272727 & 0.0909091 \\ 0.8888889 & 0.1111111 & 0 \\ 0.9803922 & 0.0196078 & 0 \end{pmatrix} \times \begin{matrix} v1 & v2 & v3 \text{ (junk)} \\ -0.864104 & 0.1573294 & 1 \\ 0.9387707 & -1.200058 & 1 \\ 1.5263913 & 2.0273455 & 1 \end{pmatrix} =$$

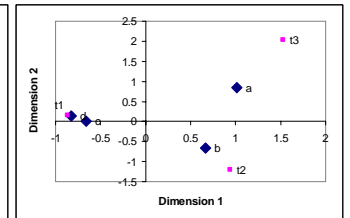
$$\begin{matrix} A \\ B \\ C \\ D \end{matrix} \begin{pmatrix} f1 & f2 & f3 \\ 1.0170003 & 0.8380852 & 1 \\ 0.6643953 & -0.65986 & 1 \\ -0.663785 & 0.0065086 & 1 \\ -0.828754 & 0.130714 & 1 \end{pmatrix}$$

6. Why is rescaling cool?

- Rare types (cols) end up with more weight (stretching the space in their direction).
- Observations (rows) scores are weighted averages of the type coefficients for types that occur in them, where the type relative frequencies are the weights. The eponymous ‘‘correspondence’’.



Before rescaling



After rescaling

6. Why is rescaling cool (continued)?

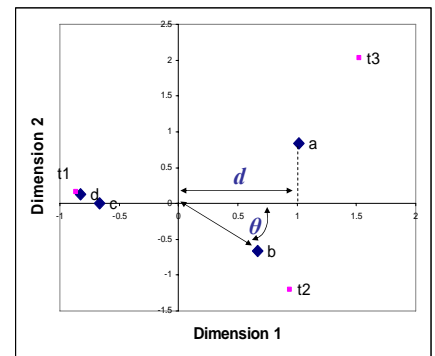
3. Euclidean distances on the CA axes among observations equal chi-square distances among the observations.

$$d_{\chi^2,1,2} = \sqrt{\sum_{j=1}^c \frac{\left( \frac{p_{1j}}{p_{1+}} - \frac{p_{2j}}{p_{2+}} \right)^2}{\frac{p_{+j}}{p_{++}}}}$$

Squared differences among proportions of each type at the two sites.

Column marginals – (weighted) mean proportion of each type

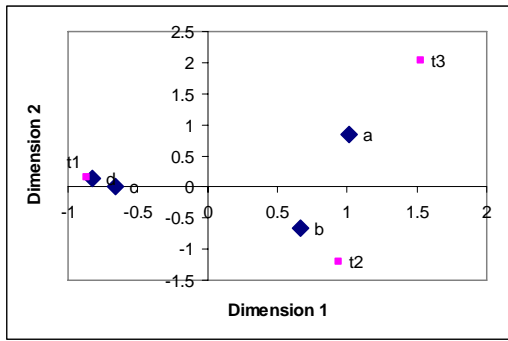
7. CA diagnostics.



$d^2 =$  partial contribution to inertia of dimension 1 from row a – large values mean the row is important in determining the orientation of the dimension.

$COS(\theta)^2 =$  correlation between row b and dimension 1 – how well the dimension 1 score capture row b’s true position.

7. CA Biplot interpretation.



How should we interpret the positions of the points in the joint space:  
 -distances among rows?  
 -relationship between rows and cols?  
 -distances among cols?

8. Other forms of scaling.

2. Switch rows and columns in the analysis as above:
  - column scores now become weighted averages of the row scores.
  - distances among columns are chi-squared distances
  - row positions are
3. Combination of 1 and 2\*:
  - keep the row scores from 1.
  - keep the columns scores from 2.
  - distances between rows *AND* columns are chi-squared distances.
  - relationships among them are only interpreted in terms of directions.

\*Default for most software, including SAS. But 1 usually makes more sense in archaeology and ecology – why?

CA as Reciprocal Averaging

1. Direct Gradient Analysis

The Gaussian response model:

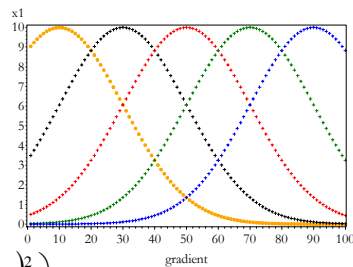
-An environmental gradient (e.g. elevation, rainfall)

-abundance or probability of each taxon determined by position on the gradient

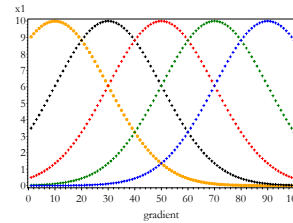
-taxon abundance follows a Gaussian response to the gradient:

$$x_{ij} = \exp \left( \ln(a_j) - .5 \frac{(g_i - u_j)^2}{t_j^2} \right)$$

$x_{ij}$  = the count for the j'th taxon at the i'th sample location – or *prob(1)*.  
 $a_j$  = the maximum abundance of the j'th taxon.  
 $u_j$  = the optimum gradient value for the j'th taxon – the location of its mode.  
 $t_j^2$  = the tolerance (variance) of the j'th taxon.  
 $g_i$  = the location of the sample along the gradient.

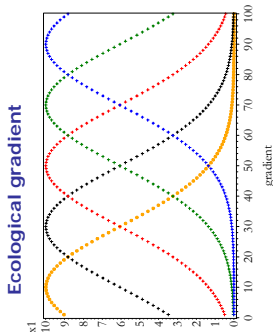


1. Direct Gradient Analysis (continued).

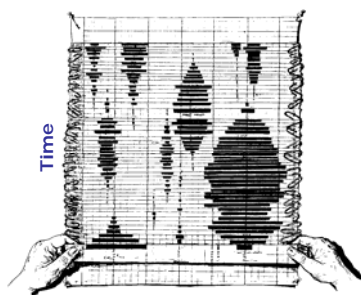


Look familiar?

1. Direct Gradient Analysis (continued).



taxa



types

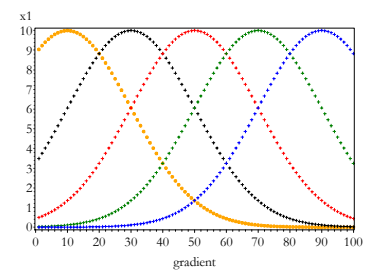
1. Direct Gradient Analysis (continued).

$$\hat{u}_j = \frac{\sum_{i=1}^n x_{ij} g_i}{\sum_{i=1}^n x_{ij}}$$

*Naïve estimate of the optimum species location on the gradient*

$$\hat{g}_i = \frac{\sum_{j=1}^m x_{ij} u_j}{\sum_{j=1}^m x_{ij}}$$

*Naïve estimate of the sample location on the gradient*



$$\hat{g}_i = \frac{\sum_{j=1}^m x_{ij} u_j}{\sum_{i=1}^n x_{ij}}$$

Naïve estimate of the sample location on the gradient

Look familiar?

## 2. Reciprocal Averaging (after M.O. Hill 1973).

1. Start with random site scores ( $x_i$ ).
2. Compute new taxon scores ( $u_i$ ) as weighted averages of site scores.
3. Compute new taxon scores as weighted averages of site scores ( $x_i$ ).
4. For the first axis, go to Step 5. For second and higher axes, make the site scores ( $x_i$ ) uncorrelated with the previous axes by the orthogonalization procedure described below.
5. Keep going until convergence.
6. Standardize the site scores ( $x_i$ ). See below for the standardization procedure.

$$\hat{u}_i = \frac{\sum_{j=1}^n x_{ij} g_j}{\sum_{i=1}^n x_{ij}}$$

Taxon scores

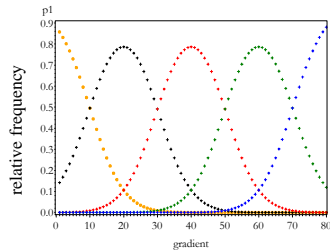
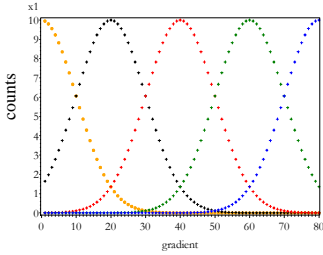
$$\hat{g}_i = \frac{\sum_{j=1}^m x_{ij} u_j}{\sum_{i=1}^n x_{ij}}$$

Site scores

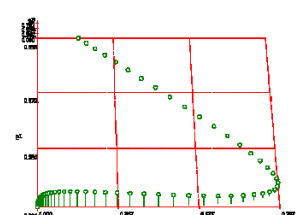
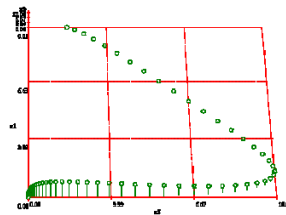
4.1 Orthogonalization: regress site scores for the  $i$ 'th axis on site score for the previous axis. Compute residuals. Use these above.

5.1 Standardization: Compute mean and standard deviation the the site scores. *Note that the standard deviation equals the eigenvalue!!!* Convert the site scores to z-scores.

## 3. The Arch



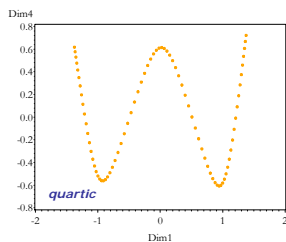
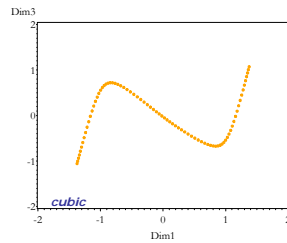
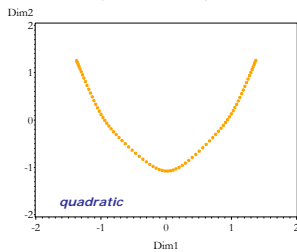
## The Arch (continued)



Taxon space: counts

Taxon space: relative frequencies

## The Arch (continued): CA .



-etc. etc. etc...

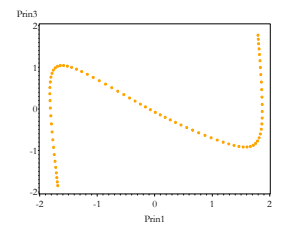
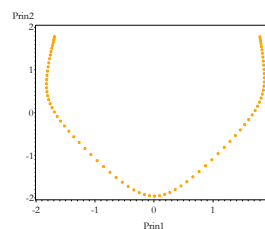
-occurs with abundance and occurrence data

-with real data, random variation swamps higher-order polynomial relationships (>2), rendering them invisible.

-amount of arch is (other things being equal) a function of the amount of taxonomic turnover along the gradient.

-CA Dimension 1 recovers the correct gradient order!!

## The Arch (continued): PCA makes it a horseshoe.



-etc. etc. etc...

-occurs with abundance and occurrence data

-with real data, random variation swamps higher-order polynomial relationships (>2), rendering them invisible.

-amount of arch is (other things being equal) a function of the amount of taxonomic turnover along the gradient.

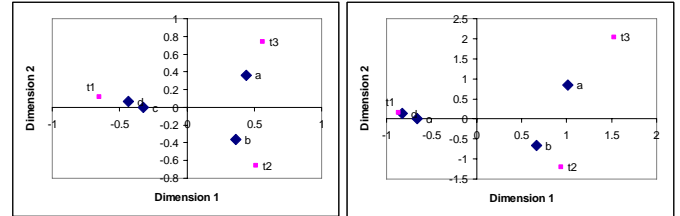
-PCA Dimension 1 does NOT recovers the correct gradient order!!

**Why is CA better than PCA with Gaussian responses?**

**Arch vs. Horseshoe?**

**6. Why is rescaling cool?**

1. Rare types (cols) end up with more weight (stretching the space in their direction) . *Under Gaussian response, rare types occur at the ends of gradients.*
2. Observations (rows) scores are weighted averages of the type coefficients for types that occur in them, where the type relative frequencies are the weights.



**Before rescaling**

**After rescaling**

**6. Why is rescaling cool (continued)?**

3. Euclidean distances on the CA axes among observations equal chi-square distances among the observations.

$$d_{\chi^2, 1, 2} = \sqrt{\sum_{j=1}^c \frac{\left( \frac{p_{1j} - p_{2j}}{p_{1+} - p_{2+}} \right)^2}{\frac{p_{+j}}{p_{++}}}}$$

*Squared differences among proportions of each type at the two sites.*

*Column marginals – (weighted) mean proportion of each type.*

*Rare types get greater weight.*

**Multiple Gradients**

So far: a single gradient with Gaussian responses.

What happens if

-there are 2 (or more) underlying gradients that determine taxonomic responses simultaneously?

-AND the two gradients are to some extent independent.

Archaeological examples:

- time and social status (CST)
- time and activity variation

CA should recover BOTH gradients.

- if the first gradient is short, CA dimension 2 will capture the second variant.
- if there is an arch, the second gradient will often emerge on dimension 3, although it may be partially visible on dimension 1 as well.

**CA: the Big Question:**

**When will estimates from CA approximate what we would get from M-L estimation of site locations, along one or more gradients, based on known taxon parameters from the Gaussian response model?**

Four conditions\*:

- equal or uniformly distributed tolerances along each gradient
- equal or uniformly distributed maxima along each gradient
- equally spaced or uniformly distributed optima along each gradient
- equally spaced or uniformly distributed sample points along each gradient.

If these conditions hold, CA provides an *approximate* solution for the Gaussian models.

\*ter Braak, Cajo. J.F. 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. *Biometrics* 41:859-873

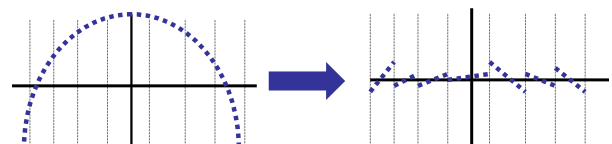
**4. Dretrended CA: solution or new problem?**

**CA problems?**

- The arch
- compression of scores along axis ends.

**DCA solution\*?**

- Chop 2nd axis up into N sub-segments (default = 26). Then rescale points in each segment to have common mean.
- Rescale 1<sup>st</sup> axis scores so within-sample variance is constant across axis 1, Assumes "species packing model" fits the data.



Hill, M.O. & Gauch, H.G. 1980. Detrended Correspondence analysis, an improved ordination technique. *Vegetatio*, 42, 47-58

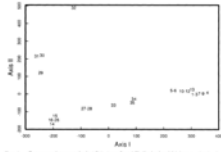


Fig. 1. Correspondence analysis (CA) Axis I and II displaying 25 tubes on the basis of 846 continuity presence-absence data.

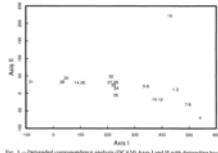


Fig. 3. Detrended correspondence analysis (DCA) Axis I and II with detrending based on the division of axes into 25 segments.

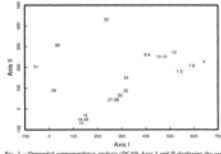


Fig. 2. Detrended correspondence analysis (DCA) Axis I and II displaying the same data as in Fig. 1. No detrending was employed in this analysis, however, the axes were rescaled.

Effect of rescaling only.

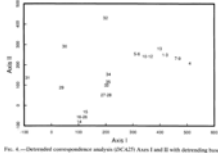


Fig. 4. Detrended correspondence analysis (DCA) Axis I and II with detrending based on the division of axes into 24 segments.

Detrending with 24 vs. 25 segments.

Jackson, DA, and KM Somers. 1991. Putting things in order: the ups and downs of detrended correspondence analysis. *American Naturalist* 137:704-712..