

Major ordination methods

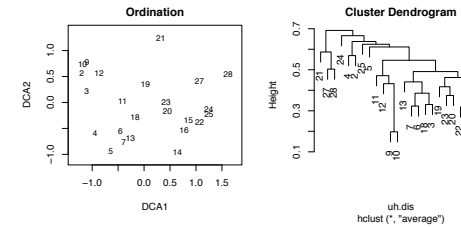
- Principal Components Analysis (PCA)
 - Factor Analysis (FA)
 - Principal Co-ordinates Analysis (PCoA)
- Correspondence Analysis (CA)
 - Detrended Correspondence Analysis (DCA)
- Non-metric Multidimensional Scaling (NMDS)

Principal Components Analysis (PCA)

- Rotates the data in species space so that major part of the variation is shown on the first axes.
- Linear representation of the data: usually inadequate for community analysis but good for reducing environmental measurements.

Why ordination?

- Nobody should **want** to make ordination, but they are desperate with multivariate data.
- Show as much as possible with only a few axes, and explain those dimensions.

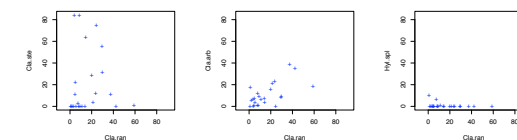


```

1111111122222222
234567901234568901234578
Cal.vul ++5.1..14185352.+5.+...
Emp.nig 652+66573746167667574674
Led.pal .....+.....54...42
Vac.myr .....1...+24.4.76...67
Vac.vit 65456677776675767877776
Vac.uli .4+.6.....4.2.4..431
Dic.sp .....1+.23..31+.87.1
Dic.fus +2113221114778414185.752
Dic.pol ...+.+111...+1.+5.+4..1
Hyl.spl .....4...+66
Ple.sch 3+4+++537666875774878899
Pol.jun +.+2+1111..+11++4+1.416.+
Pti.cil ++.+.1.3.1.+++1.6+4+1++
Cla.arb 267788236587768663665641
Cla.ran 887989568784678575566463
Cla.ste 99837199897+1+58+71++4.
Cla.unc 1131521143385311514363++
Cla.cor +1+111111+21+11+++4122+
Cla.cri ++++1+11123311++114+4+++
Cet.niv .+61+131.....++...+...
Ste.sp .136+3.13.+23+21...4+++.
Cla.def +++11+11.124331232+42+++
    
```

Species space

- Graphical presentation of data matrix: Species are axes and span the space where sites are points.
- Some species show more of the configuration of sites than others.

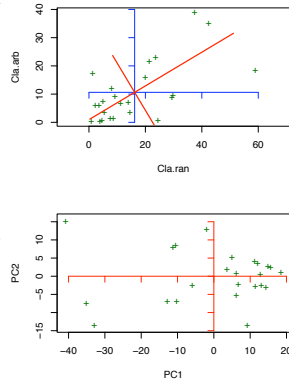


```

1111111122222222
234567901234568901234578
Cal.vul ++5.1..14185352.+5.+...
Emp.nig 652+66573746167667574674
Led.pal .....+.....54...42
Vac.myr .....1...+24.4.76...67
Vac.vit 65456677776675767877776
Vac.uli .4+.6.....4.2.4..431
Dic.sp .....1+.23..31+.87.1
Dic.fus +2113221114778414185.752
Dic.pol ...+.+111...+1.+5.+4..1
Hyl.spl .....4...+66
Ple.sch 3+4+++537666875774878899
Pol.jun +.+2+1111..+11++4+1.416.+
Pti.cil ++.+.1.3.1.+++1.6+4+1++
Cla.arb 267788236587768663665641
Cla.ran 887989568784678575566463
Cla.ste 99837199897+1+58+71++4.
Cla.unc 1131521143385311514363++
Cla.cor +1+111111+21+11+++4122+
Cla.cri ++++1+11123311++114+4+++
Cet.niv .+61+131.....++...+...
Ste.sp .136+3.13.+23+21...4+++.
Cla.def +++11+11.124331232+42+++
    
```

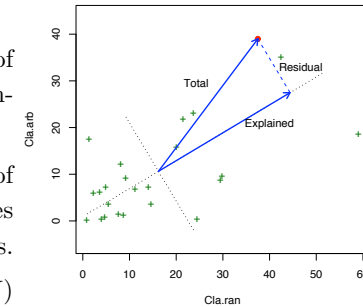
Rotation in species space

1. Move the axis origin to the centroid of the species space (species averages).
 2. Rotate the axes so that the first axis becomes
 - (a) As close to all observations as possible, which means that it
 - (b) Explains as much of the variance as possible.
- First rotated axes show the configuration as well as possible.



Explaining the variation

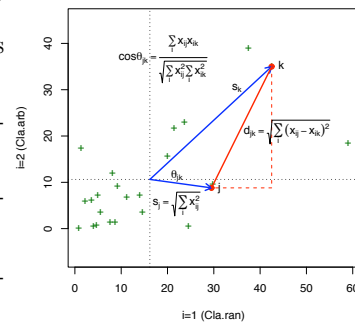
- **Total variation:** Sum of squared distances from the origin.
- **Explained variation:** Sum of squared projections onto principal components = λ .
- **Residual variation:** Sum of squared orthogonal distances from the principal components.
- Only a rotation: All $\min(S, N)$ axes explain everything.



How computer sees the configuration?

The configuration of points in species space can be defined either

1. As distances s_j, s_k and angles θ_{jk} from the origin.
 - Both included in covariances and variances.
 - $\cos \theta$ is Pearson's correlation coefficient.
2. As Euclidean distances d_{jk} between pairs of points.



Singular Value Decomposition (SVD)

$$x_{ij}^{(M)} = \bar{x}_j + \sum_{m=1}^M v_{im} \lambda_m^{1/2} u_{jm}$$

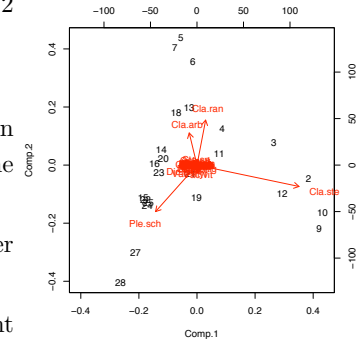
- An M -order least squares approximation of data matrix as a product of orthonormal loading matrices for sites v_{im} and species u_{jm} and corresponding eigenvalue λ_m
- Zero-order approximation: All species at their average abundance at the origin.
- When $M = \min(S, N)$ then approximation exact and $\sum \lambda =$ total variance.
- A numerical way of saying that PCA is only a rotation.

Loadings and scores

- Matrices **V** and **U** are orthonormal in SVD.
 - Average is zero $\sum_i u_{im} = 0$
 - Sum of squares is one $\sum_i u_{im}^2 = 1$
 - Components are uncorrelated $\sum_i u_{im}u_{in} = 0; \forall m \neq n$
- Normal scaling in PCA:
 - Singular values σ are scaled to covariance eigenvalues $\lambda = \sigma^2 / (n - 1)$.
 - Species loadings v_{jm} kept orthonormal: Rotation matrix.
 - Site scores scaled by singular values $\sigma_m^2 u_{im}$: Reflect the importance of axes and show the real configuration.
 - Sum of squares of scaled site scores = $\lambda_m(n - 1)$.

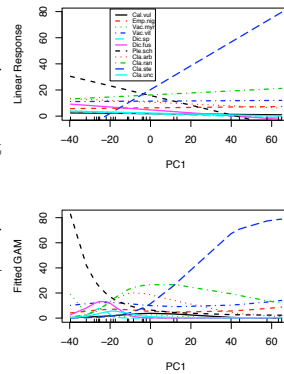
Biplot: Graphical SVD

- Species loadings and site scores in the same plot: Graphical order 2 approximation of the data.
- Origin: species averages.
- Species increases to the direction of the arrow, and decrease to the opposite direction.
- The longer the arrow, the stronger the increase.
- Approximation: Project site point onto species vector.



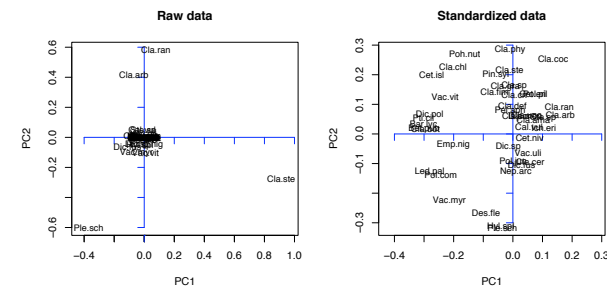
Linear response model

- PCA assumes that species have a linear regression with PCs.
- Loading is the regression coefficient between PC and the species.
- PCA may “fail”: Real species responses rather unimodal despite PCA assumptions.



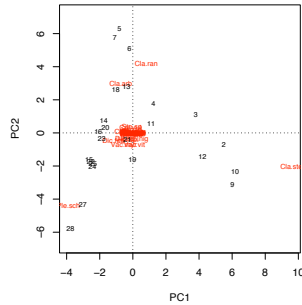
Standardized PCA

Standardizing species to unit variance (using correlation coefficient) makes all species equally important, instead of concentrating on the abundant species with largest variances.



PCA plot

- Axes must have identical scales.
- Origin is special and should be marked: Points near the origin are either average or poorly explained.
- Direction from the origin: The fastest change in species abundances.
- Distance from the origin: The magnitude of change.
- Angles between vectors: Similarity of the response.

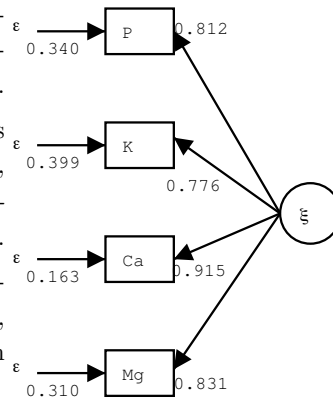


Factor Analysis (FA)

- Factor Analysis instead of PCA in many programs.
- Factor Analysis is a **statistical** method, whereas PCA is only a rotation.
 - World has two components: Common and unique.
 - FA tries to explain only the common component.
 - PCA discards later axes in hope of discarding unique variation, but that is mixed with all axes – kept and rejected.
- Latent variables ξ generate the observed variables x which have unique errors ε .
- Exploratory FA: Rotation to a simple structure.

Confirmatory Factor Analysis

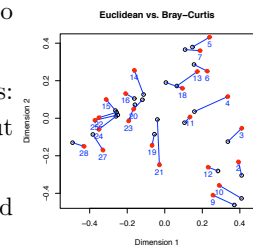
- The real variables ξ are non-observable (latent), but they influence several observable variables x .
- Select a set of observable variables as indicators of the latent variable, and use them to build a *measurement model* for the latent variable.
- The latent variables are truer explanatory variables than single, noisy observed variables and can be used in further modelling.



Principal Co-ordinates Analysis (PCoA)

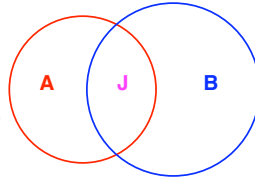
A.K.A. metric or classic multidimensional scaling

- With Euclidean distances identical to PCA: Eigen analysis.
- Allows use of other dissimilarity measures: These may be mathematically naïve but ecologically elegant.
- Classical in vegetation ecology – still used in U.K.



Dissimilarities for community data

- Presence/absence indices based on the number common species J on two sites compared to species richness A, B of sites.



- Similarity index s can be transformed to a dissimilarity index $d = 1 - s$.

- Quantitative generalizations:

1. **Manhattan style:** use common part of abundance and sums.

$$\text{Jaccard: } \frac{J}{A+B-J}$$

$$\text{Sørensen: } \frac{2J}{A+B}$$

$$\text{Ochiai: } \frac{J}{\sqrt{AB}}$$

$$\text{Kulczyński: } \frac{1}{2} \left(\frac{J}{A} + \frac{J}{B} \right)$$

2. **Euclidean style:** use cross products and sums of squares.

The number of indices is a legio

Manhattan style

$$J = \min_i(x_{ij}, x_{ik})$$

$$A, B = \sum_i x_{ij}$$

Euclidean style

$$J = \sum_i x_{ij}x_{ik}$$

$$A, B = \sum_i x_{ij}^2$$

	0/1	Manhattan style	Euclidean style
$\frac{J}{A+B-J}$	Jaccard	Jaccard	Similarity Ratio
$\frac{2J}{A+B}$	Sørensen	Steinhaus (Bray–Curtis)	Ružička
$\frac{J}{\sqrt{AB}}$	Ochiai	?	Cosine

Metric properties of indices

An index d is a **metric** if:

1. if $a = b$, then $d_{ab} = 0$ minimum 0
2. if $a \neq b$, then $d_{ab} > 0$ always positive
3. $d_{ab} = d_{ba}$ symmetric
4. $d_{ac} + d_{cb} \geq d_{ab}$ triangle inequality

Semimetric: Does not obey triangle inequality

Correspondence Analysis (CA)

Invented independently numerous times:

1. **Correspondence Analysis:** Weighted Principal Components with Chi-squared metric.
2. **Optimal Scaling:** Find site and species scores so that (i) all species occurring in one site are as similar as possible, but (ii) species of different sites are as different as possible, and (iii) sites are dispersed as widely as possible relative to species scores.
3. **Reciprocal Averaging:** Species scores are weighted averages of site scores, and simultaneously, site scores are weighted averages of species scores.

Chi-squared metric

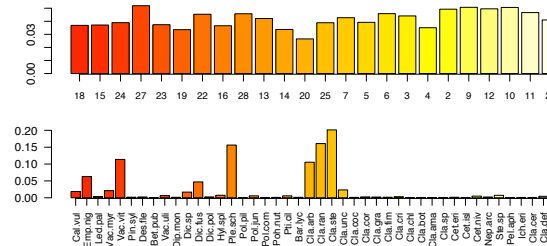
- *Metric* is a ‘yardstick’ to measure dissimilarities among points.
- PCA uses constant (‘Euclidean’) metric, but CA uses expected abundances as a metric.
- Expected abundances from marginal totals: Exactly like in χ^2 analysis of contingency tables.
- *Species profile* is the average proportion of species in the data, and *site profile* is the average proportion of sites.

Euclidean: $d_{ij} = f_{ij} - e_{ij}$

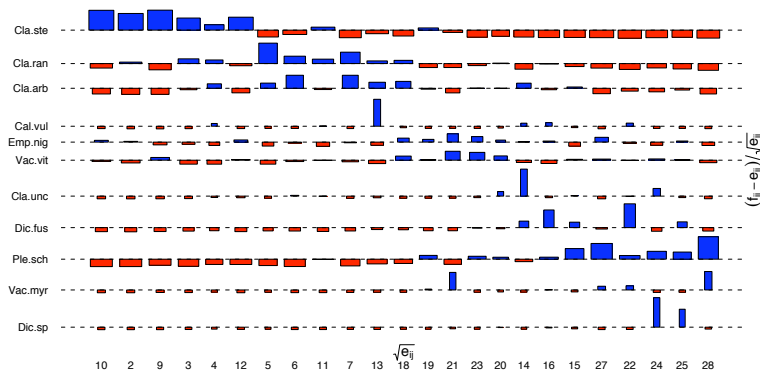
Chi-squared: $\chi_{ij} = (f_{ij} - e_{ij})/\sqrt{e_{ij}}$

Species and site profiles

- All sites should have all species in the same proportions.
- Chi-squared distance is the difference between expected profile and real abundance distributions – both for the species and the sites.

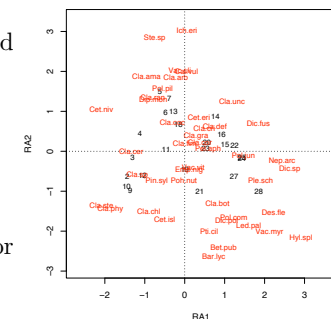


Chi-squared transformation...



... Weighted principal components rotation

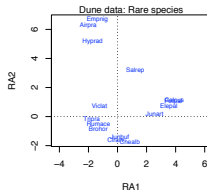
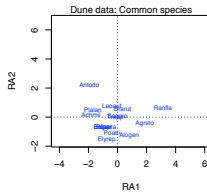
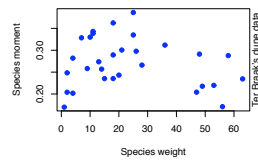
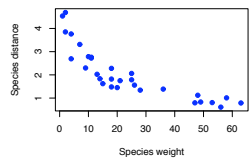
- Weighted rotation in Chi-squared space:
 - Chi-squared transformation.
 - Marginal totals as weights.
- SVD in Chi-squared metric.
- Inbuilt double standardization for sites and species simultaneously.



Rare species

Rare species cannot have an average profile and are extreme in CA – but have small weights. Solutions to the “problem”:

1. Remove rare species.
2. Downweight rare species.
3. Do not show rare species: Popular.

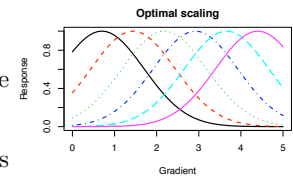
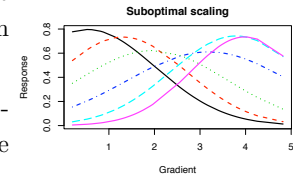


Ordered vegetation table

Raw	PCA	CA
1111111122222222	2212221212 1 12 11 1 1	1 1 1 1112221122222
234567901234568901234578	875245634078531691432290	029342561738913046572458
Cal.vul ++5.1..14185352..+5.+..	Ple.sch 998887767+5+64+774+6353	Cet.niv 1.3+6.1+.1.+++.
Emp.nig 652+66573746167667574674	Dic.fus 2578.7857424141311121+21	Cla.ste 99998937817587+++141+..
Led.pal	Cla.arb 147656667688783866765223	Cla.ran 685877989898556747665643
Vac.myr	Vac.myr 7626..4+.17.4. . . .1.	Ste.sp 1..13.6+33+21.4.2+3.
Vac.vit 6545667776675767877776	Dic.sp 1.2+873..3.1.	Cla.arb 322675786888636676746561
Vac.uli .4+.6.4.2.4. . . .31	Cla.unc ++5463338521131514313111	Vac.uli . . .4.6.4. . . .2. . . .34.+1
Dic.sp	Hyl.spl 66.4.	Cal.vul 1+.51.14.82. . . .553.5+.
Dic.fus +211322114778414185.752	Cal.vul .35+.5.5+.2.8.1.45+1+.1	Emp.nig 765527+63647677666175464
Dic.pol . . .+.411. . . .1+.45.4+.1.	Pol.jun +.1.16+41+1+2+1+41+. . . .11	Vac.vit 76754756766768876577776
Hyl.spl	Vac.uli 13.4+. . . .264.4.	Cla.cor 1+11+1111+11+4+2+12+12+
Ple.sch 3+4++53766687577487899	Led.pal 24.4.5.	Cla.cri 1+1+1+11+21+14+313+1+++
Pol.jun +.2+1111.1+11+4+1.416+. . . .	Ste.sp +.3.42.326+. . . .1331. . . .1	Pti.cil 1+1+.3.1+164.
Pti.cil ++. . . .1.3.1.6+4+1+	Cla.def ++3+2+3443+112212. . . .1+1+1	Cla.def 1+1+111.212243433+2+
Cla.arb 267788236587768636666463	Pti.cil ++++1+4.16+13.+1.	Dic.pol 1.1.1+.5+.1+.4.1
Cla.ran 887989568784678575566463	Cla.cri ++31+143+1+21+1+1+1+1	Cla.unc 1111331542311135835+463+
Cla.ste 99837199897+1+58+71+4.	Dic.pol 1+.4.1+.+5.1.11	Pol.jun 1+.1.2+1+1+41+1+. . . .16+
Cla.unc 1131521143385311514363+	Cla.cor +21+12+42+11+1+1+1+1+1	Led.pal
Cla.cor +1+111111+21+11+4122+	Cet.niv	Dic.fus 1+2211131244115478758.72
Cla.cri ++++1+1112331+1+114+++	Vac.vit 67777586767568667457677	Ple.sch 335+46+7+65747767898889
Cet.niv ++1+131.	Emp.nig 471546676667+47663257657	Vac.myr .1.1.47.4266.
Ste.sp .136+3.13.+23+21.4+.	Cla.ran 366564764798985858787856	Dic.sp1.+1.3.32.+871
Cla.def +++1+11.124331232+42+++	Cla.ste .411+.1537778889999	Hyl.spl

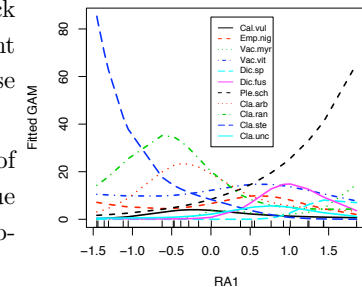
When scaling is optimal?

- The species optima should be widespread: Measured by between species variance SS_B .
- The species responses should be narrow: Measured by within site variance of species optima SS_W .
- Total variance is $SS_T = SS_B + SS_W$.
- If most variance is between sites, the scaling is optimal.
- The criterion variable $\lambda = SS_B/SS_T$ is maximized in CA.



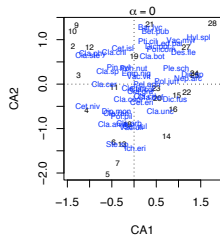
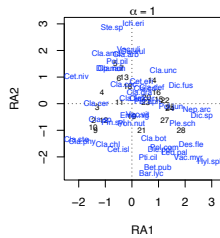
Unimodal response

- Optimal scaling tries to pack species occurrences into tight parcels: Unimodal response model.
- Eigenvalue λ tells the success of packing – but too high a value (~ 1) indicates disjunct subsets of sites.



Reciprocal weighted averages

- Repeated application of weighted averaging leads into CA.
- Weighted averages shrink by factor λ and need restoring.
- Species scores can be weighted averages of site scores ($\alpha = 1$), or vice versa ($\alpha = 0$), or restored symmetrically ($\alpha = \frac{1}{2}$).

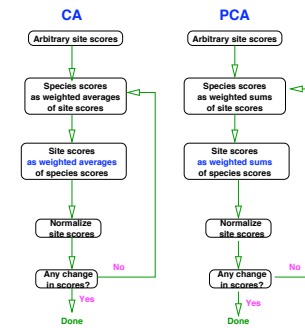


$$u_i = \lambda^{\alpha-1} \frac{j x_{ij} v_j}{j x_{ij}}$$

$$v_j = \lambda^{-\alpha} \frac{i x_{ij} u_i}{i x_{ij}}$$

Power algorithm

Arbitrary vector multiplied with high power of a matrix converges to the eigenvector of the matrix.

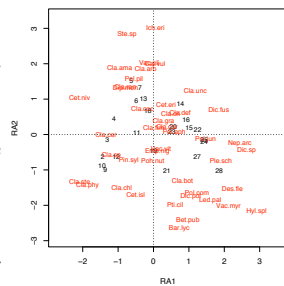


- Repeated application of weighted averaging gives CA solution.
- Repeated application of weighted sums gives PCA solution.
- Both methods simple, but numerically poor.
- In fact, CA and PCA both are best made using [weighted] SVD of [Chi-square transformed] matrix.

CA: Joint plots

Biplot with weighted Chi-squared metric: Species and sites in the same plot (but scaling α counts).

- Distance from the origin: Chi-squared difference from the profile.
- Points at the origin either average or poorly explained.
- Distant species often rare, close species common.
- Unimodal interpretation: Species optima and gradient values – at least for well explained species.



Eigenvalue in CA

1. The factor of shrinking in weighted averaging: $\lambda = 0 \dots 1$.
 2. The proportion between site variation from total variation: $\lambda = 0 \dots 1$.
- Sum of all eigenvalues = mean squared contingency coefficient: The statistic decomposed into linear components in CA.
 - No ecological meaning: “Proportion of ‘variance’ explained” neither meaningful.
 - Variance is never evaluated in CA.
 - Eigenvalues close to 1 suspicious: Probably a disjunct subset of points.

Detrended Correspondence Analysis (DCA)

Mission to correct three 'artefacts' in CA:

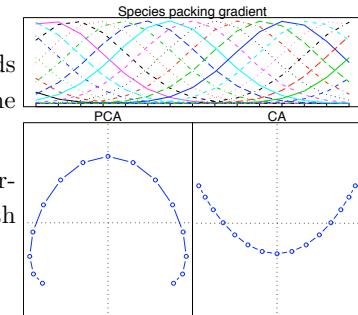
1. **Detrending** to remove 'spurious' curvature in ordination.
2. **Rescaling** to correct shrinking at the ends of ordination axes.
3. **Downweighting** to reduce the influence of rare species.

Program DECORANA the only implementation – difficult to separate a program from the method.

Detrending CA: The argument

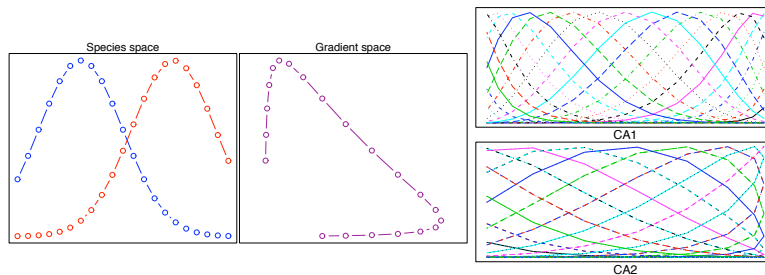
PCA and CA both produce a curve from a single, ideal gradient, but the shapes have one important difference:

- Horseshoe in PCA curved inwards at ends: Wrong order along the first axis.
- Arc in CA preserves the correct ordering along the first axis: Worth of detrending.

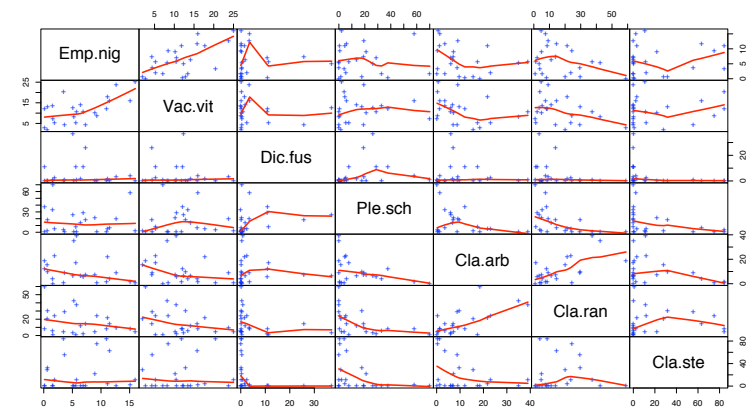


The birth of the curve

- There is a curve in the species space, and PCA shows it correctly.
- CA may be able to deal with unimodal responses, but if there is one dominant gradient, the second optimal scaling is folded first axis.

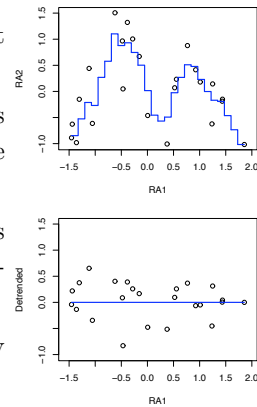


But is there a curve in species space?



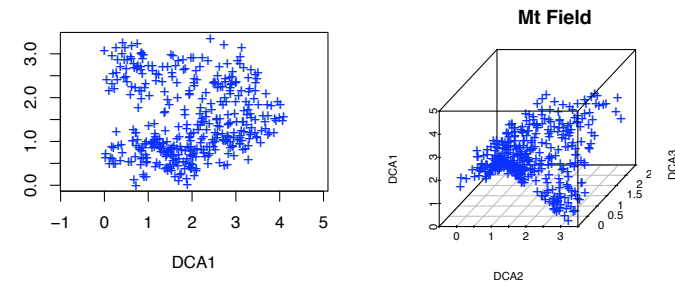
Detrending by segments

- Divide the axis into segments (default 26).
- Average using moving, weighted windows in segments and take the residuals as the new 2nd axis.
- Ideally the direction of 2nd axis changes to an important, non-linearly independent gradient.
- May remove information as well, not only the arc like desired.



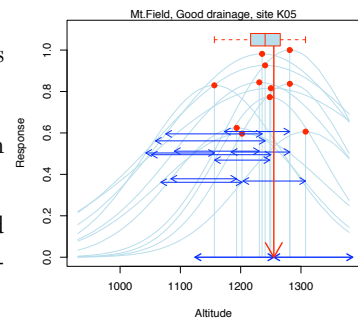
Detrending artefacts

- DCA configuration usually a triangle or a diamond: an artefact.
- DCA detrends twisting the space: ‘Lollypop’ or ‘lasagna’ effect.



Hill indices of beta diversity

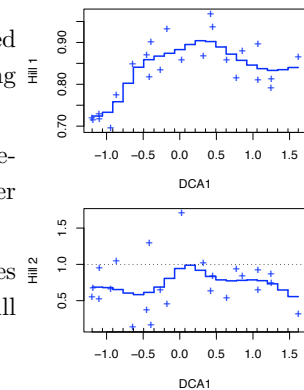
1. Average width of species responses.
 2. Variance of optima of species occurring in one site.
- Used with scaling of ordination axes.
 - The first index discussed and described, but the second applied.
 - Equal only to degenerated species packing gradients.



Gaussian responses fitted to species occurring in one site.

Rescaling and downweighting

- Rescale gradients so that the weighted variance of species scores = 1 along the axes.
- Often said that average width of responses is 1 ‘sd’ unit, but that is never done.
- Downweighting: Reduce abundances of rare species so that they have still less weight on ordination.

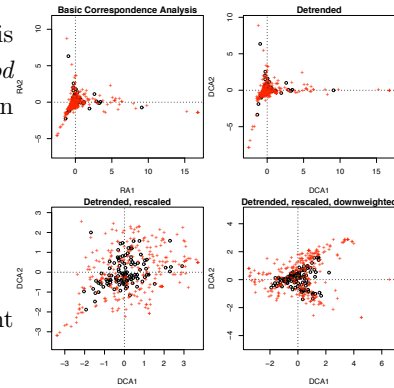


Method or Programme?

It seems that DECORANA rather is a piece of software than a *method* of ordination, since it is a collection of three different tricks:

1. Detrending
2. Rescaling
3. Downweighting

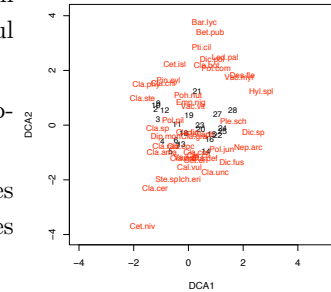
Rescaling often is more important than detrending...



DCA plot

DCA plots often taken more seriously than other ordinations:

- Axes taken as gradients, and even scaled in ecologically meaningful 'sd' units.
- Species scores taken as species optima.
- Species scaled so that site scores are their direct weighted averages ($\alpha = 1$).

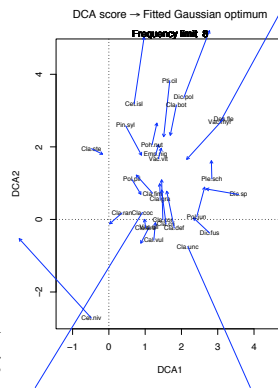


Is DCA based on Gaussian response model?

CAJO TER BRAAK:

“Four conditions (equal tolerances, equal or independent maxima, and equally-spaced or uniformly distributed optima and sample points) are needed to show that (detrended) correspondence analysis provides an *approximate* solution to the unimodal models.” *Biometrics* **41**, p. 870 (1985).

= (D)CA can *approximate* Gaussian model – if we have infinite species packing gradients.



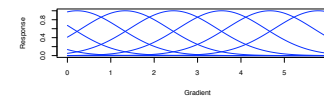
Weighted averages are good estimates ...

... of species optima if:

1. Sites x are evenly distributed about optimum u
2. Sites are close to each other

... of gradient values if:

1. Species optima u evenly distributed about site x
2. All species have equal response widths t
3. All species have equal maximum abundance h
4. Optima u are close to each other



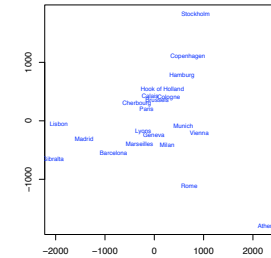
Conditions true only for infinite species packing gradients.

Non-metric Multidimensional Scaling (NMDS)

- Rank-order relation with community dissimilarity and ordination distance: 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.

MDS is a map

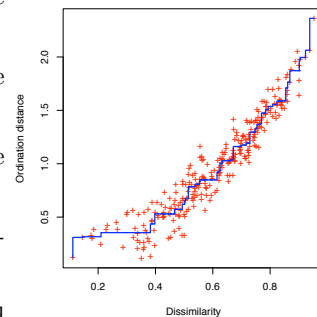
- MDS tries to draw a map using distance data.
- MDS tries to find an underlying configuration from dissimilarities.
- Metric MDS (PCoA) assumes linear relation, but NMDS finds any relation.
- Only the configuration counts:
 - No origin, but only the constellations.
 - No axes or natural directions, but only a framework for points.



Map of Europe from road distances.

Monotone regression

- Measured community dissimilarities and ordination distances have similar rank ordering.
- No specified shape, but can cope with different response shapes.
- Sum of squared residuals from the regression: *Stress*.
- The model behind: Finds gradients if dissimilarities meaningful.
- Iterative solution: No guaranteed convergence.



Recommended procedure

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.

Good dissimilarity measures for gradients

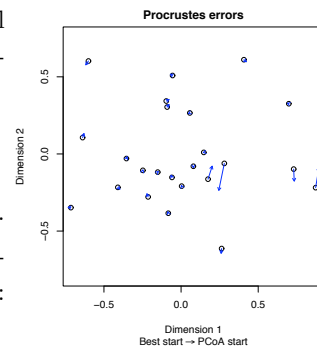
- The Model: Find a a dissimilarity measure that describes correctly gradient separation.
- Bray–Curtis (Steinhaus), Jaccard and Kulczyński.
- ‘Wisconsin double standardization’ often helpful: First standardize each species to equal maximum, then sites to equal totals.
- Euclidean distance and Chi-squared dissimilarity are poor.

Starting MDS

- Most software packages start with metric MDS (PCoA), but this is dangerous:
 1. Metric MDS may be close to a local minimum which traps the iteration.
 2. Hides the uncertainty of convergence from the user (and that is unmoral).
- Start with random configurations, and compare minimum stress solutions.

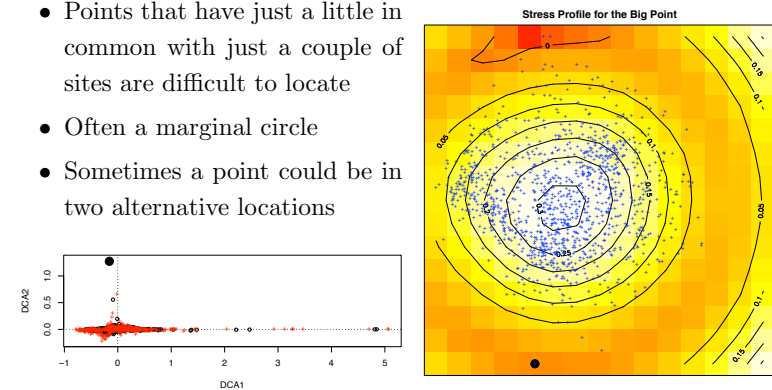
Comparing configurations: Procrustes rotation

- 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.



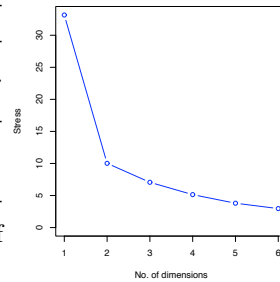
Outliers in the outskirts

- Points that have just a little in common with just a couple of sites are difficult to locate
- Often a marginal circle
- Sometimes a point could be in two alternative locations



Number of dimensions

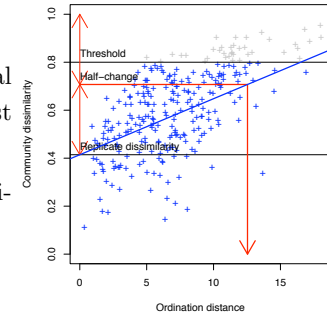
- In eigenvector methods axes are orthogonal and previous axes remain unchanged when new axes are evaluated.
- MDS Solutions for each number of dimensions separate.
- Adequate number of dimensions difficult to know: After sudden drop of stress is a good idea.



Scaling of axes

MDS axes have no unique scaling or direction, but all rotations and scalings are equally good solutions.

- Customary to rotate to principal components: First dimension most important.
- Half-change scaling give ecologically meaningful units.



MDS plot

- All that counts is the configuration: Axes are nothing.
- No origin, but just a map.
- Species can be added as their weighted averages.

