Multivariate Analysis in Ecology

I: Unconstrained Ordination

Jari Oksanen

Oulu

January 2016

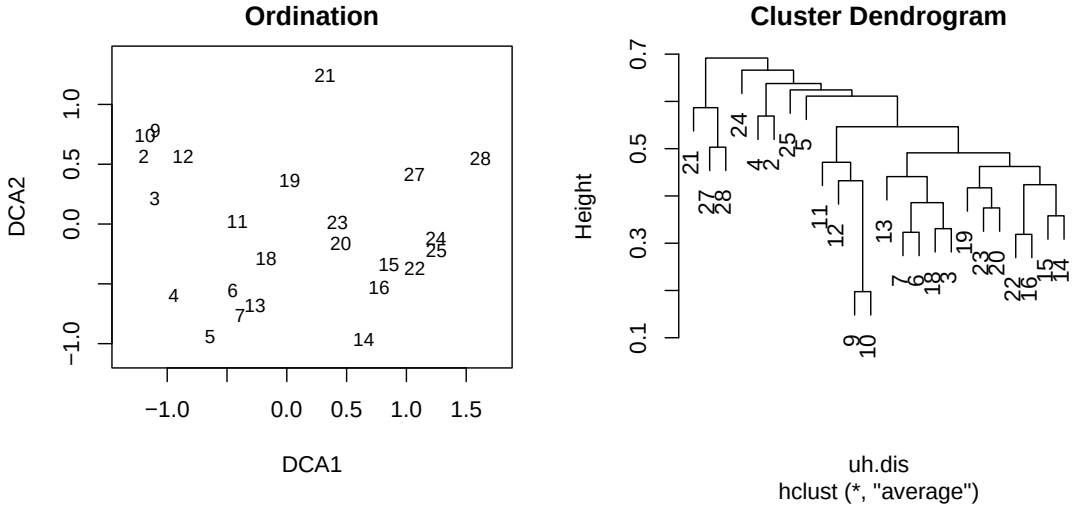
Multivariate Analysis and Ordination

- Basic ordination methods to simplify multivariate data into low dimensional graphics
- Analysis of multivariate dependence and hypotheses
- Analyses can be performed in **R** statistical software using **vegan** package and allies
- Course homepage `http://cc.oulu.fi/~jarioksa/opetus/metodi/`
- **Vegan** homepage `https://github.com/vegandevs/vegan/`

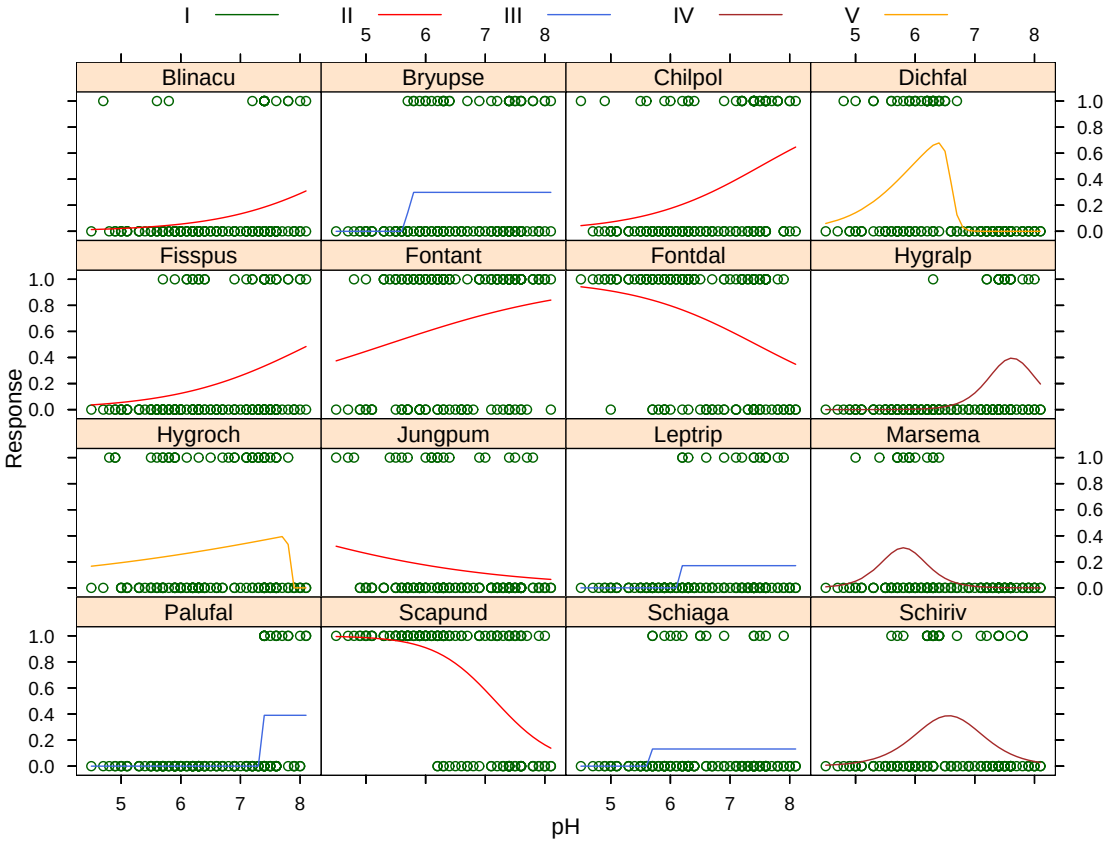
Outline

Why Ordination?

- **Nobody** should want to make an ordination, but they are desperate with multivariate data
- Map multidimensional table into low-dimensional display



Two Ways of Analysing Data



Gradient Analysis

- Gradient Analysis developed in 1950s in USA, with R. H. Whittaker as the main founding father
- Only two or three environmental variables, or *Gradients* needed to explain complicated community patterns
- Against classification: Species responses smooth along gradients
- Against organism analogies: Species responses individualistic
- The basis of modern theory and praxis: Ordination and Gradient modelling of communities

The Gradient Model

R. H. Whittaker (1956) Vegetation of The Great Smoky Mountains. *Ecological Monographs* **26,** 1–80.

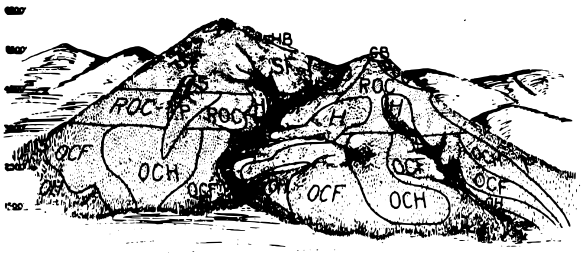
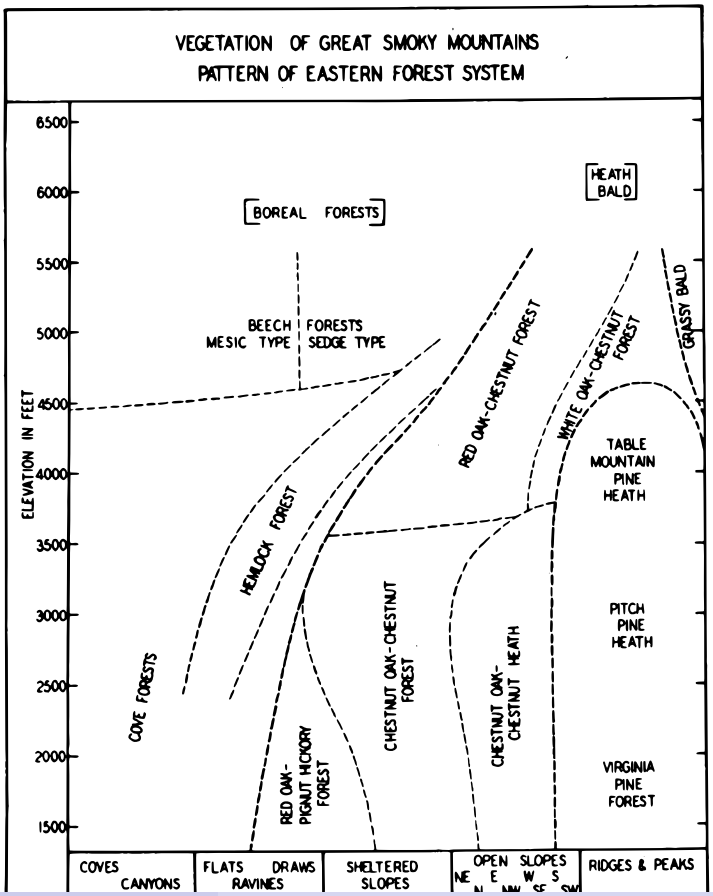


FIG. 21. Topographic disposition of vegetation types. View of idealized mountain and valley, looking east, with 6500-ft peak bearing subalpine forest on left, lower 5500-ft peak covered up to summit bald with deciduous forest on right. Vegetation types:

BG—Beech Gap
CF—Cove Forest
F—Fraser Fir Forest
GB—Grassy Bald
H—Hemlock Forest
HB—Heath Bald
OCF—Chestnut Oak-Chestnut Forest
OCH—Chestnut Oak-Chestnut Heath
OH—Oak-Hickory Forest
P—Pine Forest and Pine Heath
ROC—Red Oak-Chestnut Forest
S—Spruce Forest
SF—Spruce-Fir Forest
WOC—White Oak-Chestnut Forest

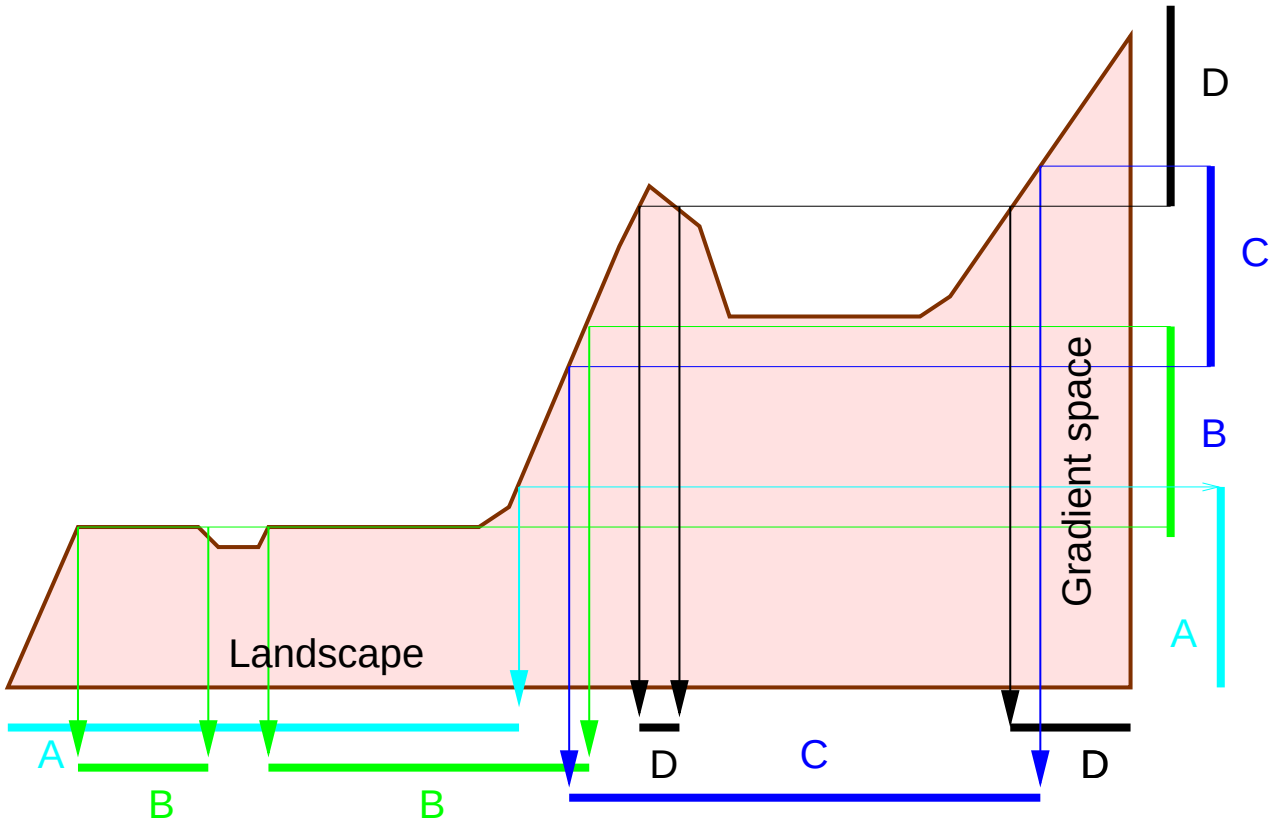


Types of Gradients

1. **Direct gradients:** Influence organims but are not consumed.
   - Correspond to conditions.
2. **Resource gradients:** Consumed
   - Correspond to resources.

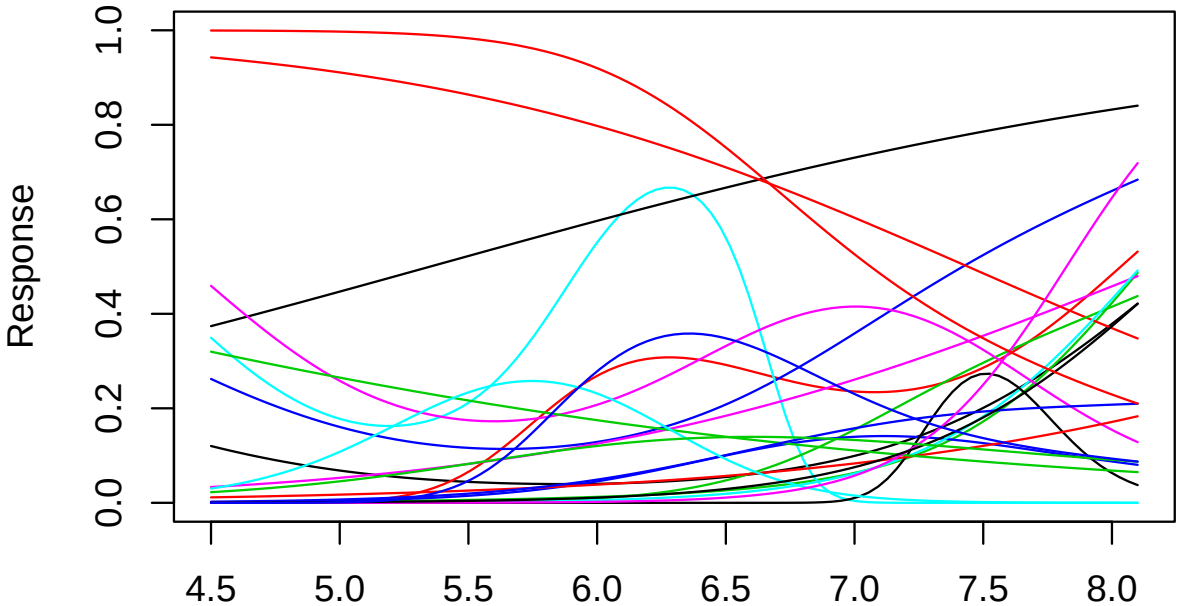
- **Complex gradients.** Covarying direct and/or resource gradients: Impossible to separate effects of single gradients.
  - Most observed gradients.

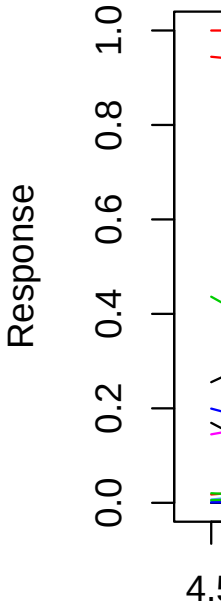
Landscapes and Gradients



Species responses

- Species have non-linear responses along gradients.
- Often assumed to be Gaussian...

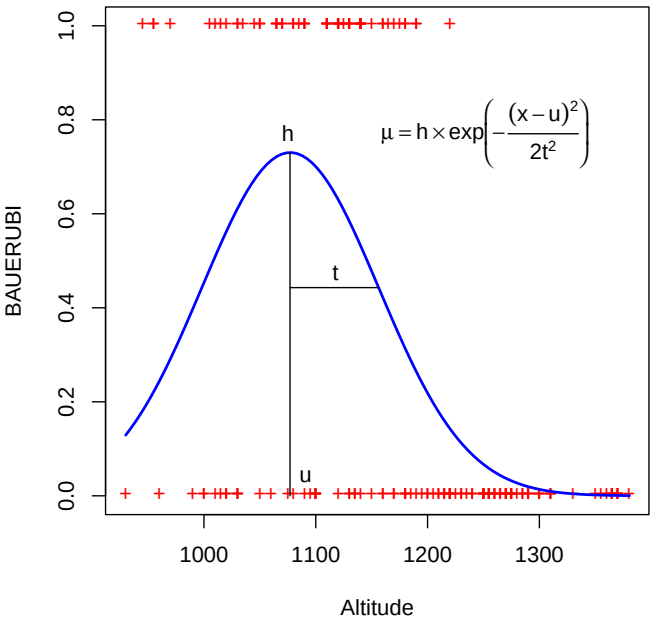




Gaussian Response Function

Gaussian Response Function has three interpretable parameters that define the expected response μ along the gradient x

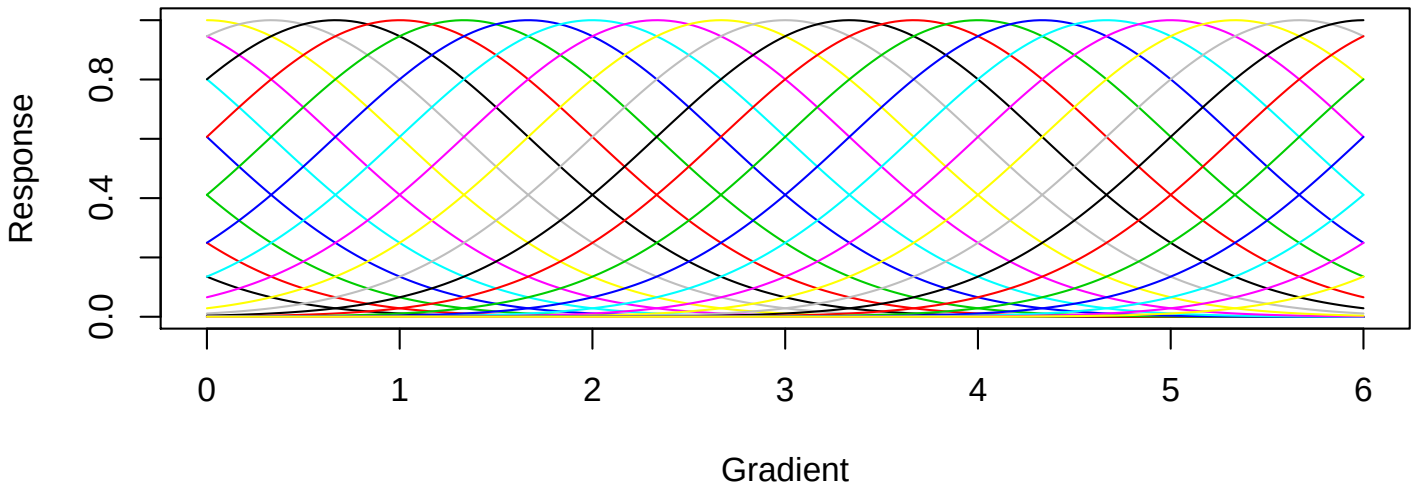
- Location of the optimum u on the gradient x
- Width of the response t in the units of gradient x
- Height of the response h in the units of response height μ



Dream of species packing

Species have Gaussian responses and divide the gradient optimally:

- Equal heights h.
- Equal widths t.
- Evenly distributed optima u.



Evidence for Gaussian Responses

- Whittaker reported a large number of different response types
- Only a small proportion were symmetric, bell shaped responses
- Still became the standard of our times
- Comparison of ordination methods based on simulation, and many of those use Gaussian responses
- We need to use simulation because then we know the truth that should be found

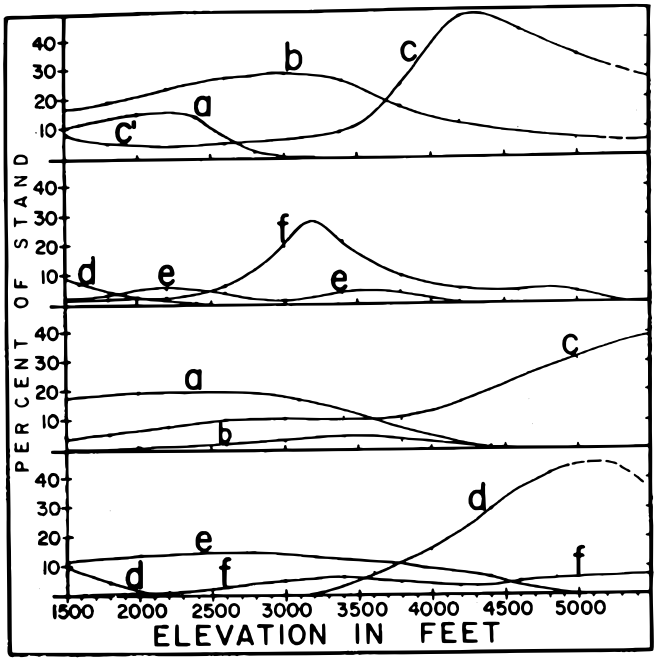


FIG. 9. Elevation transects in submesic and subxeric sites, smoothed curves for tree species. Above—submesic sites: a, *Cornus florida*; b, *Acer rubrum*; c and c', *Quercus borealis* and var. *maxima*; d, *Carya tomentosa*; e, *Carya glabra*; f, *Hamamelis virginiana*. Below—subxeric sites: a, *Quercus prinus*; b, *Sassafras albidum*; c, *Castanea dentata*; d, *Quercus alba*; e, *Oxydendrum arboreum*; f, *Robinia pseudoacacia*.

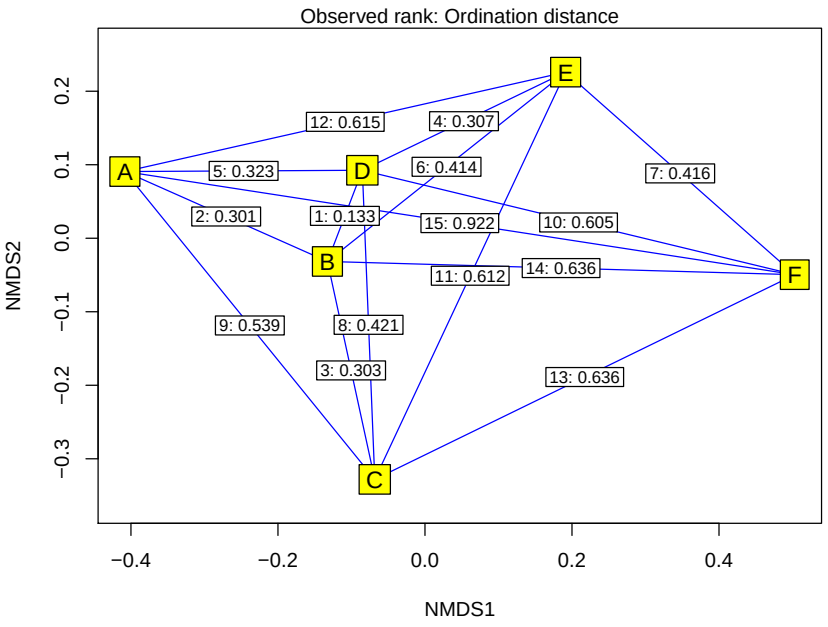
Ordination

- Ordination maps multivariate data onto low dimensional displays: "Most data sets have 2.5 dimensions"
- Gradients define vegetation: ordination tries to find the underlying gradients
- Basic ordination uses only community composition: *Indirect Gradient Analysis*
- Constrained ordination studies only the variation that can be explained by the available environmental variables: Often called *Direct Gradient Analysis*
- Distinct flavours of tools:
  - Nonmetric MDS the most robust method
  - PCA duly despised
  - Flavours of Correspondence Analysis popular
  - Canonical method: Constrained Correspondence Analysis

Nonmetric Multidimensional Scaling

- Rank-order relation with (1) community dissimilarities and (2) ordination distances: No specified form of regression, but the the best shape is found from the data.
- Non-linear regression can cope with non-linear species responses of various shapes: Not dependent on Gaussian model.
- Iterative solution: No guarantee of convergence.
- Must be solved separately for each number of dimensions: A lower dimensional solutions is not a subset of a higher, but each case is solved individually.
- A test winner, and a natural choice. . .

From Ranks of Dissimilarities to Ordination Distances

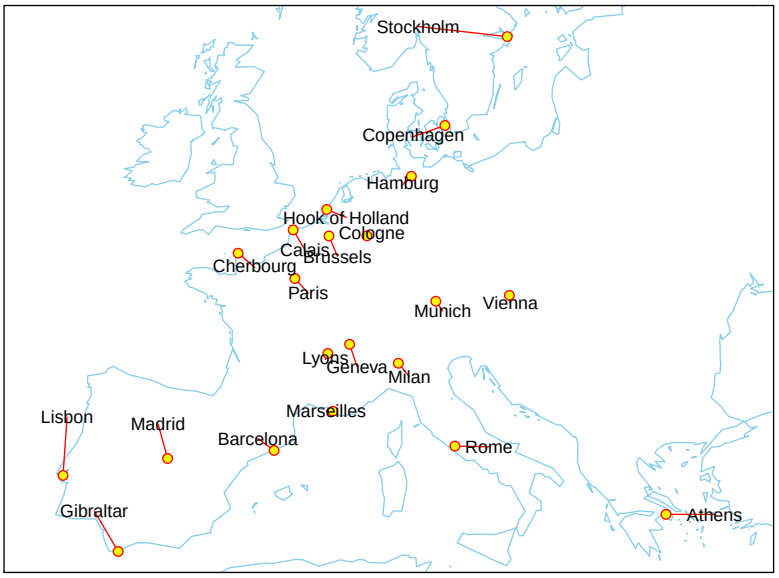



```
Observed dissimilarities:
      A     B     C     D     E
B 0.467
C 0.636 0.511
D 0.524 0.356 0.634
E 0.843 0.600 0.753 0.513
F 0.922 0.893 0.852 0.667 0.606

Ranks of observed dissimilarities:
   A  B  C  D  E
B  2
C  9  3
D  5  1  8
E 12  6 11  4
F 15 14 13 10  7

Ordination distances:
      A     B     C     D     E
B 0.301
C 0.539 0.303
D 0.323 0.133 0.421
E 0.615 0.414 0.612 0.307
F 0.922 0.636 0.636 0.605 0.416
```

MDS is a map

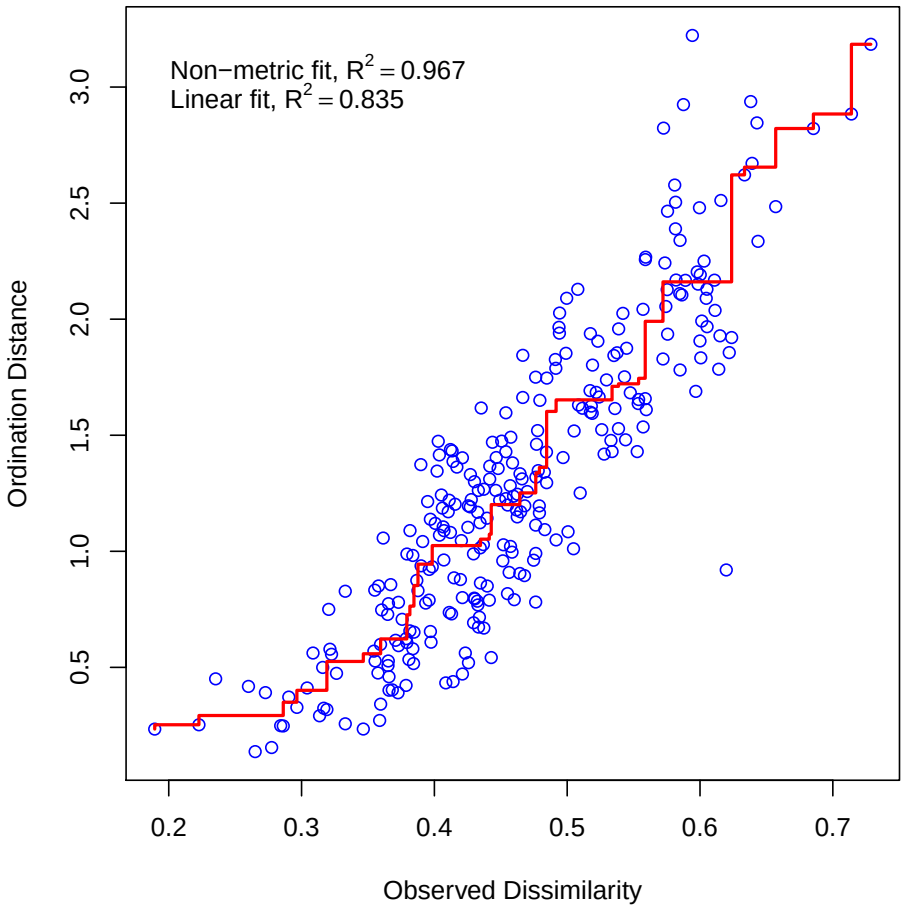
- MDS tries to draw a map using distance data.
- MDS tries to find an underlying configuration from dissimilarities.
- Only the configuration counts:
  - No origin, but only the constellations.
  - No axes or natural directions, but only a framework for points.

Map of Europe from road distances



Lambert conical conformal projection

Shepard Diagram



Iterative Optimization

Recommended procedure

NMDS may be good, but its use needs special care: Not every NMDS automatically is good

1. Use adequate dissimilarity indices: An adequate index gives a good rank-order relation between community dissimilarity and gradient distance.
2. No convergence guaranteed: Start with several random starts and inspect those with lowest stress.
3. Satisfied only if minimum stress configurations are similar.

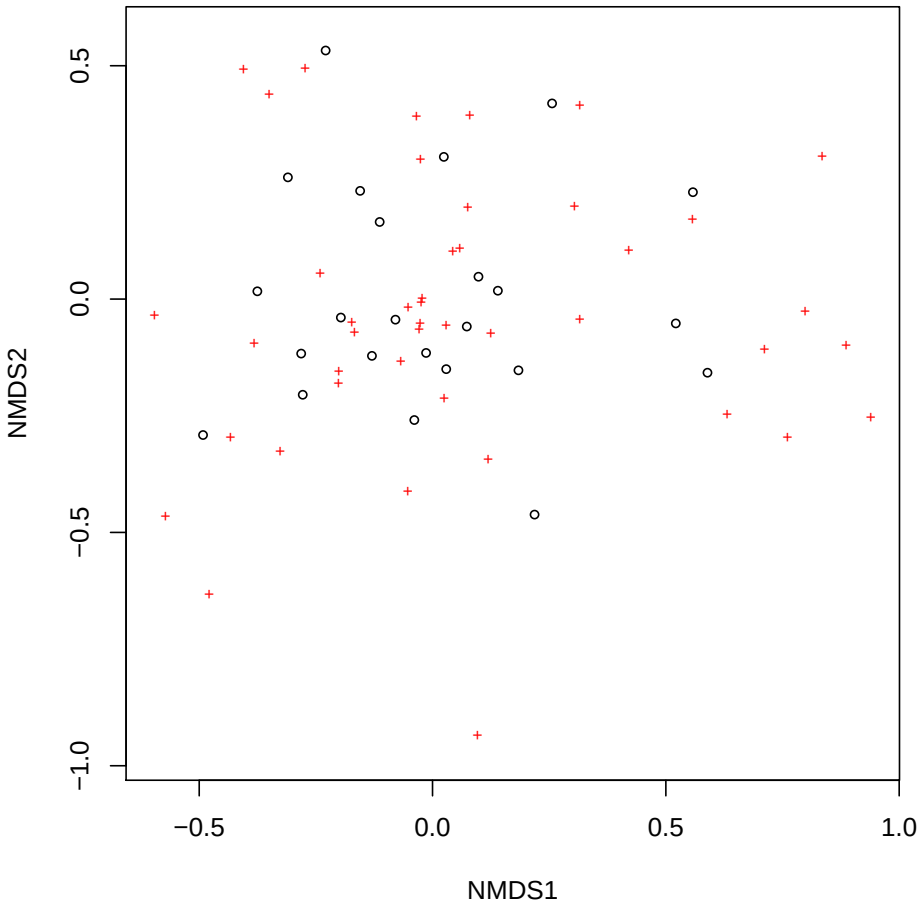
metaMDS I

```
> vare.mds <- metaMDS(varespec)

Square root transformation
Wisconsin double standardization
Run 0 stress 0.184
Run 1 stress 0.196
Run 2 stress 0.185
... procrustes: rmse 0.0494  max resid 0.158
Run 3 stress 0.209
Run 4 stress 0.215
Run 5 stress 0.235
Run 6 stress 0.196
Run 7 stress 0.234
Run 8 stress 0.196
Run 9 stress 0.222
Run 10 stress 0.185
Run 11 stress 0.195
Run 12 stress 0.229
Run 13 stress 0.184
... New best solution
... procrustes: rmse 3.6e-05  max resid 0.000139
*** Solution reached
```


metaMDS II

```
> vare.mds

Call:
metaMDS(comm = varespec)

global Multidimensional Scaling using monoMDS

Data:     wisconsin(sqrt(varespec))
Distance: bray

Dimensions: 2
Stress:     0.184
Stress type 1, weak ties
Two convergent solutions found after 13 tries
Scaling: centring, PC rotation, halfchange scaling
Species: expanded scores based on 'wisconsin(sqrt(varespec))'
```

Plot metaMDS

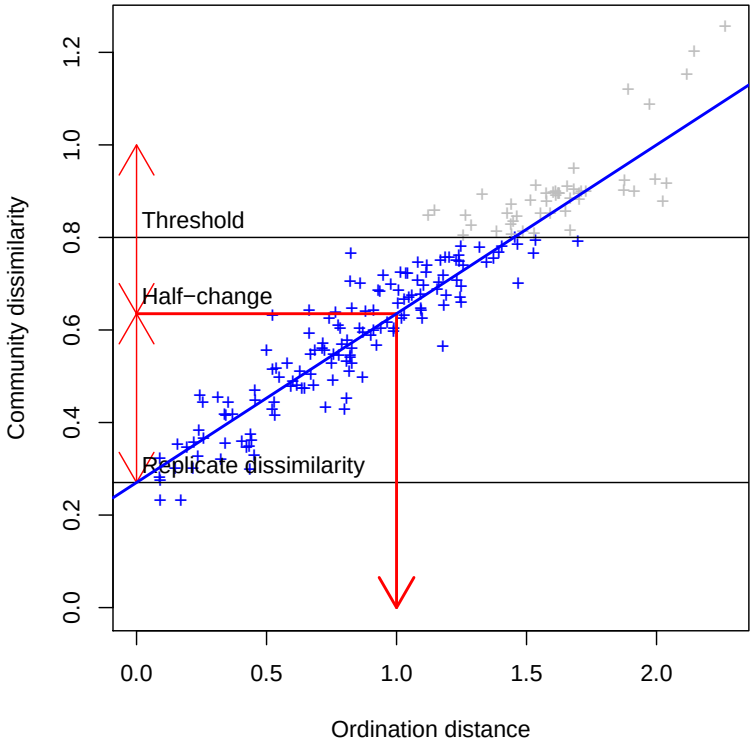


Numbers

- Badness of fit measure **stress** is based on the residuals from the non-linear regression
  - A proportional measure in the range 0 (perfect) . . . 1 (desperate) related to goodness of fit measure $1 - R^2$
  - Random configuration typically ≈ 0.4 and 0 degenerate
  - Often given in percents (but omitting the percent sign: $15 = 0.15$, since cannot be > 1)
- Orientation, rotation, scale and origin of the coordinates (scores) are indeterminate: only the constellation matters
- Vegan arbitrarily fixes some of these:
  - Axes are centred, but the origin has no special meaning
  - Axes are rotated so that the first is the longest (technically: rotated to principal components)
  - Axes are scaled so that one unit corresponds to halving of similarity from the "replicate similarity"
  - The sign (direction) of the axes still undefined

Half-change Scaling in NMDS

- Replicate similarity: dissimilarity at ordination distance = 0
- Maximum dissimilarity = 1: nothing in common
- Linear area of ordination distance – dissimilarity: 0 . . . 0.8

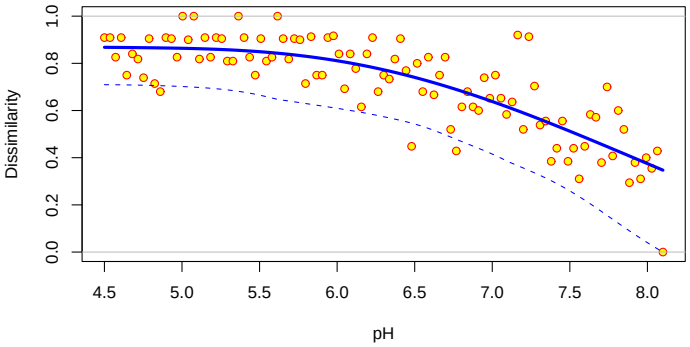


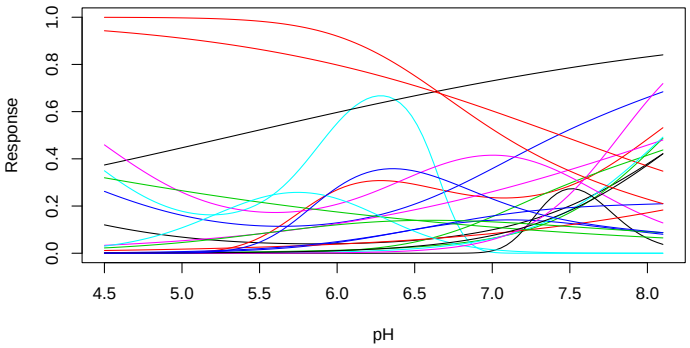
What happened in metaMDS?

1. Square root transformation and Wisconsin double standardization
2. Bray–Curtis dissimilarities
3. `monoMDS` with several random starts and stopping after finding two identical minimum stress solutions
4. Solution rotated to PCs
5. Solution scaled to half-change units
6. Species scores as weighted averages

Dissimilarity measures

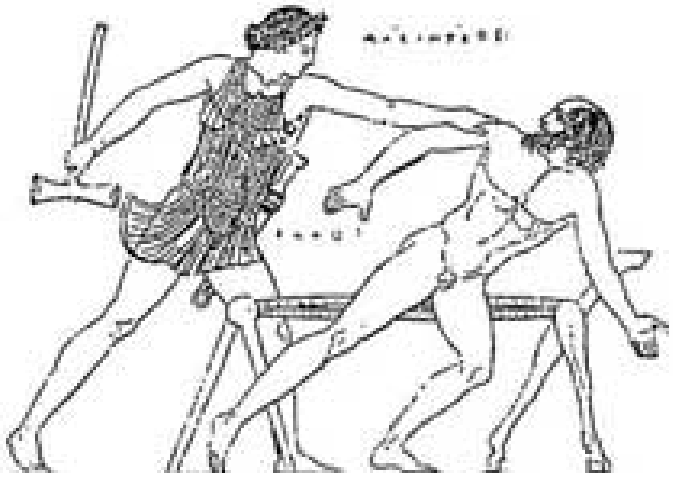
- Use a dissimilarity that describes correctly gradient separation
- Bray–Curtis (Steinhaus), Jaccard, Kulczyński
- Wisconsin double standardization often helpful
- Should use dissimilarities which reach their maximum (1) when no species are shared (like those listed above)
- Indices with no bound maximum are usually bad (Euclidean distance etc.)





Procrustes rotation

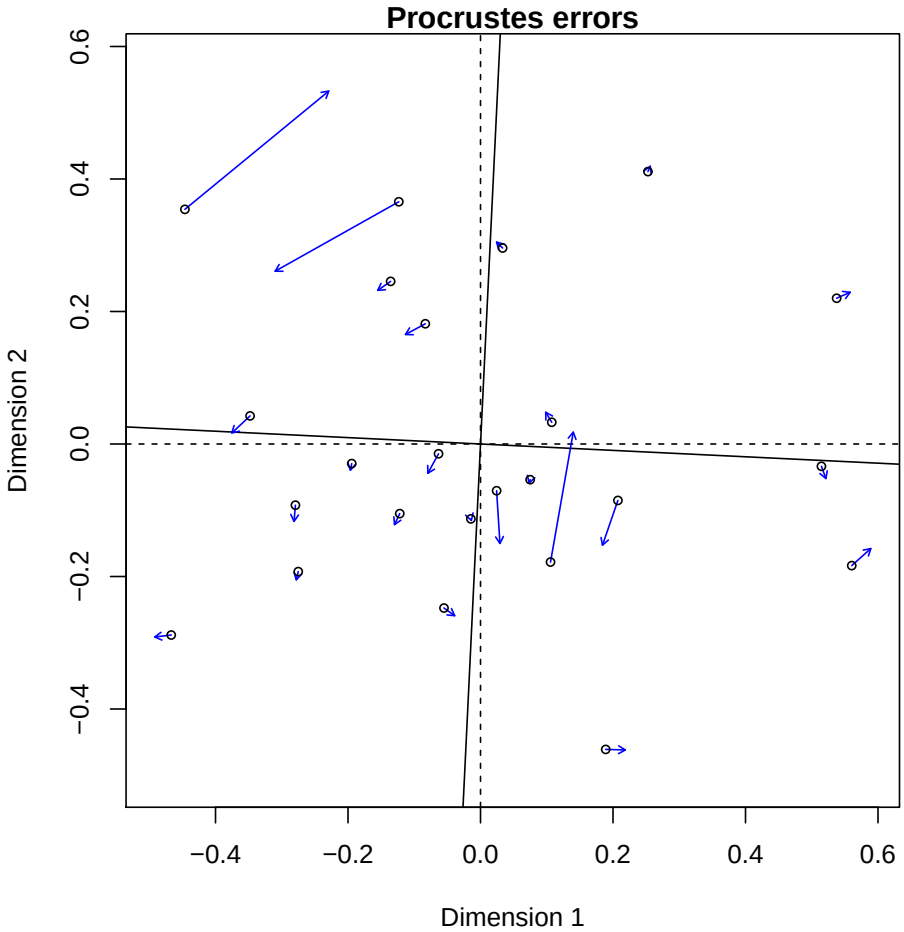
- Procrustes rotation to maximal similarity between two configurations:
  - Translate the origin.
  - Rotate the axes.
  - Deflate or inflate the axis scale.
- Single points can move a lot, although the stress is fairly constant: Especially in large data sets.



Procrustes Rotation

```
> tmp <- wisconsin(sqrt(varespec))
> dis <- vegdist(tmp)
> vare.mds0 <- monoMDS(dis, trace = 0)
> pro <- procrustes(vare.mds, vare.mds0)
> pro

Call:
procrustes(X = vare.mds, Y = vare.mds0)

Procrustes sum of squares:
0.186
```

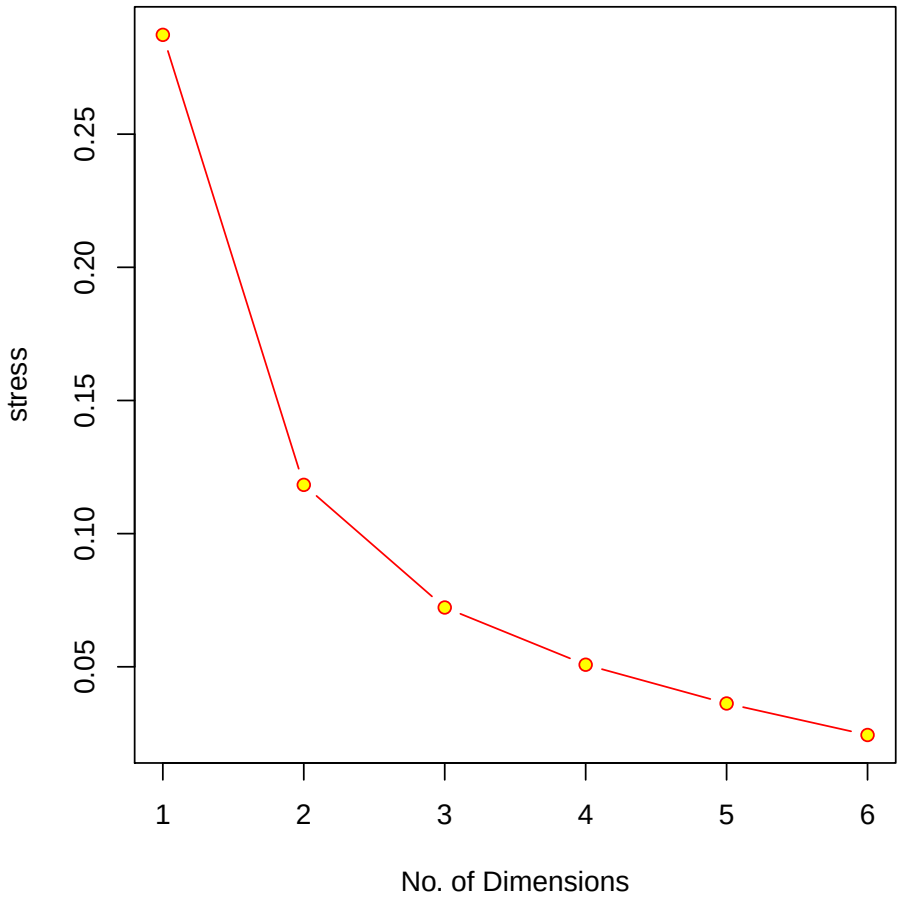

Plot Procrustes Rotations



Number of dimensions

- In NMDS, 2D solution is not a plane in 3D space
- Solution must be found separately for each dimensionality
- Some people very disturbed: how do they *know* the correct number
- Answer is easy: there is no correct number, although some numbers may be worse than others
- "Most data sets have 2.5 dimensions"
- Typically you try with 2 and 3
- Do you need more dimensions to explain species patterns and environmental data?
- Is convergence very slow? Try another number of dimensions
- *Scree plot* or stress against the number of dimensions often suggested but rarely works

Scree Plot



Simplified mapping: Eigen analysis

- NMDS uses *non-linear* mapping for *any* dissimilarity measure: This is very difficult
- Things are much simpler if we accept only certain dissimilarity indices and map them linearly onto ordination
- Linear mapping is only a rotation, and can be solved using eigenvector techniques
- Sometimes said that certain methods are *model-based* (CA), but they also employ a distance

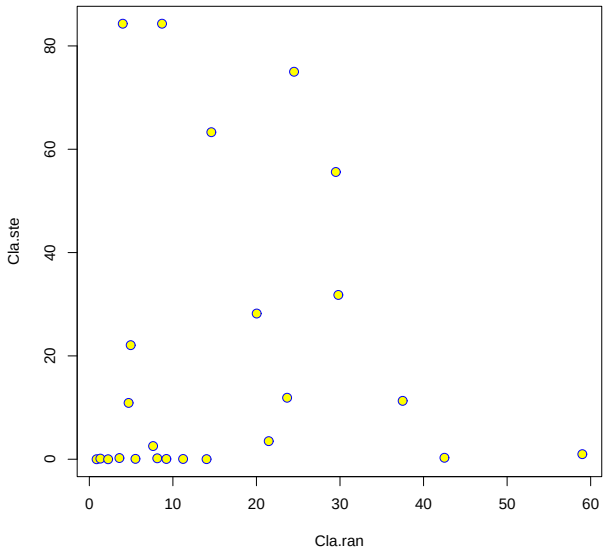
method	metric	mapping
NMDS	any	nonlinear
MDS	any	linear
PCA	Euclidean	linear
CA	Chi-square	weighted linear

Why Not PCA?

- We admit that PCA is just a rotation, but it is a linear method
- PCA works with species space, but we boldly go to gradient space
- CA is an optimal scaling method
  - Sites with similar species composition packed close to each other
  - Species that occur together simultaneously packed close to each other
- CA can handle unimodal species responses, even approximate one dimensional species packing model

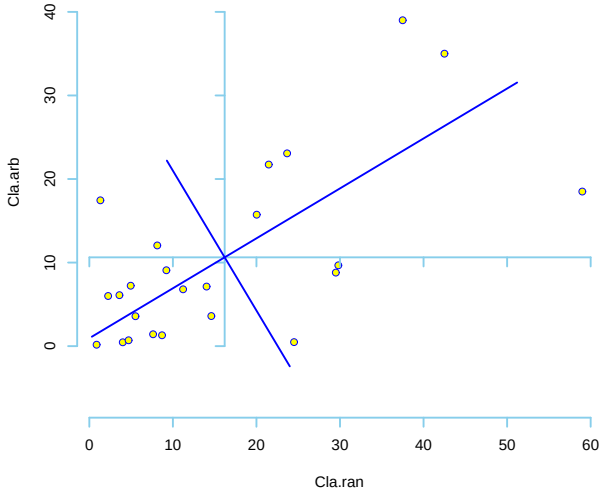
Species space

- Graphical presentations of data matrix: Species are axes and span the space where sites are points
- Some species show more of the configuration than others
- What is the ideal viewing angle to the species space?
- Shows as much as possible of all species in just two or three dimensions



Rotation in species space

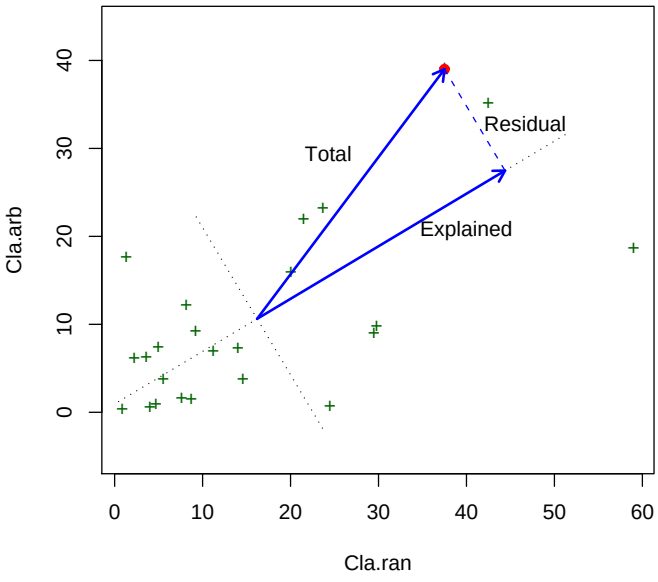
1. Put sites into species space
2. Move the origin to the centroid
3. Rotate the axes so that the first axis (1) is as close to all points as possible, and (2) explains as much of the variance as possible

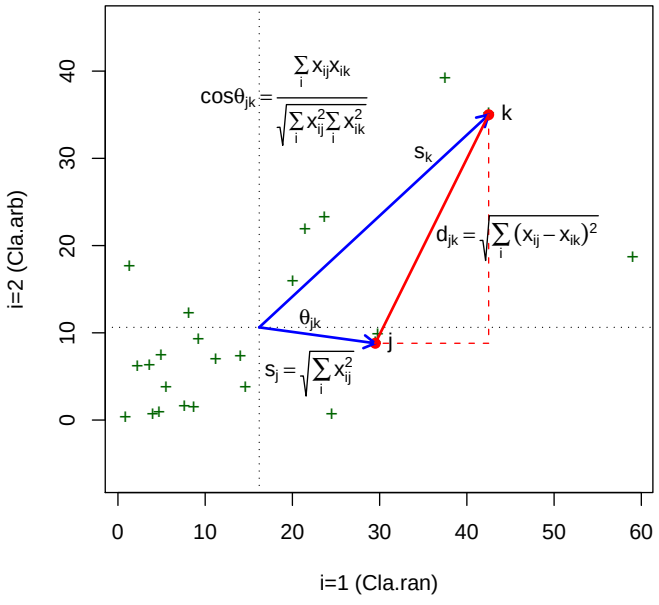


Goodness of Fit

- The total variation (Λ) is the sum of squared distances of points from the origin
- Λ can be expressed as the sum of squares (SS) or variance (SS/n or $SS/(n-1)$)
- The points are projected on the axis, and the sum of projected squared distances is the eigenvalue of the axis (λ_i)
- The eigenvalues are ordered and non-negative $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, and sum up to total variance $\Lambda = \sum_{i=1}^{p} \lambda_i$
- λ_i/Λ gives the proportion that an axis explains of the total variance, and λ_1 explains the largest proportion
- Cumulative sum gives the proportion of variance explained by the first axes: often emphasized but rather useless statistic
- PCA is often used to reduce data into a few linearly independent components that explain the most of the original variables

Euclidean Metric of PCA





Running PCA I

```
> (ord <- rda(dune))

Call: rda(X = dune)

              Inertia Rank
Total            84.1
Unconstrained    84.1   19
Inertia is variance

Eigenvalues for unconstrained axes:
  PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8
24.80 18.15  7.63  7.15  5.70  4.33  3.20  2.78
(Showed only 8 of all 19 unconstrained eigenvalues)

> head(summary(ord), 3, 1)
```


Running PCA II

```
Call:
rda(X = dune)

Partitioning of variance:
              Inertia Proportion
Total            84.1          1
Unconstrained    84.1          1

Eigenvalues, and their contribution to the variance

Importance of components:
                          PC1    PC2    PC3   PC4    PC5    PC6
Eigenvalue             24.795 18.147 7.6291 7.153 5.6950 4.3333
Proportion Explained    0.295  0.216 0.0907 0.085 0.0677 0.0515
Cumulative Proportion   0.295  0.510 0.6011 0.686 0.7539 0.8054
                          PC7    PC8    PC9  PC10   PC11   PC12
Eigenvalue             3.199 2.7819 2.4820 1.854 1.7471 1.3136
Proportion Explained   0.038 0.0331 0.0295 0.022 0.0208 0.0156
Cumulative Proportion  0.843 0.8765 0.9060 0.928 0.9488 0.9644
                         PC13    PC14    PC15    PC16    PC17
Eigenvalue             0.9905 0.63779 0.55083 0.35058 0.19956
Proportion Explained   0.0118 0.00758 0.00655 0.00417 0.00237
```

Running PCA III

```
Cumulative Proportion 0.9762 0.98377 0.99032 0.99448 0.99686
                         PC18    PC19
Eigenvalue            0.14880 0.11575
Proportion Explained  0.00177 0.00138
Cumulative Proportion 0.99862 1.00000

Scaling 2 for species and site scores
* Species are scaled proportional to eigenvalues
* Sites are unscaled: weighted dispersion equal on all dimensions
* General scaling constant of scores:  6.3229


Species scores

             PC1    PC2      PC3    PC4     PC5     PC6
Achimill -0.6038  0.124  0.00846  0.160  0.4087  0.1279
Agrostol  1.3740 -0.964  0.16691  0.266 -0.0877  0.0474
Airaprae  0.0234  0.251 -0.19477 -0.326  0.0557 -0.0796
....
Callcusp  0.5385  0.180  0.17509  0.239  0.2553  0.1692
```

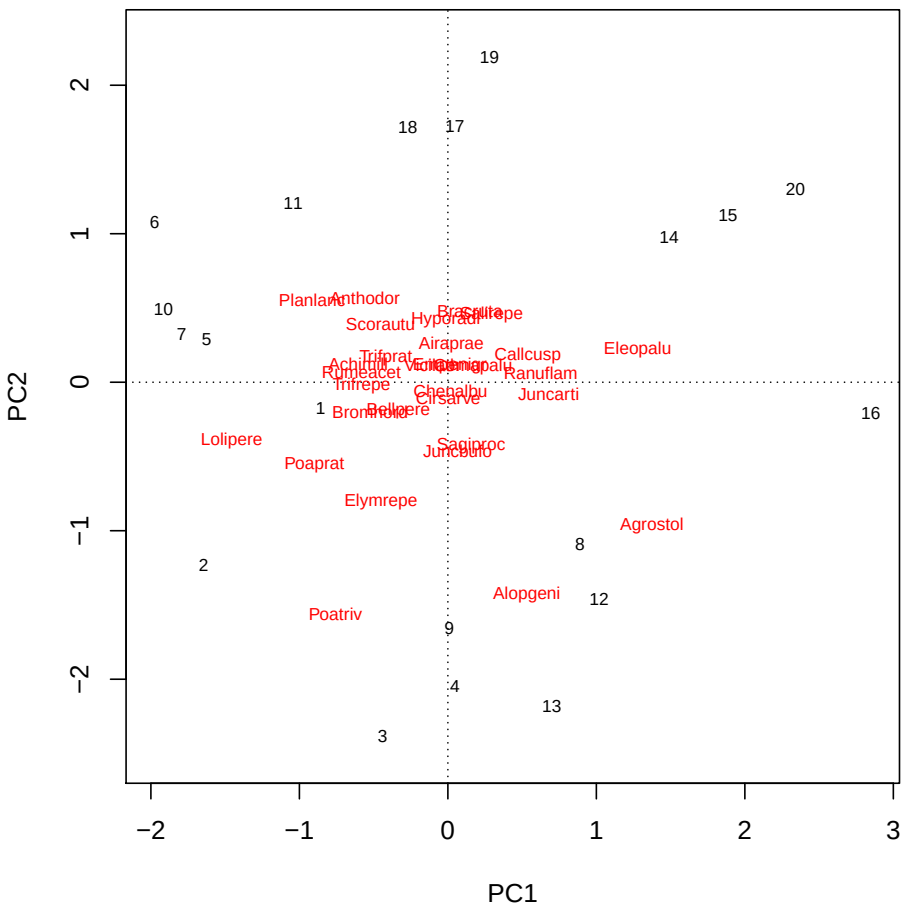

Running PCA IV

```
Site scores (weighted sums of species scores)

        PC1    PC2   PC3    PC4     PC5     PC6
1    -0.857 -0.172 2.608 -1.130  0.4507 -2.4911
2    -1.645 -1.230 0.887 -0.986  2.0346  1.8106
3    -0.440 -2.383 0.930 -0.460 -1.0278 -0.0518
....
20    2.341  1.299 0.903  0.718 -0.0757 -0.9691
```

Row and Column scores

- The scores are centred (= their mean is zero) and either normalized (= all have equal spread) or proportional to eigenvalues (= spread is higher when eigenvalue is high)
- Normalized scores give the regression coefficients between the axis and the variables: often used for species
- Scores proportional to the eigenvalue give the true configuration of points in the space defined by normalized scores: often used for sites (hence in species space)
- Together these scores give a linear least square approximation of the data
- Graphical presentation called **biplot**
- However, there are many alternative scaling systems

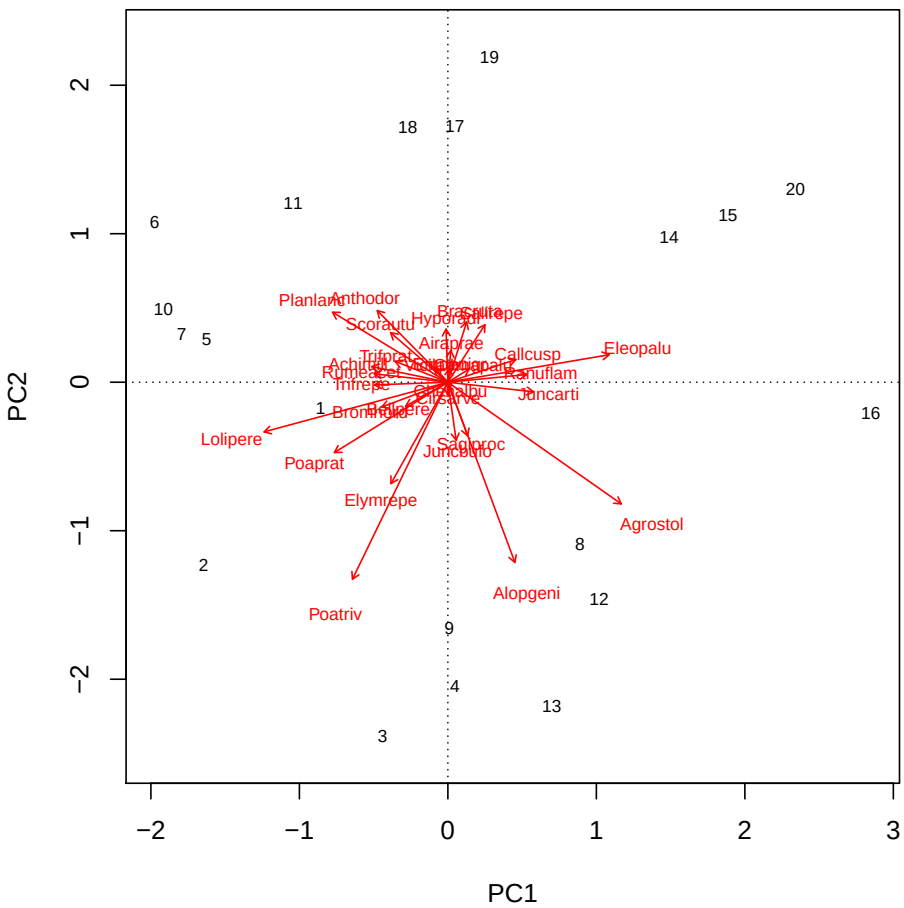
Default Plot



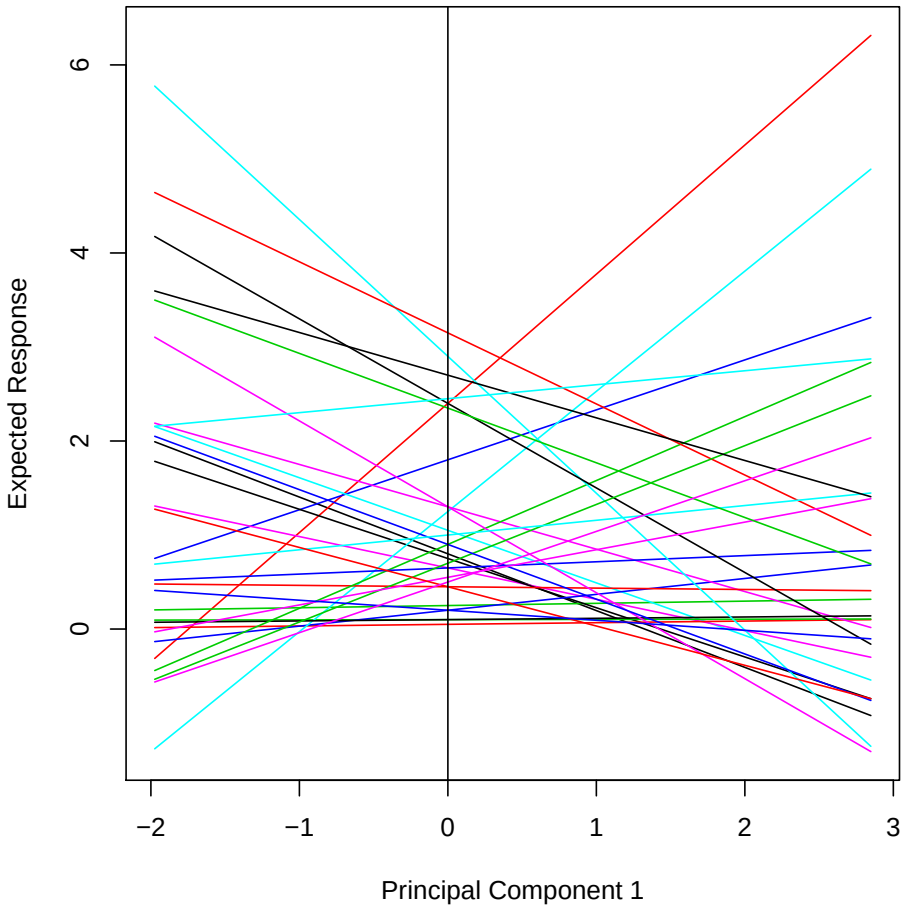
Reading the Plot

- Origin: all species (variables) at their average values
- The *distance* from the origin for a row (site) implies how much the point differs from the average
- The *distance* from the origin for a column (species, variable) implies how much the point increases to that *direction*
- The change is measured in absolute scale: big changes, long distances from the origin
- Implies a linear model of species response against axes
- The *angle* between two points implies correlations
- 90° means zero correlation, $< 90^\circ$ positive correlation, $> 90^\circ$ negative correlation, 0° implies $r = 1$
- Arrow biplots often used instead of point biplot

Arrow Biplot



Linear Model



Variances and Correlations

- Analysis of raw data explains variances: variables with high variance are most important
- If the variables are standardized to unit variance before analysis $z = (x - \bar{x})/s_x$ all variables are equally important and the analysis explains correlations among variables
- Standardization can be used when we want all variables to have equal weights
- Standardization must be used when variables are measured in different scales, such as for environmental measurements

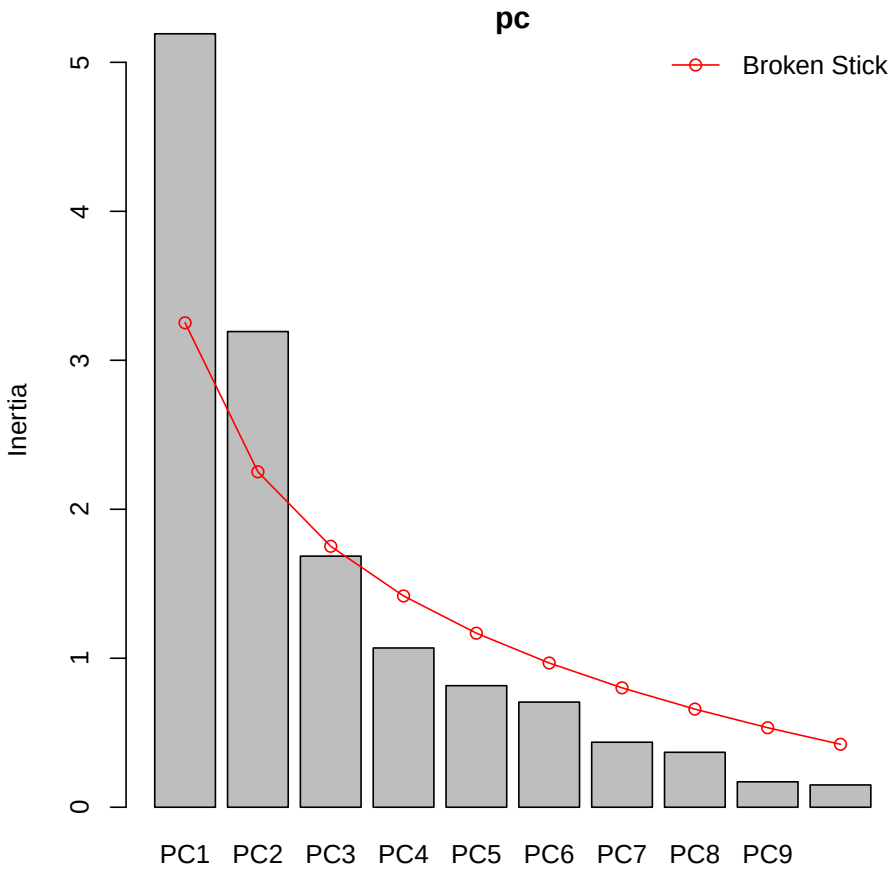
Reducing the Number of Correlated Environmental Variables I

```
> (pc <- rda(varechem, scale=TRUE))

Call: rda(X = varechem, scale = TRUE)

              Inertia Rank
Total              14
Unconstrained      14   14
Inertia is correlations

Eigenvalues for unconstrained axes:
 PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9 PC10 PC11 PC12
5.19 3.19 1.69 1.07 0.82 0.71 0.44 0.37 0.17 0.15 0.09 0.07
PC13 PC14
0.04 0.02
```

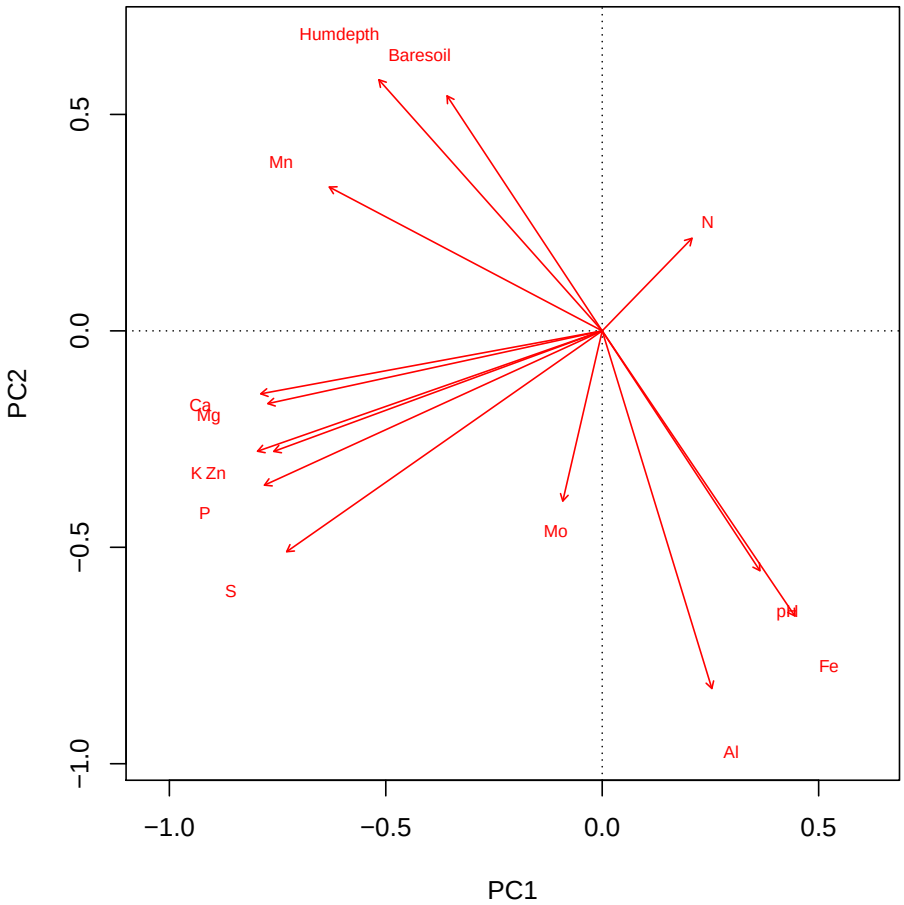

The Number of Components

- PCA is a rotation in species (character) space and retains the original configuration
- The number of PC's is $\min(N, S)$, and all together give the original data
- First axes are most important and we may ignore the minor axes
- We can either use the axes as variables in other models, or use them to identify major (almost) independent variables
- Often we want to retain a certain proportion of the variance, say 50 %
- Sometimes we would like to retain "significant" axes
- There really is no way of doing this, but some people suggest comparing eigenvalues against *broken stick* distribution

Broken Stick and Eigenvalues



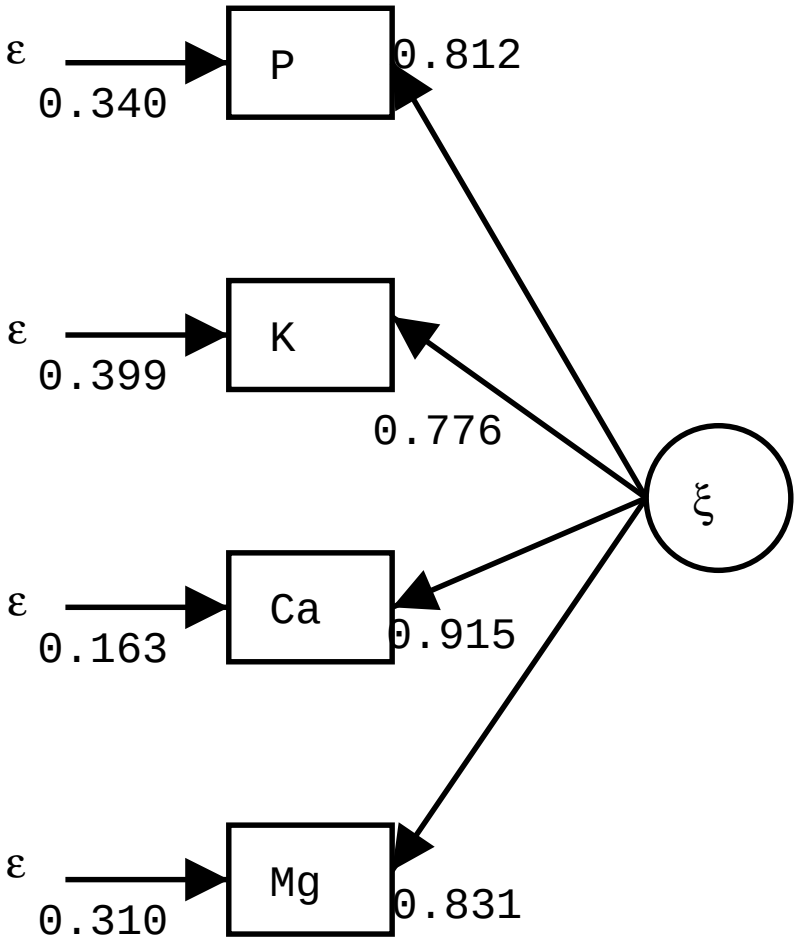
Two Dimensions, but which?



Methods Related to PCA

- **Metric Scaling** a.k.a. **Principal Coordinates Analysis**
  - Used dissimilarities instead of raw data
  - With Euclidean distances equal to PCA, but can use other dissimilarities
- **Factor Analysis**
  - A *statistical* method that makes a difference between systematic components and random error
  - In PCA we just ignore latter components, but here we really identify the real components
  - Much used in human sciences and often referred to in ecology (but usually misunderstood)

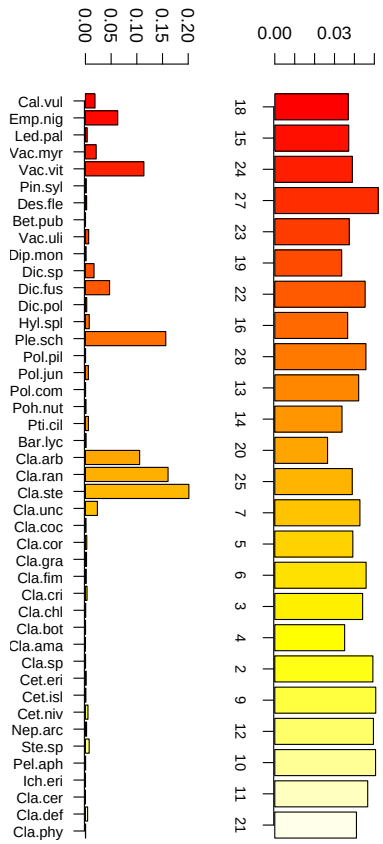
Confirmatory Factor Analysis



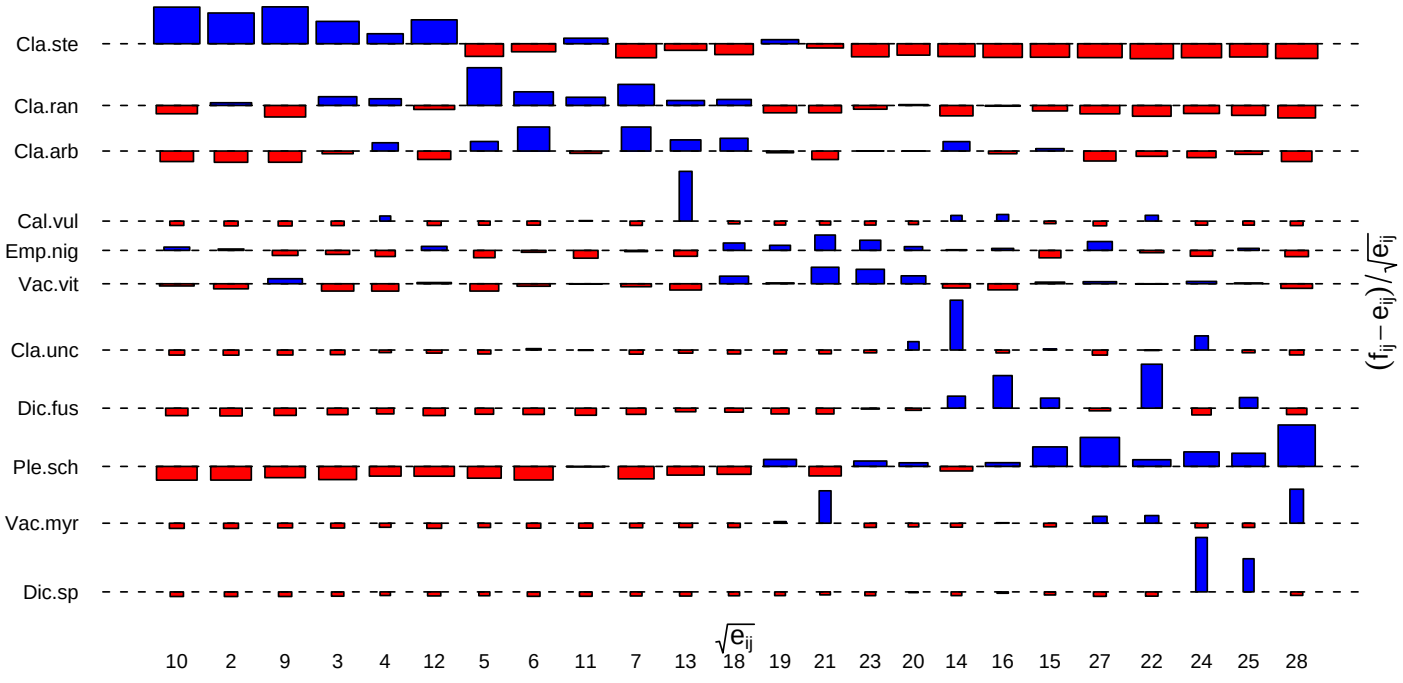
Correspondence Analysis

- Minor variant of PCA: Weighted Principal Components with Chi-square metric
- All sites should have all species in in the same proportions as in the whole data
- Site and species marginal profiles define the expected abundances
- Null model: Species *composition* is identical in all sampling units
- Chi-square transformation tells how much the observed proportions f_{ij} differ from the expected proportions e_{ij}:

$$\chi_{ij} = \frac{f_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

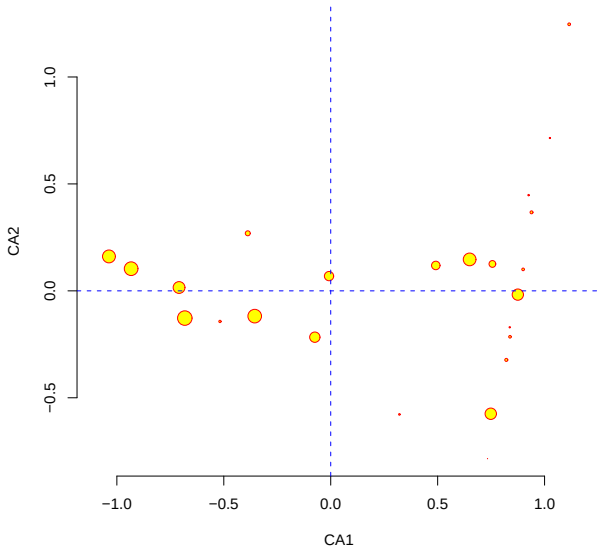


Chi-squared metric



CA Rotation

1. Sites in a species space
2. Relative proportions are axes and points have weights
3. Chi-square transformation
4. Weighted rotation
5. De-weighting



Running CA I

```
> (ord <- cca(dune))

Call: cca(X = dune)

              Inertia Rank
Total            2.12
Unconstrained    2.12   19
Inertia is mean squared contingency coefficient

Eigenvalues for unconstrained axes:
  CA1   CA2   CA3   CA4   CA5   CA6   CA7   CA8
0.536 0.400 0.260 0.176 0.145 0.108 0.092 0.081
(Showed only 8 of all 19 unconstrained eigenvalues)

> head(summary(ord), 2, 1)
```

Running CA II

```
Call:
cca(X = dune)

Partitioning of mean squared contingency coefficient:
              Inertia Proportion
Total            2.12          1
Unconstrained    2.12          1

Eigenvalues, and their contribution to the mean squared contingency coefficient

Importance of components:
                         CA1   CA2   CA3    CA4    CA5   CA6
Eigenvalue            0.536 0.400 0.260 0.1760 0.1448 0.108
Proportion Explained  0.253 0.189 0.123 0.0832 0.0684 0.051
Cumulative Proportion 0.253 0.443 0.565 0.6486 0.7170 0.768
                          CA7    CA8    CA9   CA10   CA11
Eigenvalue            0.0925 0.0809 0.0733 0.0563 0.0483
Proportion Explained  0.0437 0.0382 0.0347 0.0266 0.0228
Cumulative Proportion 0.8117 0.8500 0.8847 0.9113 0.9341
                         CA12   CA13    CA14    CA15    CA16
Eigenvalue            0.0412 0.0352 0.02053 0.01491 0.00907
Proportion Explained  0.0195 0.0167 0.00971 0.00705 0.00429
```


Running CA III

```
Cumulative Proportion 0.9536 0.9702 0.97995 0.98700 0.99129
                          CA17    CA18    CA19
Eigenvalue            0.00794 0.00700 0.00348
Proportion Explained  0.00375 0.00331 0.00164
Cumulative Proportion 0.99505 0.99836 1.00000

Scaling 2 for species and site scores
* Species are scaled proportional to eigenvalues
* Sites are unscaled: weighted dispersion equal on all dimensions


Species scores

            CA1     CA2    CA3      CA4    CA5     CA6
Achimill -0.909  0.0846 -0.586 -0.00892 -0.660 -0.1888
Agrostol  0.934 -0.2065  0.282  0.02429 -0.139 -0.0226
....
Callcusp  1.952  0.5674 -0.859 -0.09897 -0.557  0.2328


Site scores (weighted averages of species scores)
```

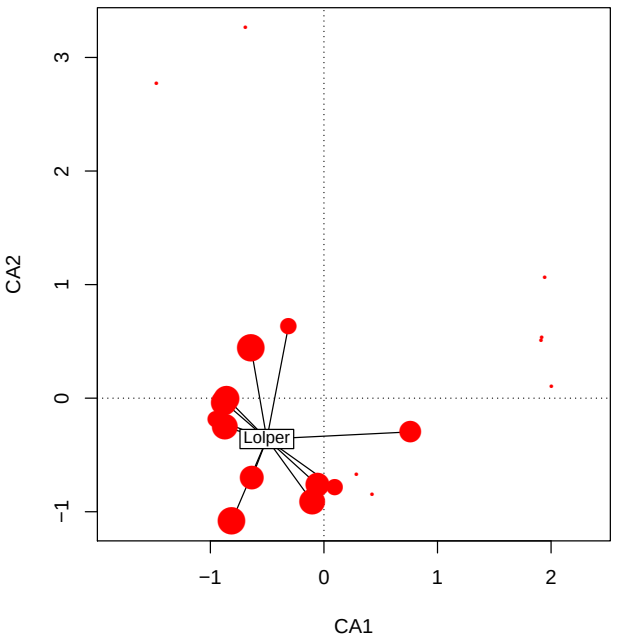
Running CA IV

```
        CA1    CA2     CA3    CA4    CA5     CA6
1    -0.812 -1.083 -0.1448 -2.107 -0.393 -1.8346
2    -0.633 -0.696 -0.0971 -1.187 -0.977  0.0658
....
20    1.944  1.069 -0.6660 -0.553  1.596 -1.7029
```


Goodness of Fit of Scores

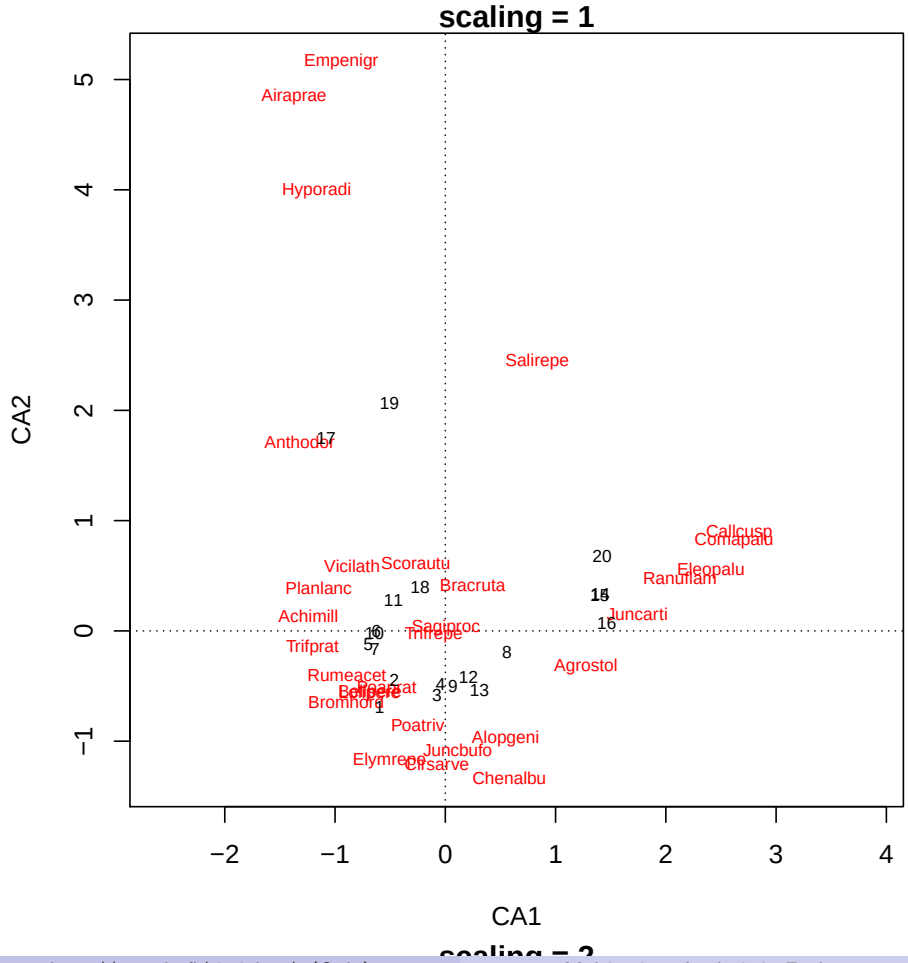
- Inertia is "mean square contingency coefficient": Chi-squared of a matrix standardized to unit sum, or Chi-square of $\frac{x}{\sum x}$
- Eigenvalues are non-negative and ordered like in PCA, but they are bound to maximum 1
- The origin gives the expected abundances for all species and all sites
- The deviant species and deviant sites are far away from the origin
- CA is weighted analysis, and the weighted sum of squared scores is the eigenvalue
- The species and site scores are (scaled) weighted averages of each other: proximity matters
- Rare species have low weights: they are further away from the origin

Weighted Average?



- For presence/absence data: weighted average of a species is in the middle ("barycentre") of plots where the species occurs
- For quantitative data: plots wheres species is abundant are heavier and the weighted average is closer to them
- Sampling units (SU) are close to species that occur on them
- CA is a weighted average method: it tries to put SUs close to the species that occur in them, and all SUs with similar species composition close to each other: Unimodal response

Default Plot and Effect of Scaling

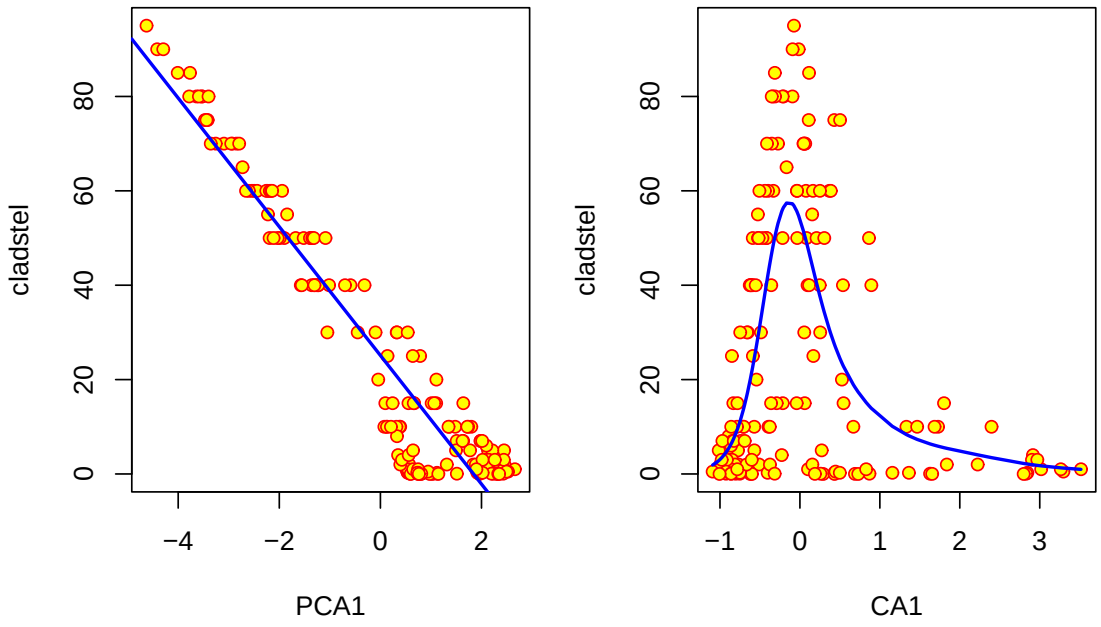


Weighted Averages

- Species scores are [proportional to] weighted averages of site scores, and simultaneously
- Site scores are [proportional to] weighted averages of species scores
- Either one (but not both) of these can be a direct weighted average of other
- If sites scores are weighted averages of species scores, site point is in the middle of points of species that occurs in the site
- The *location* of the point is meaningful whereas in PCA the main things were *distance* and *direction* from the origin (but these, too, matter)
- Can approximate unimodal response model and therefore CA is **much better** for community ordination than PCA

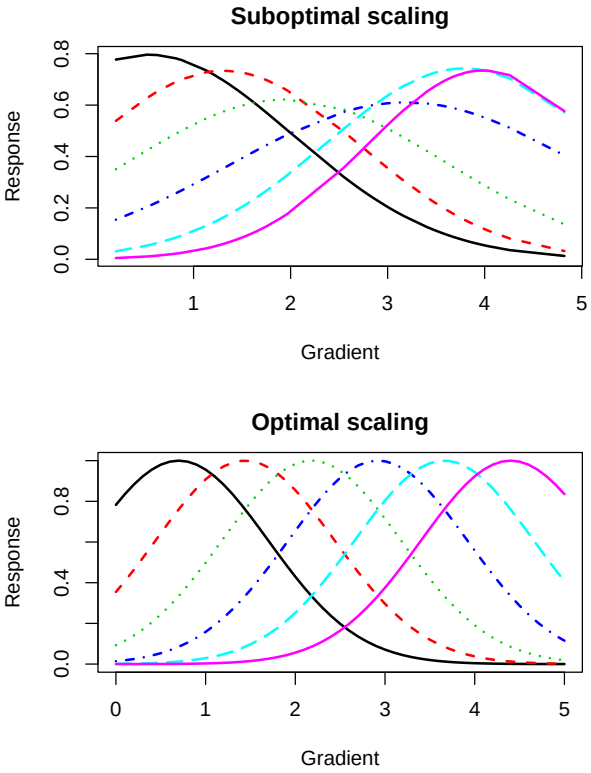
Linear and Unimodal Models

- PCA implies linear relations between axes and species abundances
- CA packs species and approximates a unimodal model



Optimal Scaling

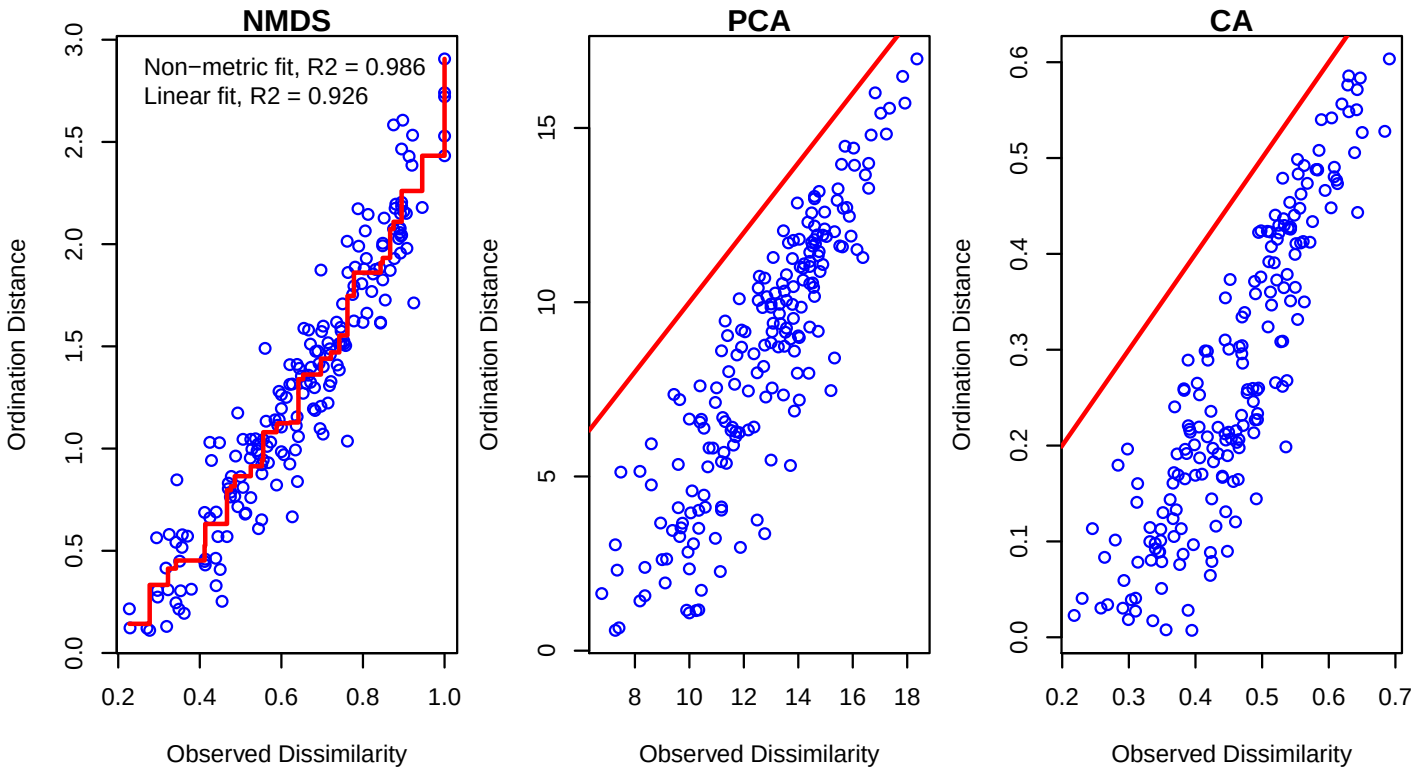
- The locations of species optima (tops) should be widespread: spread is measured as SS_B
- The species responses should be narrow: width is measured as SS_w
- The total variance is their sum $SS_T = SS_B + SS_w$
- High SS_B means that species have different optima, and low SS_w means that species have narrow tolerance
- Scaling is optimal if most of variance is between species and SS_B is high
- The criterion of variance is the eigenvalue maximized in CA: $\lambda = SS_B/SS_T$



Goodness of Fit Statistics: Repetition

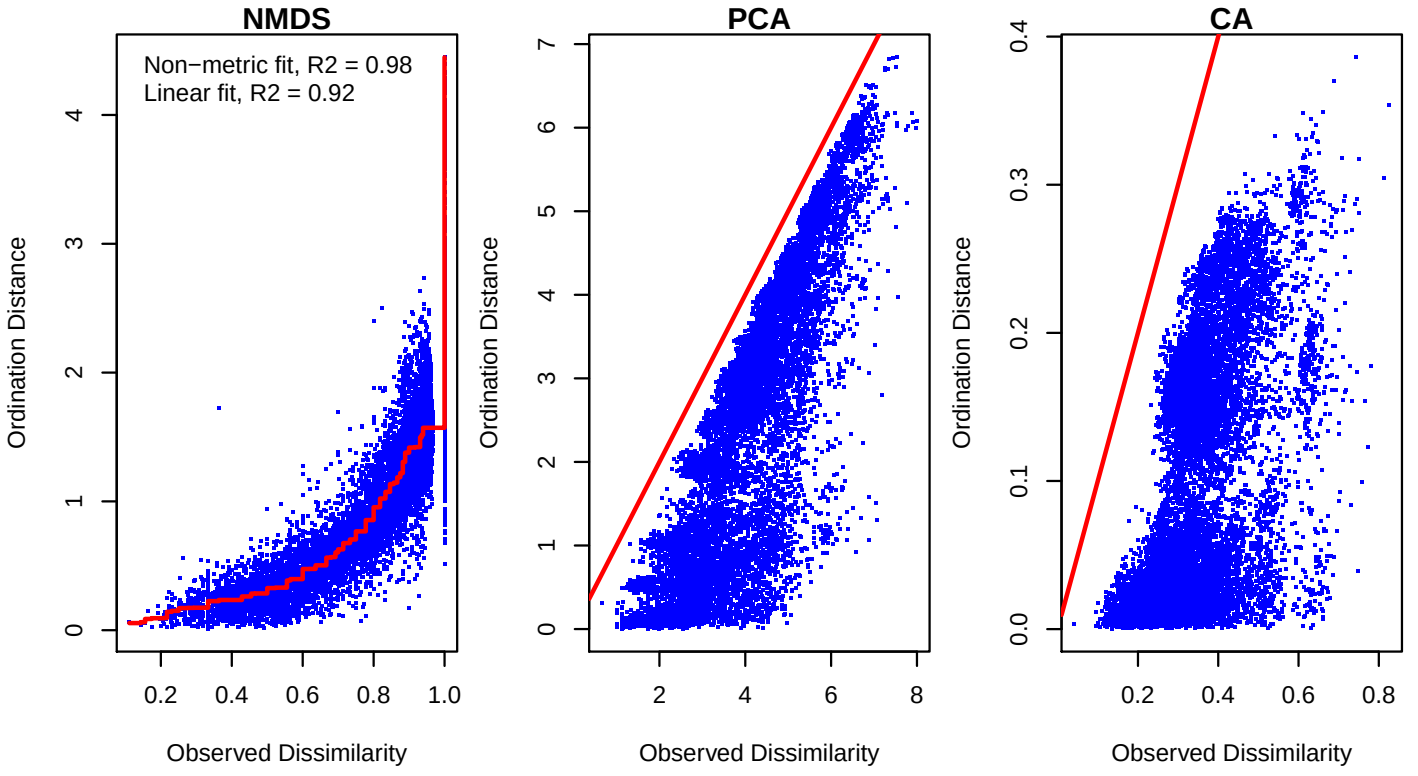
- **NMDS**: stress of nonlinear transformation from observed dissimilarities to ordination distances
  - In range 0 . . . 1 (0 . . . 100 %), but in practice 0.4 for random configuration
  - 0.1 is good, and 0.2 is not bad, 0 is suspect
- **PCA**: sum of eigenvalues is variance (or SS)
  - Upper limit is total variance, large is good
- **CA**: sum of all eigenvalues is (scaled) Chi-square
  - Single eigenvalue maximum 1
  - high is good, but $\lambda < 0.2$ may not be bad
  - Eigenvalues $\lambda > 0.7$ are suspect: disjunct or very heterogeneous data

Nonlinear and Linear Mapping



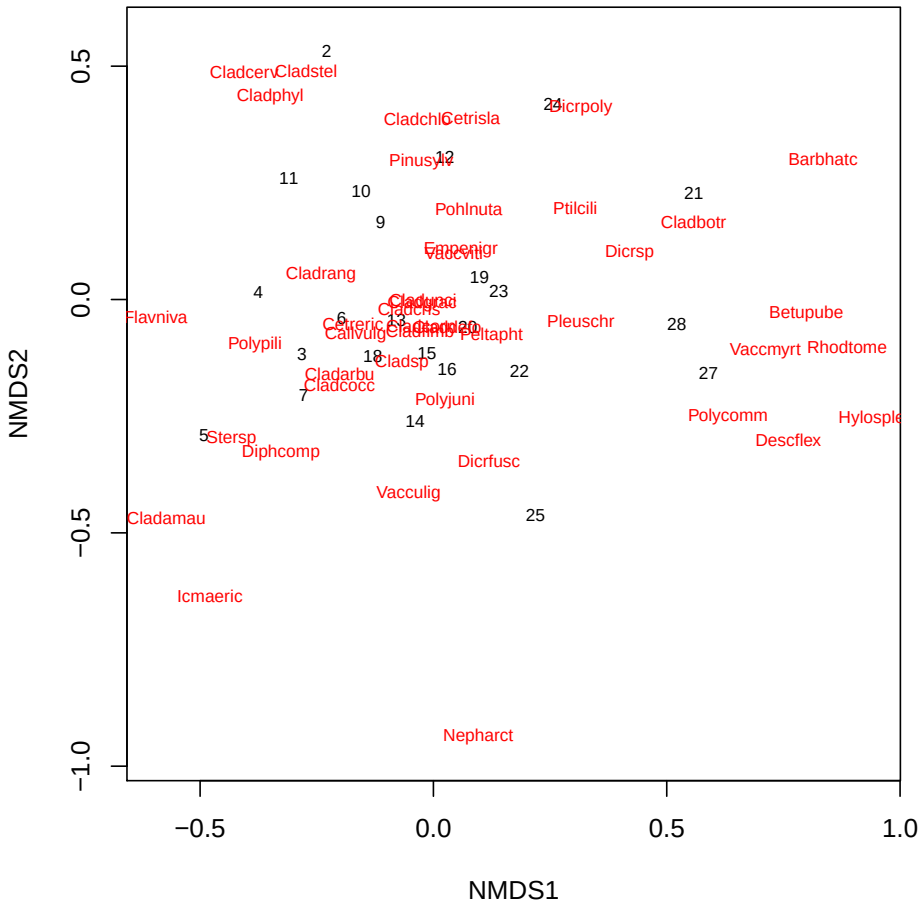
Dune Meadow data

Nonlinear and Linear Mapping: A Difficult Case



Bryce Canyon

Anatomy of a Plot



Plotting functions

- All vegan ordination functions have a `plot` function, and `ordiplot` can be used for other functions as well
- For full control, use first `plot(x, type="n")` and then add configurable `points` or `text`
- Congested plots can displayed with `orditorp` or edited with `orditkplot`
- Lattice graphics can be made with `ordixyplot`, `ordicloud` or `ordisplom`
- Dynamic, spinnable 3D plots can be made with `ordirgl` function in the **vegan3d** package
- Items can be added to the plots with `ordiarrows`, `ordihull`, `ordispider`, `ordihull`, `ordiellipse`, `ordisegments`, or `ordigrid`

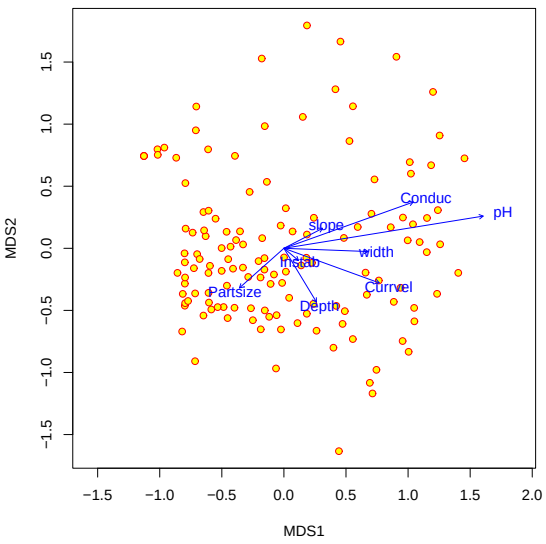
Ordination and Environment

We take granted that vegetation is controlled by environment, so

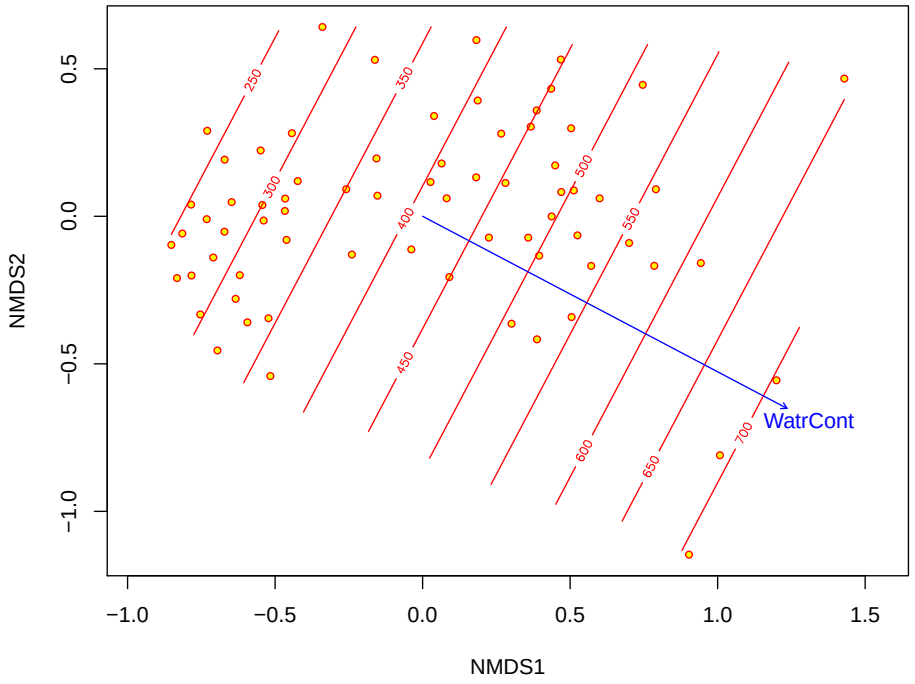
1. Two sites close to each other in ordination have similar vegetation
2. If two sites have similar vegetation, they have similar environment
3. Two sites far away from each other in ordination have dissimilar vegetation, and perhaps
4. If two sites have different vegetation, they have different environment

Fitted Vectors

- **Direction** of fitted vector shows the gradient of the environmental variable, **length** shows its importance.
- For every arrow, there is an equally long arrow into opposite direction: Decreasing direction of the gradient.
- Implies a linear model: Project sample plots onto the vector for expected value.

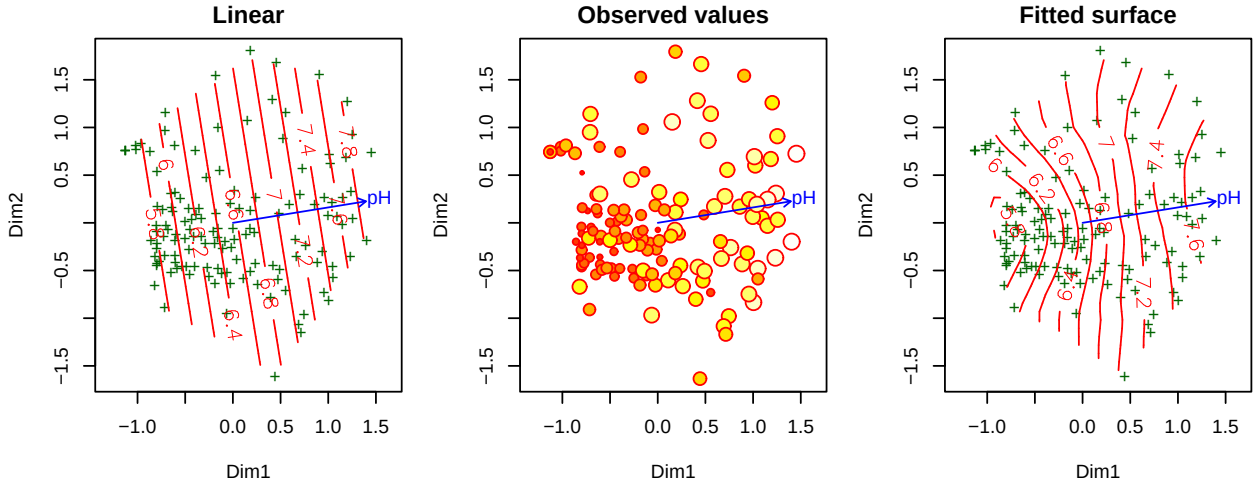


Interpretation of Arrow



Alternatives to Vectors

- Fitted vectors natural in constrained ordination, since these have linear constraints.
- Distant sites are different, but may be different in various ways: Environmental variables may have a non-linear relation to ordination.

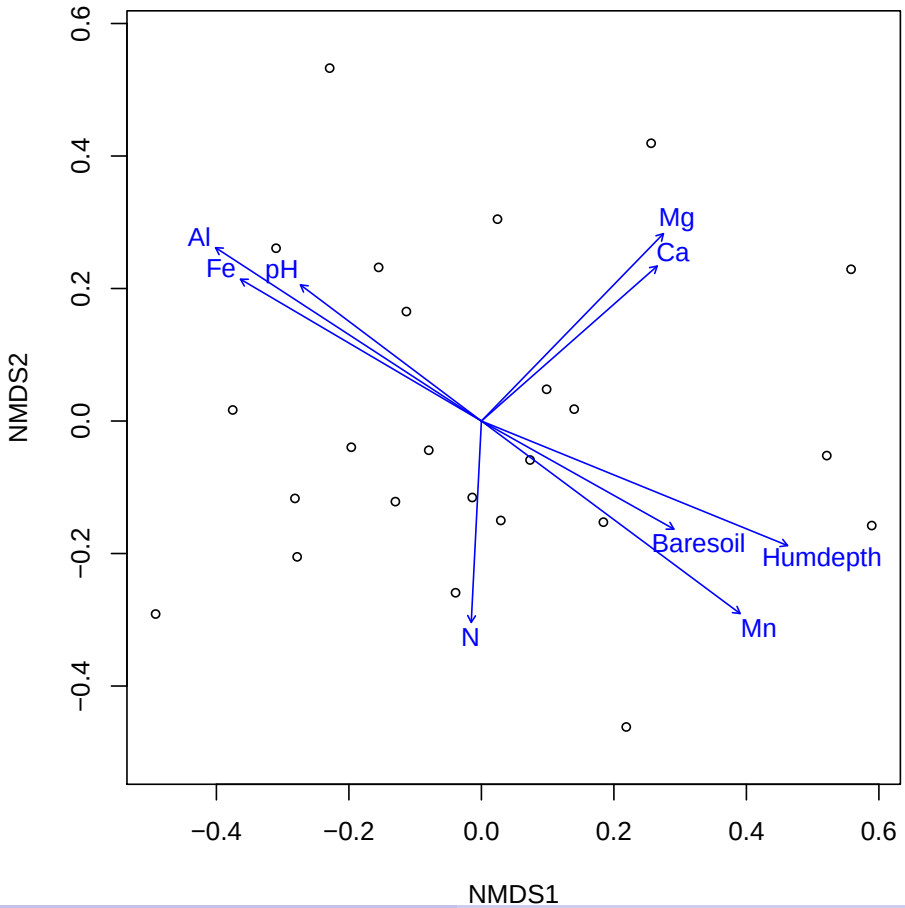


Fitting Environmental Vectors I

```
> (ef <- envfit(vare.mds, varechem, permu = 999))

***VECTORS

           NMDS1  NMDS2   r2 Pr(>r)
N         -0.050 -0.999 0.21  0.098 .
P          0.687  0.727 0.18  0.135
K          0.827  0.562 0.17  0.147
Ca         0.750  0.661 0.28  0.029 *
Mg         0.697  0.717 0.35  0.015 *
S          0.276  0.961 0.18  0.143
Al        -0.838  0.546 0.52  0.002 **
Fe        -0.862  0.507 0.40  0.013 *
Mn         0.802 -0.597 0.53  0.001 ***
Zn         0.665  0.747 0.18  0.146
Mo        -0.849  0.529 0.05  0.581
Baresoil   0.872 -0.490 0.25  0.035 *
Humdepth   0.926 -0.377 0.56  0.001 ***
pH        -0.799  0.601 0.26  0.042 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Permutation: free
Number of permutations: 999
```


Plotting Environmental Vectors

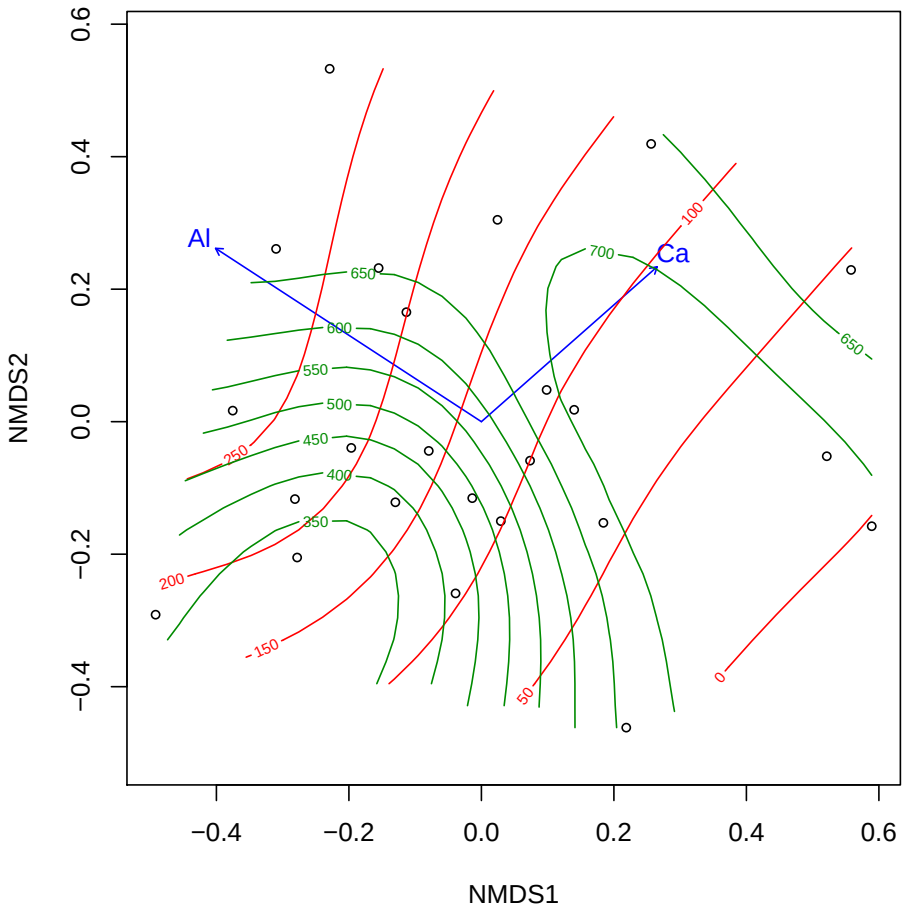
Limit $p < 0.1$



Fitting Environmental surfaces

```
> ef <- envfit(vare.mds ~ Al + Ca, varechem)
> plot(vare.mds, display = "sites")
> plot(ef)
> tmp <- with(varechem, ordisurf(vare.mds, Al, add = TRUE))
> tmp <- with(varechem, ordisurf(vare.mds, Ca, add = TRUE, col = "green4"))
```


Plotting Environmental Surfaces



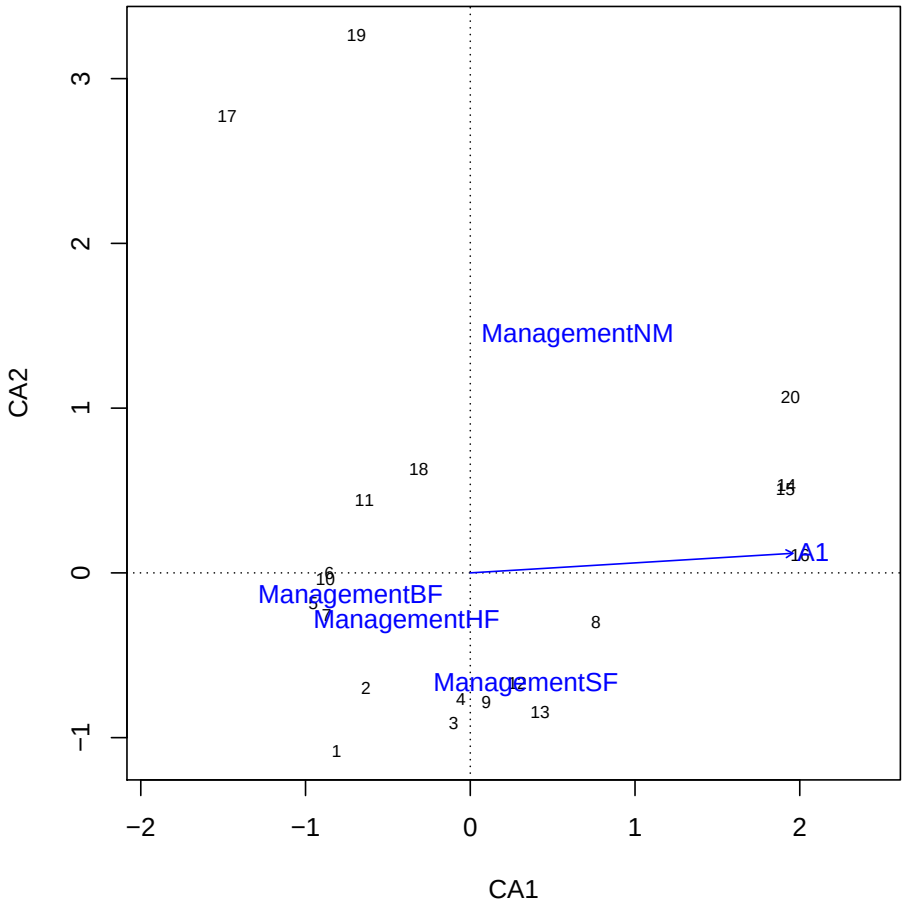
Factor Fitting I

```
> dune.ca <- cca(dune)
> ef <- envfit(dune.ca ~ A1 + Management, data=dune.env, perm=999)
> ef

***VECTORS

      CA1    CA2   r2 Pr(>r)
A1 0.9980 0.0606 0.31  0.052 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Permutation: free
Number of permutations: 999

***FACTORS:

Centroids:
               CA1   CA2
ManagementBF -0.73 -0.14
ManagementHF -0.39 -0.30
ManagementNM  0.65  1.44
ManagementSF  0.34 -0.68
```


Factor Fitting II

```
Goodness of fit:
             r2 Pr(>r)
Management 0.44  0.003 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Permutation: free
Number of permutations: 999
```

Plotting Fitted Factors

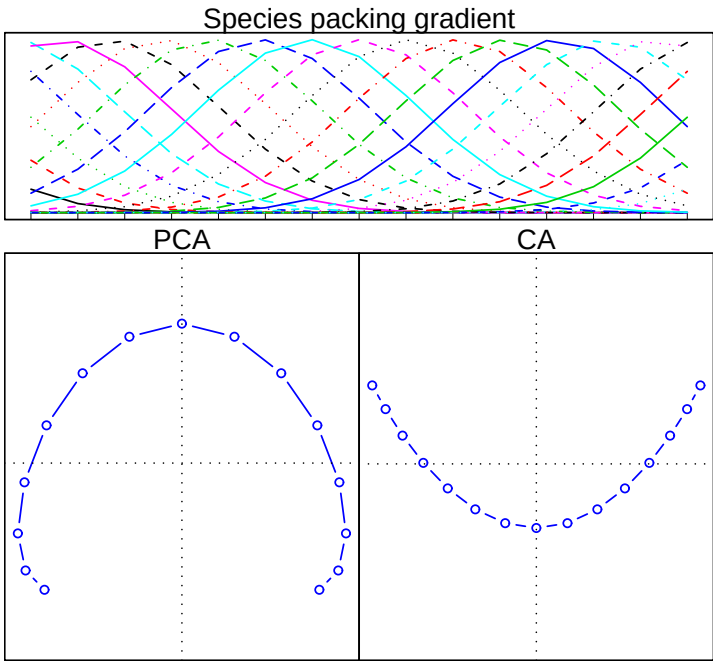


Environmental Interpretation

- Environmental variables need not be parallel to ordination axes.
- Axes cannot be taken as gradients, but gradients are oblique to axes: You cannot tear off an axis from an ordination.
- **Never** calculate a correlation between an axis and an environmental variable.
- Environmental variables need not be linearly correlated with the ordination, but locations in ordination can be exceptional.

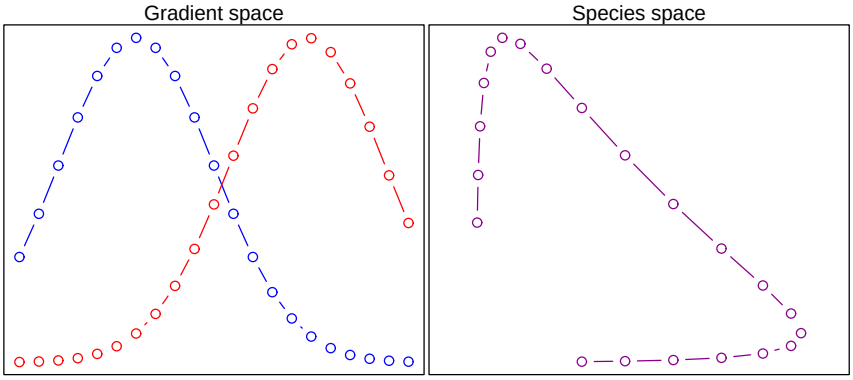
Gradient Model and Ordination

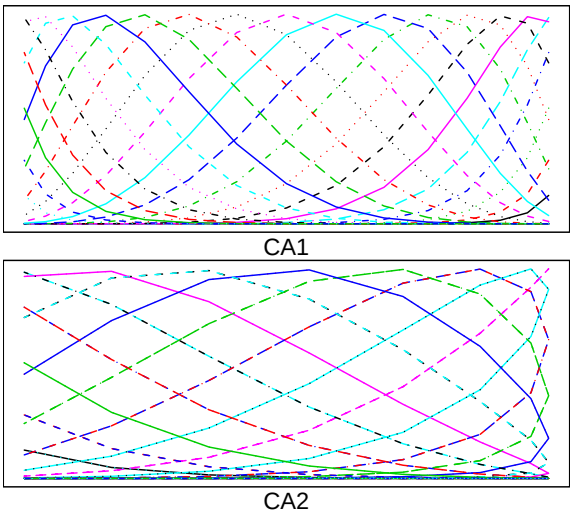
- Single gradients appear as curves in linear ordination methods
- PCA *horseshoe*: curve bends inward and gives wrong ordering of points on axis 1
- CA *arch*: axis 1 retains the correct ordering of sites despite the curve
- Environmental interpretation by vector fitting or surface bound to be biased
- Axes cannot be interpreted as "gradients"



The birth of the curve

- There is a curve in the species space and PCA shows it correctly
- CA deals better wit unimodal responses, but the second optimal scaling axis is folded first axis





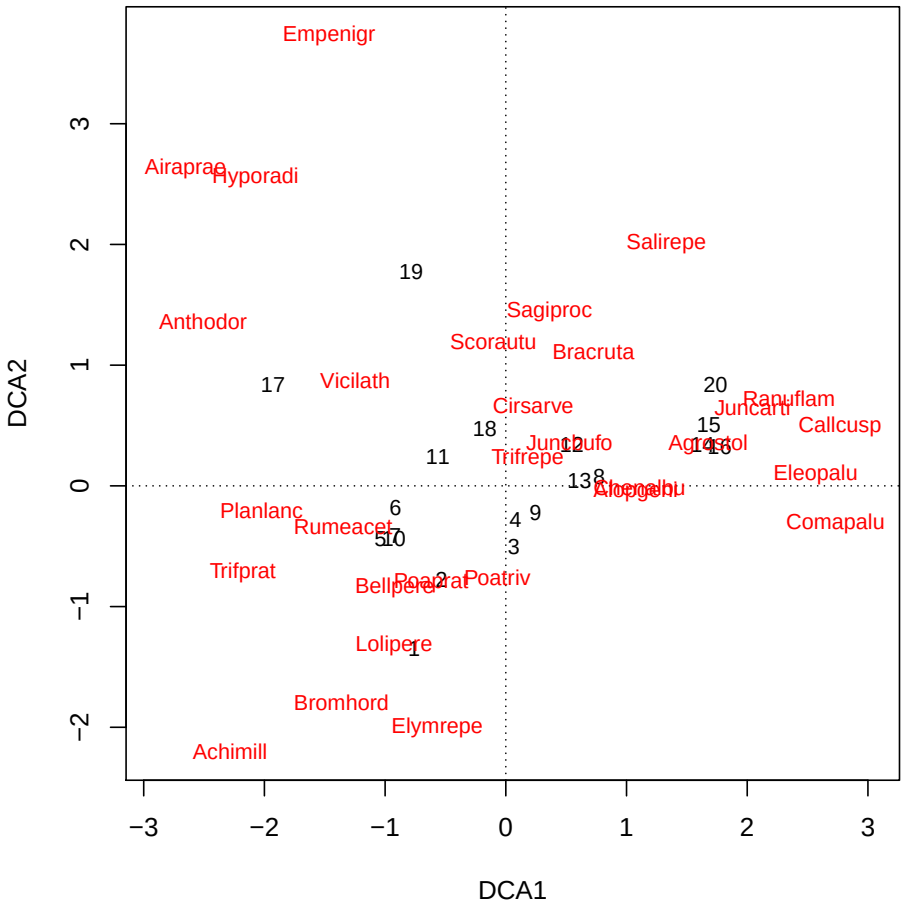
Solutions to the Curvature

- **Detrended Correspondence Analysis** (DCA)
  - CA axis retains the correct ordering: keep that, but instead of orthogonal axes, use detrended axes
  - Programme DECORANA additionally rescales axes to *sd* units approximating *t* parameter of the Gaussian model
  - Distorts space, introduces new artefacts and probably should be avoided
- **Nonmetric Multidimensional Scaling** (NMDS) should be able to cope with moderately long gradients
- Constrained ordination may linearize the responses

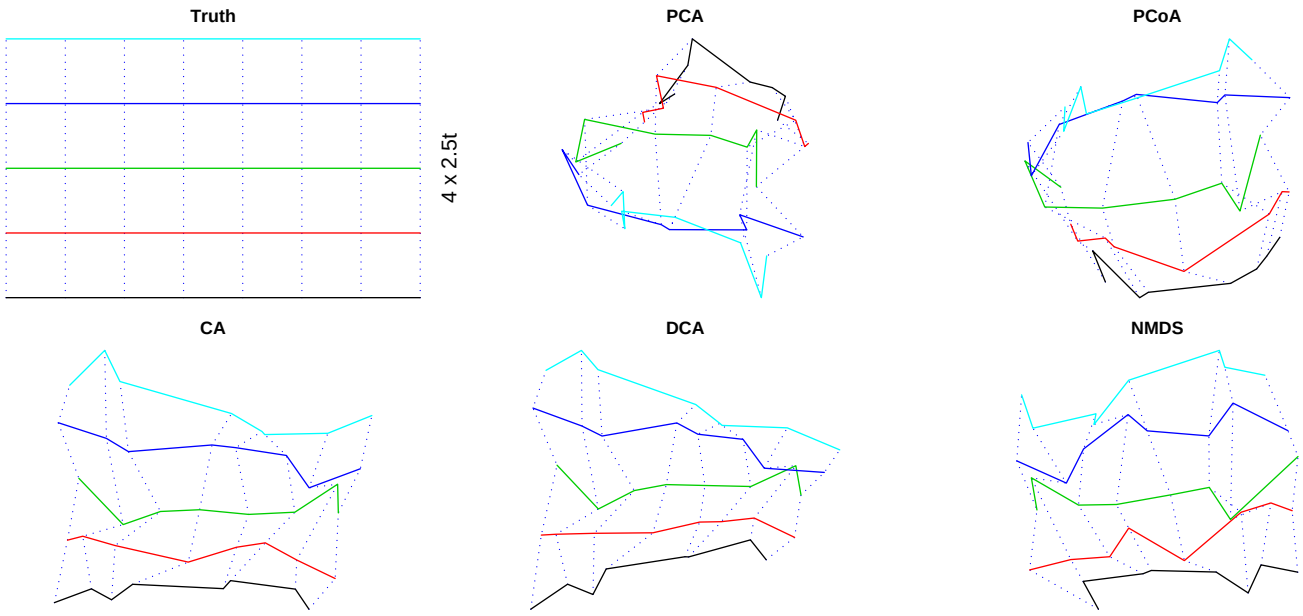
Running Detrended Correspondence Analysis

```
> (ord <- decorana(dune))

Call:
decorana(veg = dune)

Detrended correspondence analysis with 26 segments.
Rescaling of axes with 4 iterations.

                 DCA1  DCA2   DCA3   DCA4
Eigenvalues     0.512 0.304 0.1213 0.1427
Decorana values 0.536 0.287 0.0814 0.0481
Axis lengths    3.700 3.117 1.3005 1.4789
```

Default plot

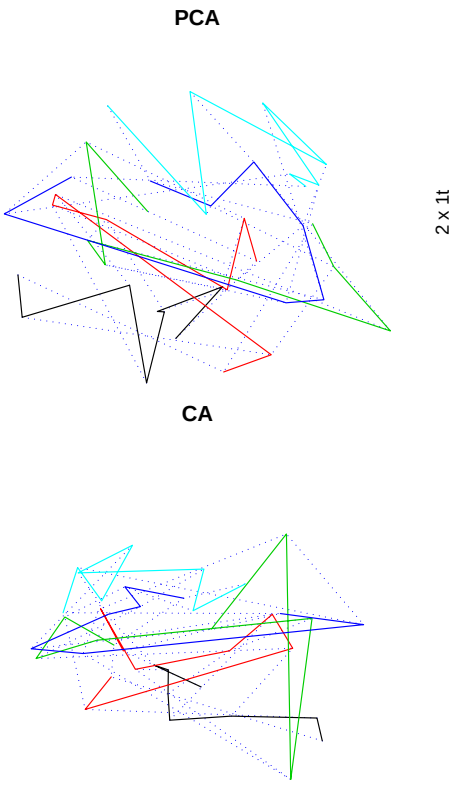


Community Pattern Simulation



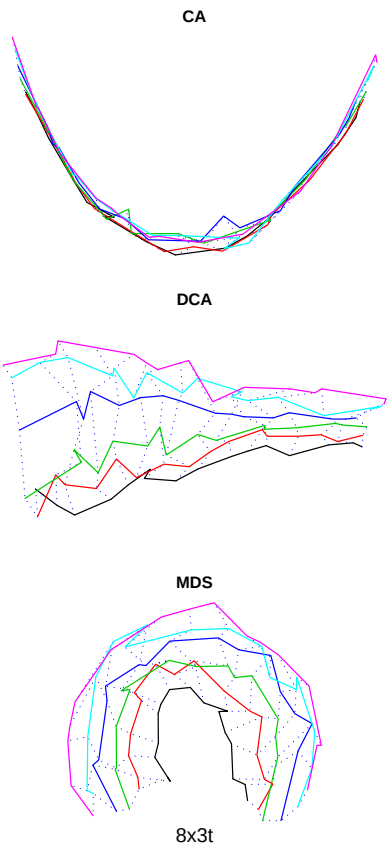
Short Gradients: Is There a Niche for PCA?

- Folklore: PCA with short gradients ($\leq 2t$).
- Not based on research, but simulation finds PCA uniformly worse than CA: With short gradients about as good as CA, but usually worse.
- There should be no species optimum within gradient: Shortness alone not sufficient.
- PCA best used for really linear cases (environment) or for reduction of variables into principal components (but see FA).
- Noise dominates over signal in homogeneous data.



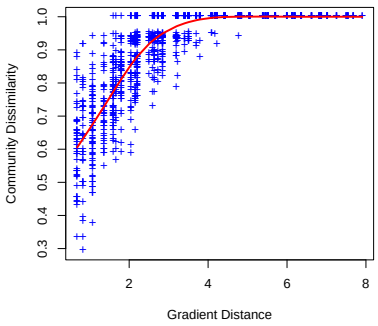
Long Gradients: DCA or NMDS

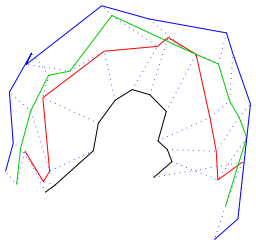
- Curvature with long gradients: Need either DCA or NMDS.
- NMDS is a test winner: More robust than DCA.
- DCA more popular.
- DCA may produce new artefacts, since it twists the space.

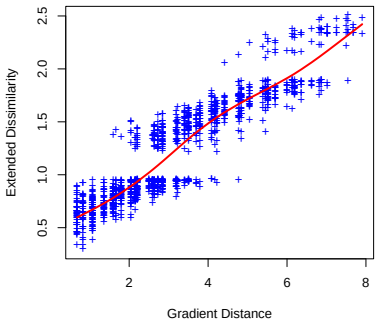


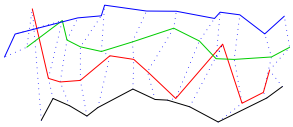
Extended Dissimilarities and Step-across

- How different are sites that have nothing in common?
- Use step-across points to estimate their distance
- Flexible shortest path or their approximations, extended dissimilarities
- Extended dissimilarity: use only one-site steps, do not update dissimilarities below a threshold
- No shared species since rare species were not observed: Swan transformation estimates the probability of finding an unobserved species



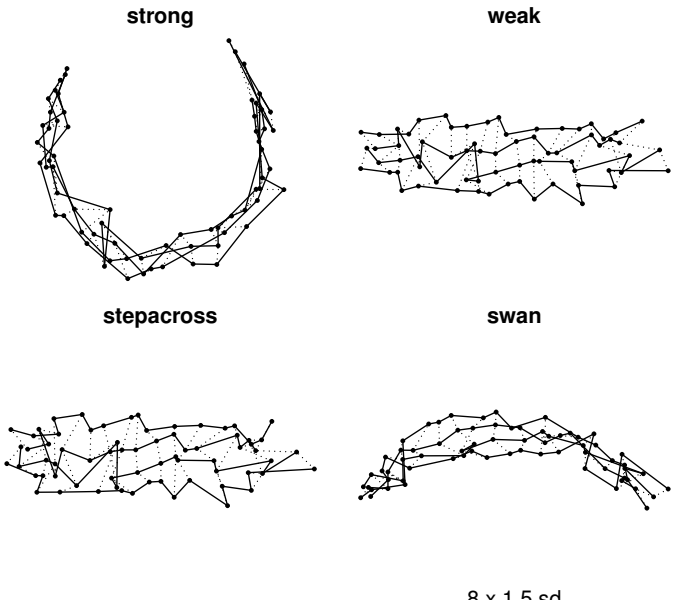






Strong and Weak Ties

- Maximum dissimilarities (no shared species) are tied
- Strong tie treatment tries to keep tied values together and puts maximum dissimilarites to a circle
- Weak tie treatment allows breaking ties and straightens the axes: now the default in vegan, whereas earlier was impossible



8 × 1.5 sd units, Gaussian binary response