

## Ordinations: Correspondence analysis



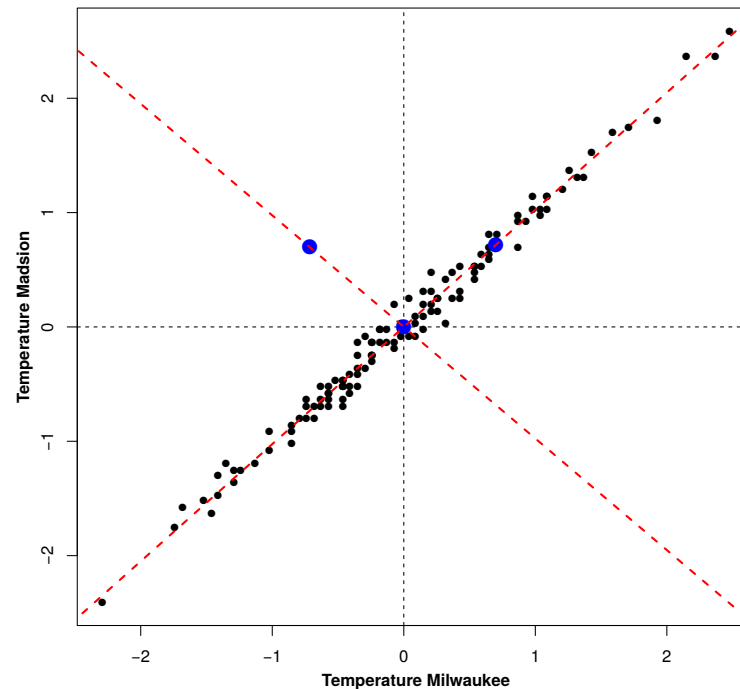
# PCA

- ▶ Find most relevant patterns in a dataset
- ▶ Summarize a dataset in a simple two dimensional scatterplot
  - ▶ relation among variables
  - ▶ distance among samples

# PCA: Geometric representation

New **coordinate system** so that:

Axes point into the direction of maximum elongation of dataset  
(data cloud)



Direction of new coordinates in old coordinate space: **Eigenvector**  
(PC loadings, species scores)

# PCA: Mathematics

...	Variable 1	Variable 2	...	Variable i
time 1	$X_{11}$	$X_{21}$	...	$X_{i1}$
time 2	$X_{12}$	$X_{22}$	...	$X_{i2}$
...	...	...	...	...
time j	$X_{1j}$	$X_{2j}$	...	$X_{ij}$

Correlation or covariance matrix

...	Variable 1	Variable 2	...	Variable i
Variable 1	1	$r_{12}$	...	$r_{1i}$
Variable 2	$r_{12}$	1	...	$r_{2i}$
...	...	...	...	...
Variable i	$r_{1j}$	$r_{2i}$	...	1

# PCA: Mathematics

Eigenvalue decomposition of correlation or covariance matrix

...	Variable 1	Variable 2	...	Variable i
Variable 1	1	$r_{12}$	...	$r_{1i}$
Variable 2	$r_{12}$	1	...	$r_{2i}$
...	...	...	...	...
Variable i	$r_{1j}$	$r_{2i}$	...	1

**Eigenvectors:** new coordinate system

...	ev1	ev2	...	ev i
species 1	$ev1_1$	$ev2_1$	...	$evl_1$
species 2	$ev1_2$	$ev2_2$	...	$evl_2$
...	...	...	...	...
species i	$ev1_i$	$ev2_i$	...	$evl_i$

# PCA: Geometric representation

Position in new coordinate system: **Principal component** (site scores)

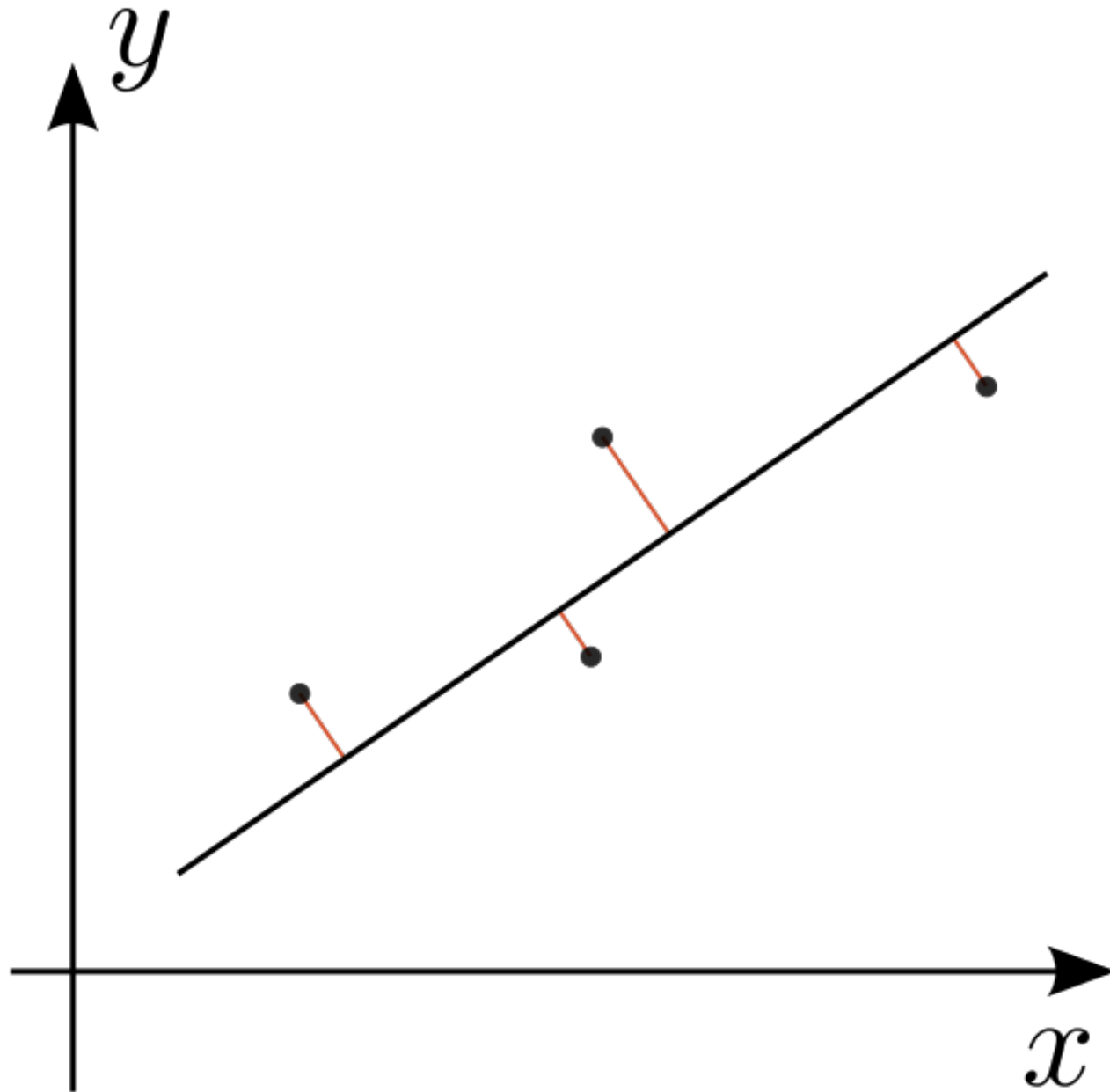
...	Var1	Var2	...	Var i	...	PC1	PC2	...	PC i
time 1	$z1_1$	$z2_1$	...	$z1_i$	...	$pc1_1$	$pc2_1$	...	$pci_1$
time 2	$z1_2$	$z2_2$	...	$z1_2$	...	$pc1_2$	$pc2_2$	...	$pci_2$
...	...	...	...	...	...	...	...	...	...
time j	$z1_j$	$z2_j$	...	$z1_j$	...	$pc1_j$	$pc2_j$	...	$pci_j$

**Principal components:** linear combination of initial variables and eigenvectors

$$pc1_1 = ev1_1z1_1 + ev1_2z2_1 + ...ev1_iz1_i$$

$$pc1_2 = ev1_2z1_2 + ev1_2z2_2 + ...ev1_iz1_2$$

## PCA as extension of linear regression





# Gradient analysis

Some important physical gradient (e.g. altitude) determines occurrences of biological species

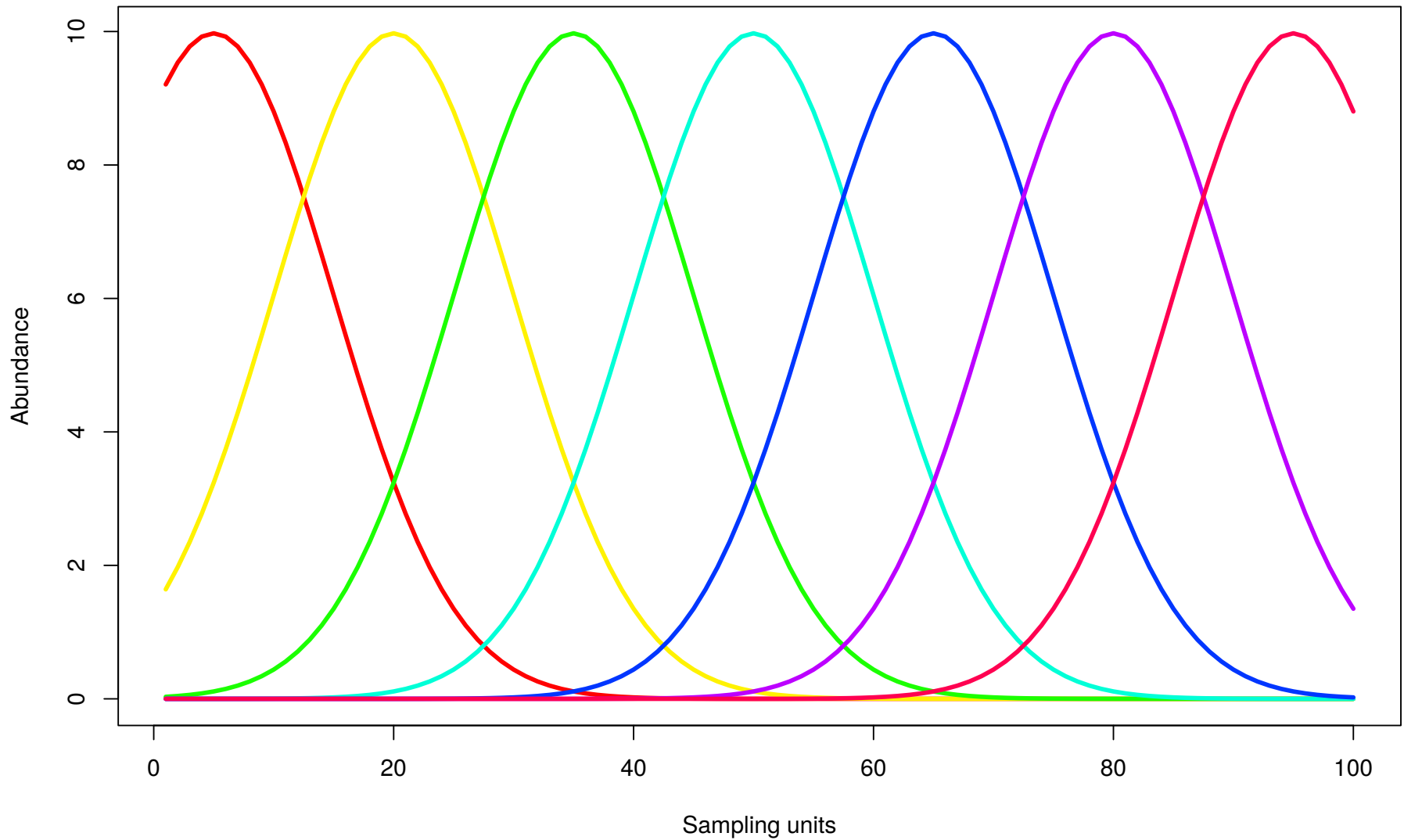
PCA:

- ▶ assume linear response of species to gradient (i.e. value of a species is low at one end of the gradient and high at the other end of the gradient)

Do biological species respond linearly to an environmental or latent gradient?

# Plant species: Unimodal response

Niche theory: plants have a unimodal response



## Gradient analysis:

There is some latent (unobserved) variable underlying that gradient

...	species 1	species 2	...	species i	$\Sigma$
site 1	$p_{11}$	$p_{12}$	...	$p_{1i}$	$p_{1+}$
site 2	$p_{21}$	$p_{22}$	...	$p_{2i}$	$p_{2+}$
...	...	...	...	...	...
site j	$p_{j1}$	$p_{j2}$	...	$p_{ji}$	$p_{j+}$
$\Sigma$	$p_{+1}$	$p_{+2}$	...	$p_{+i}$	$p_{++}$

*The method of gradient analysis is to take some well-marked gradient and to assign scores to the species according to their [...] preferences. Sites are then ordinated by taking averages of the scores of the species which occur in them. Hill, 1973*

## Gradient analysis: Reciprocal averaging

...	species 1	species 2	...	species i	$\Sigma$
site 1	$p_{11}$	$p_{12}$	...	$p_{1i}$	$p_{1+}$
site 2	$p_{21}$	$p_{22}$	...	$p_{2i}$	$p_{2+}$
...	...	...	...	...	...
site j	$p_{j1}$	$p_{j2}$	...	$p_{ji}$	$p_{j+}$
$\Sigma$	$p_{+1}$	$p_{+2}$	...	$p_{+i}$	$p_{++}$

**Site scores:**  $x$

**Species scores:**  $y$

Start with random site scores (assign a random env. variable)

$$scores_{species} = \frac{p_{11}}{p_{+1}} score_{site,1} + \frac{p_{21}}{p_{+1}} score_{site,2} + \dots + \frac{p_{j1}}{p_{+1}} score_{site,j}$$

$$y_1 = \frac{p_{11}}{p_{+1}} x_1 + \frac{p_{21}}{p_{+1}} x_2 + \dots + \frac{p_{j1}}{p_{+1}} x_j$$

## Simpler notation

$$y_1 = w_{11}x_1 + w_{21}x_2 + \dots + w_{j1}x_j$$

...	species 1	species 2	species i
site_1	$w_{11} = \frac{p_{11}}{p_{+1}}$	$w_{12} = \frac{p_{12}}{p_{+2}}$	$w_{1i} = \frac{p_{1i}}{p_{+i}}$
site_2	$w_{21} = \frac{p_{21}}{p_{+1}}$	$w_{22} = \frac{p_{22}}{p_{+2}}$	$w_{2i} = \frac{p_{2i}}{p_{+i}}$
site_j	$w_{j1} = \frac{p_{j1}}{p_{+1}}$	$w_{j2} = \frac{p_{j2}}{p_{+2}}$	$w_{ji} = \frac{p_{ji}}{p_{+i}}$

$$scores_{species} = y_i = \sum_{j=1}^{j=m} w_{ji}x_j = \text{weighted} - \text{average}$$

$$\bar{z} = \sum_{j=1}^{j=m} \frac{1}{m} a = \frac{1}{m} \sum_{j=1}^{j=m} a = \text{normal} - \text{average}$$

## New site scores

...	species 1	species 2	...	species i	$\Sigma$
site 1	$p_{11}$	$p_{12}$	...	$p_{1i}$	$p_{1+}$
site 2	$p_{21}$	$p_{22}$	...	$p_{2i}$	$p_{2+}$
...	...	...	...	...	...
site j	$p_{j1}$	$p_{j2}$	...	$p_{ji}$	$p_{j+}$
$\Sigma$	$p_{+1}$	$p_{+2}$	...	$p_{+i}$	$p_{++}$

$$scores_{site} = \frac{p_{11}}{p_{1+}} scores_{spec,1} + \frac{p_{12}}{p_{1+}} scores_{spec,2} + \dots + \frac{p_{1i}}{p_{1+}} scores_{spec,i}$$

$$x_1 = \frac{p_{11}}{p_{1+}} y_1 + \frac{p_{12}}{p_{1+}} y_2 + \dots + \frac{p_{1i}}{p_{1+}} y_i$$

## Simpler notation

$$x_1 = v_{11}y_1 + v_{21}y_2 + \dots + v_{j1}y_j$$

...	species 1	species 2	species i
site 1	$v_{11} = \frac{p_{11}}{p_{1+}}$	$v_{12} = \frac{p_{12}}{p_{1+}}$	$v_{1i} = \frac{p_{1i}}{p_{1+}}$
site 2	$v_{21} = \frac{p_{21}}{p_{2+}}$	$v_{22} = \frac{p_{22}}{p_{2+}}$	$v_{2i} = \frac{p_{2i}}{p_{2+}}$
site j	$v_{j1} = \frac{p_{j1}}{p_{j+}}$	$v_{j2} = \frac{p_{j2}}{p_{j+}}$	$v_{ji} = \frac{p_{ji}}{p_{j+}}$

$$scores_{site,j} = x_j = \sum_{i=1}^{i=n} v_{ij}y_i$$

$$\bar{z} = \sum_{j=1}^{j=m} \frac{1}{m} a = \frac{1}{m} \sum_{j=1}^{j=m} a = \text{normal} - \text{average}$$

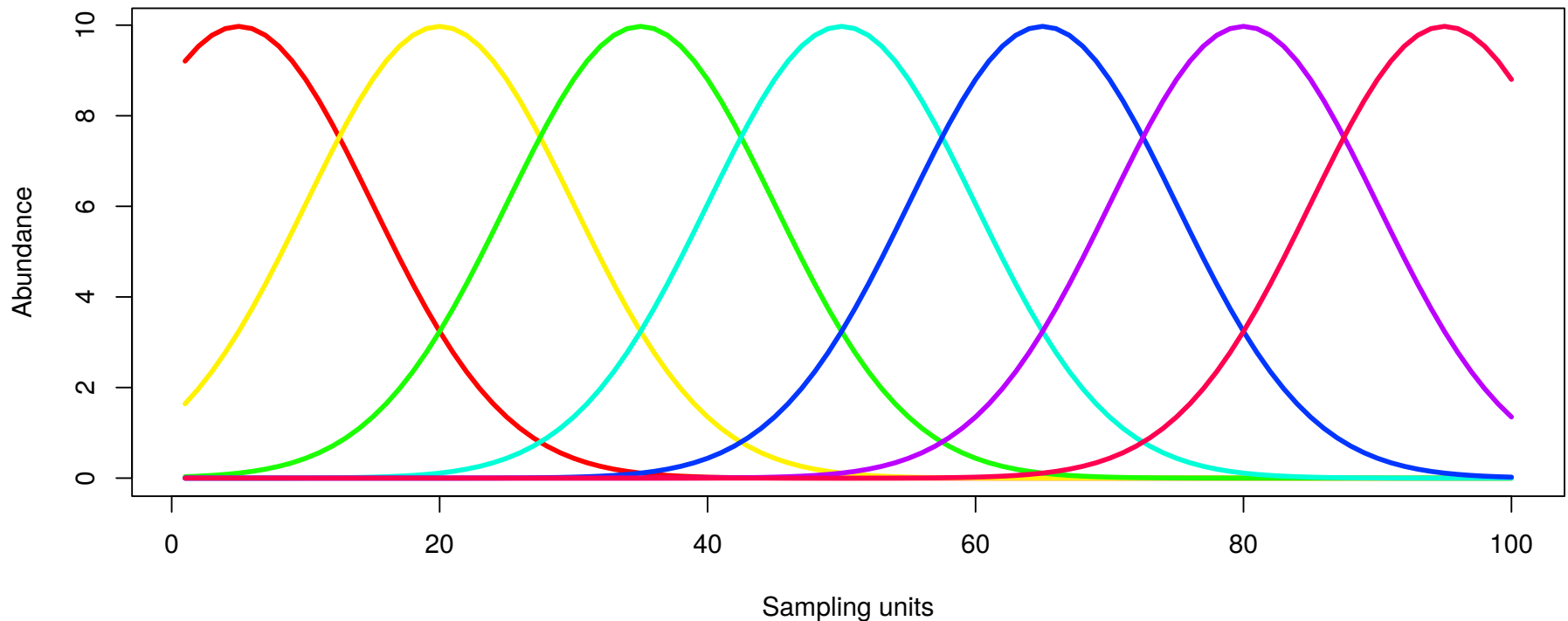
# Reciprocal averaging

Repeat until site and species scores converged (remain constant)  
Built a latent (unobserved) variable that best discriminates between species scores



# Weighted averaging representation of CA:

Consider a set of species with unimodal distributions arranged along an environmental gradient



The explanatory power of the gradient is measured by the degree of separation of the species

CA finds the theoretical (latent) variable that maximizes the separation of species - that is, best explains the species data in terms of a unimodal weighted averaging model

# Relation to PCA

CA is a PCA on transformed data

...	species 1	species 2	...	species i	$\Sigma$
site 1	$p_{11}$	$p_{12}$	...	$p_{1i}$	$p_{1+}$
site 2	$p_{21}$	$p_{22}$	...	$p_{2i}$	$p_{2+}$
...	...	...	...	...	...
site j	$p_{j1}$	$p_{j2}$	...	$p_{ji}$	$p_{j+}$
$\Sigma$	$p_{+1}$	$p_{+2}$	...	$p_{+i}$	$p_{++}$

Matrix  $\overline{Q}$

...	species 1	species 2	species i
site1	$q_{11} = \frac{p_{11} - p_{1+}p_{+1}}{\sqrt{p_{1+}p_{+1}}}$	$q_{12} = \frac{p_{12} - p_{1+}p_{+2}}{\sqrt{p_{1+}p_{+2}}}$	$q_{1i} = \frac{p_{1i} - p_{1+}p_{+i}}{\sqrt{p_{1+}p_{+i}}}$
site2	$q_{21} = \frac{p_{21} - p_{2+}p_{+1}}{\sqrt{p_{2+}p_{+1}}}$	$q_{22} = \frac{p_{22} - p_{2+}p_{+2}}{\sqrt{p_{2+}p_{+2}}}$	$q_{2i} = \frac{p_{2i} - p_{2+}p_{+i}}{\sqrt{p_{2+}p_{+i}}}$
sitej	$q_{j1} = \frac{p_{j1} - p_{j+}p_{+1}}{\sqrt{p_{j+}p_{+1}}}$	$q_{j2} = \frac{p_{j2} - p_{j+}p_{+2}}{\sqrt{p_{j+}p_{+2}}}$	$q_{ji} = \frac{p_{ji} - p_{j+}p_{+i}}{\sqrt{p_{j+}p_{+i}}}$

## PCA on $\overline{Q}$

- Site scores (Principal components, linear combinations)

...	PC1	PC2	...	PC i
site 1	$pc_{11}$	$pc_{12}$	...	$pc_{1i}$
site 2	$pc_{21}$	$pc_{22}$	...	$pc_{2i}$
...	...	...	...	...
site j	$pc_{j1}$	$pc_{j2}$	...	$pc_{jci}$

- Species scores (Loadings, Eigenvector)

...	ev1	ev2	...	ev i
species 1	$ev_{11}$	$ev_{12}$	...	$ev_{1i}$
species 2	$ev_{21}$	$ev_{22}$	...	$ev_{2i}$
...	...	...	...	...
species i	$ev_{i1}$	$ev_{i2}$	...	$ev_{ii}$

## CA site scores

...	species 1	species 2	...	species i	$\Sigma$
site 1	$q_{11}$	$q_{12}$	...	$q_{1i}$	$q_{1+}$
site 2	$q_{21}$	$q_{22}$	...	$q_{2i}$	$q_{2+}$
site j	$q_{j1}$	$q_{j2}$	...	$q_{ji}$	$q_{j+}$
$\Sigma$	$q_{+1}$	$q_{+2}$	...	$q_{+i}$	$q_{++}$

CA site scores: weighted averages of PC site scores (Principal component)

...	CA1	CA2	...	CA i
<i>site1</i>	$\frac{pc1_1}{\sqrt{\frac{q_{1+}}{q_{++}}}}$	$\frac{pc2_1}{\sqrt{\frac{q_{1+}}{q_{++}}}}$	...	$\frac{pci_1}{\sqrt{\frac{q_{1+}}{q_{++}}}}$
<i>site2</i>	$\frac{pc1_2}{\sqrt{\frac{q_{2+}}{q_{++}}}}$	$\frac{pc2_2}{\sqrt{\frac{q_{2+}}{q_{++}}}}$	...	$\frac{pci_2}{\sqrt{\frac{q_{2+}}{q_{++}}}}$
<i>sitej</i>	$\frac{pc1_j}{\sqrt{\frac{q_{j+}}{q_{++}}}}$	$\frac{pc2_j}{\sqrt{\frac{q_{j+}}{q_{++}}}}$	...	$\frac{pci_j}{\sqrt{\frac{q_{j+}}{q_{++}}}}$

## CA species scores

...	species 1	species 2	...	species i	$\Sigma$
site 1	$q_{11}$	$q_{12}$	...	$q_{1i}$	$q_{1+}$
site 2	$q_{21}$	$q_{22}$	...	$q_{2i}$	$q_{2+}$
site j	$q_{j1}$	$q_{j2}$	...	$q_{ji}$	$q_{j+}$
$\Sigma$	$q_{+1}$	$q_{+2}$	...	$q_{+i}$	$q_{++}$

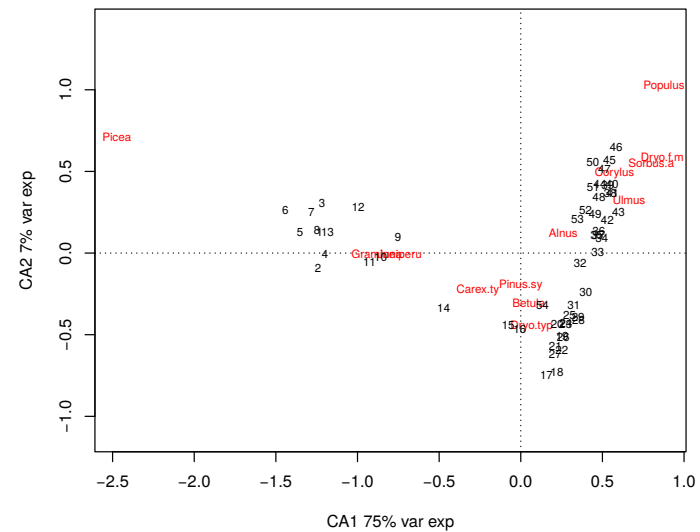
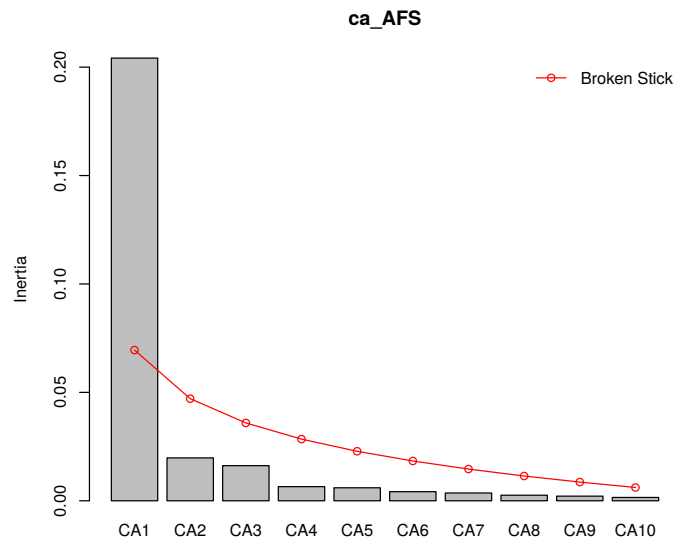
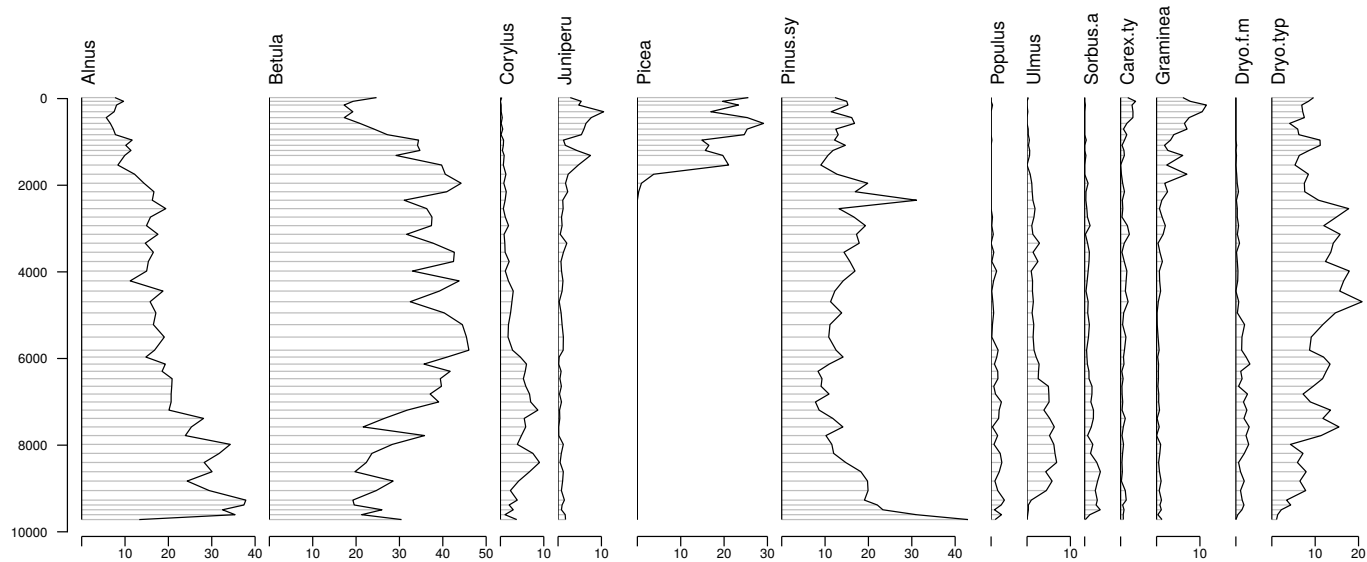
CA species scores: weighted averages of PC species scores (eigenvector)

...	CA axis 1	CA axis 2	...	CA axis i
<i>species1</i>	$\frac{ev1_1}{\sqrt{\frac{q_{+1}}{q_{++}}}}$	$\frac{ev2_1}{\sqrt{\frac{q_{+1}}{q_{++}}}}$	...	$\frac{evl_1}{\sqrt{\frac{q_{+1}}{q_{++}}}}$
<i>species2</i>	$\frac{ev1_2}{\sqrt{\frac{q_{+2}}{q_{++}}}}$	$\frac{ev2_2}{\sqrt{\frac{q_{+2}}{q_{++}}}}$	...	$\frac{evl_2}{\sqrt{\frac{q_{+2}}{q_{++}}}}$
<i>speciesl</i>	$\frac{ev1_i}{\sqrt{\frac{q_{+i}}{q_{++}}}}$	$\frac{ev2_i}{\sqrt{\frac{q_{+i}}{q_{++}}}}$	...	$\frac{evl_i}{\sqrt{\frac{q_{+i}}{q_{++}}}}$

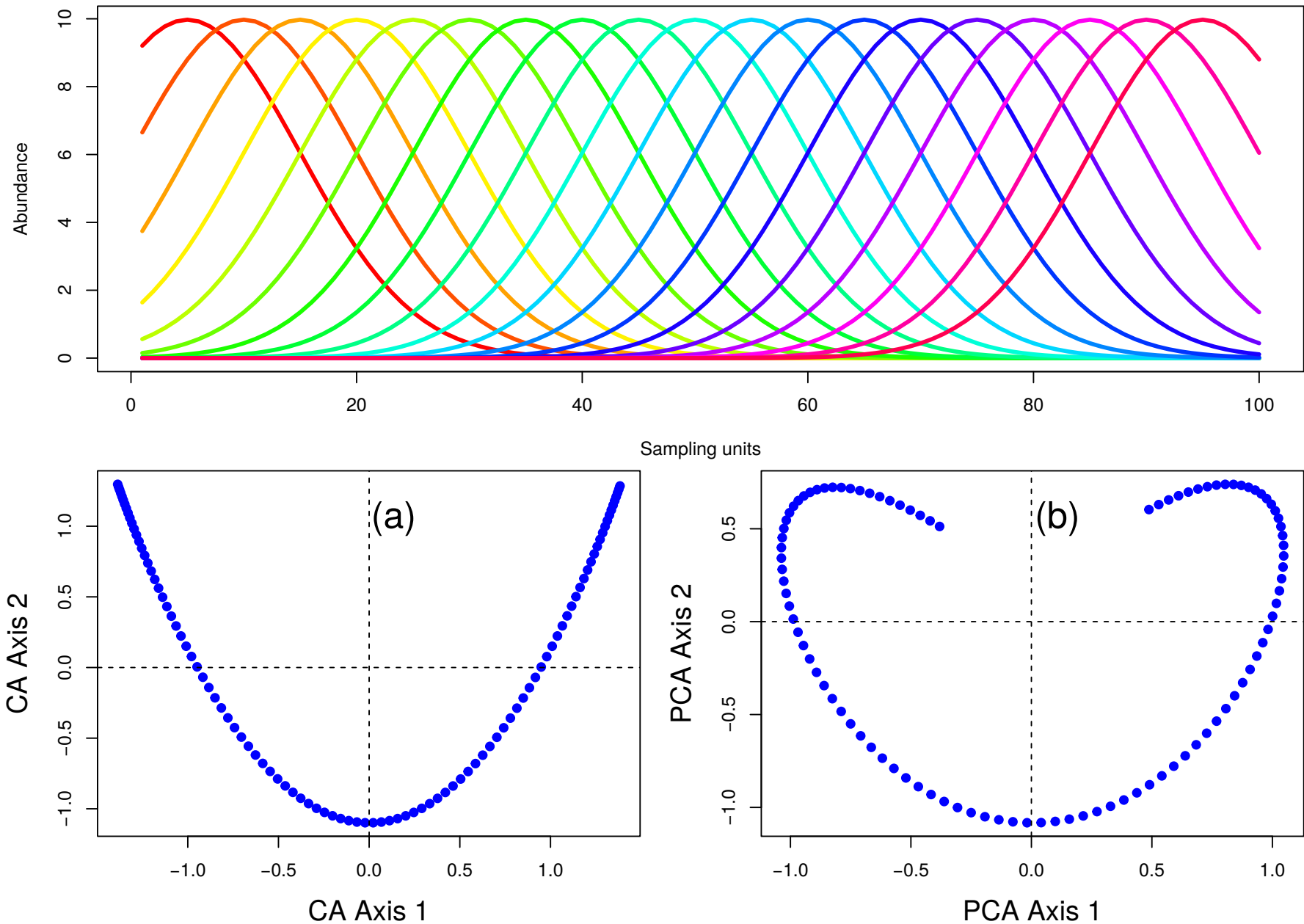
# Correspondence analysis

- ▶ Pre transform data
- ▶ Run PCA
- ▶ Post transform site and species scores (weighted average)
- ▶ Unimodal response
- ▶ Distances among samples preserved in CA:  $\chi^2$  distance

# CA example: Kinnshaugen southern Norway



# CA arch effect - PCA horse shoe effect





# CA arch effect

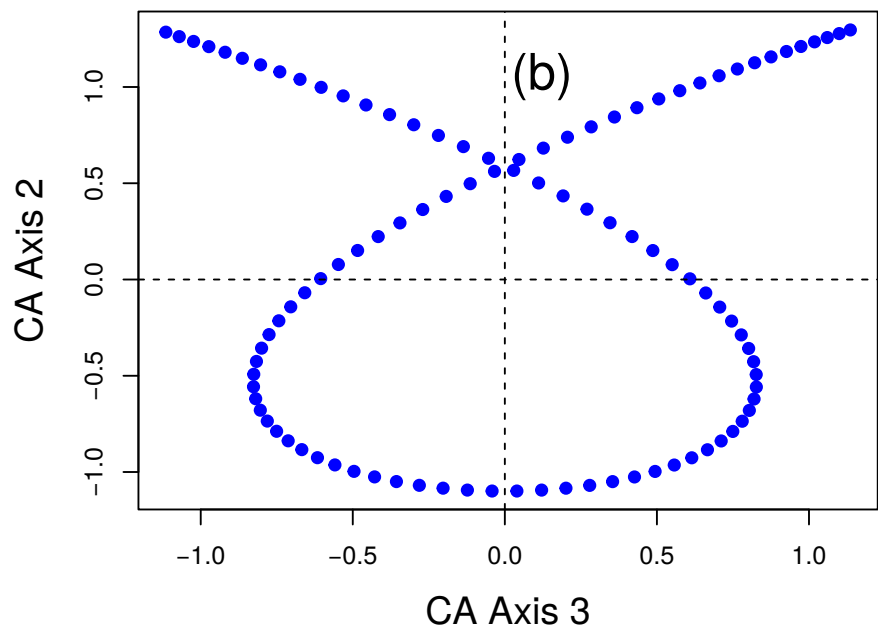
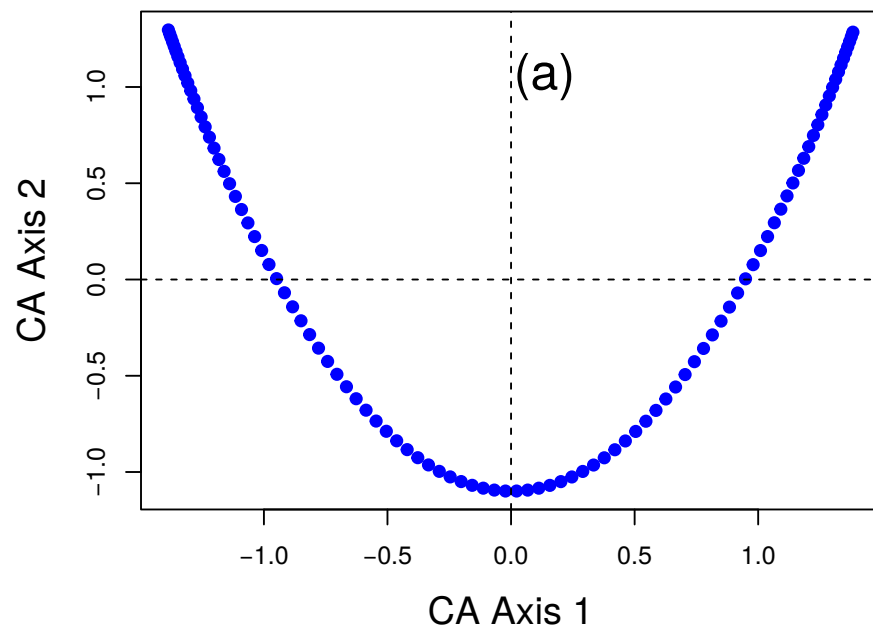
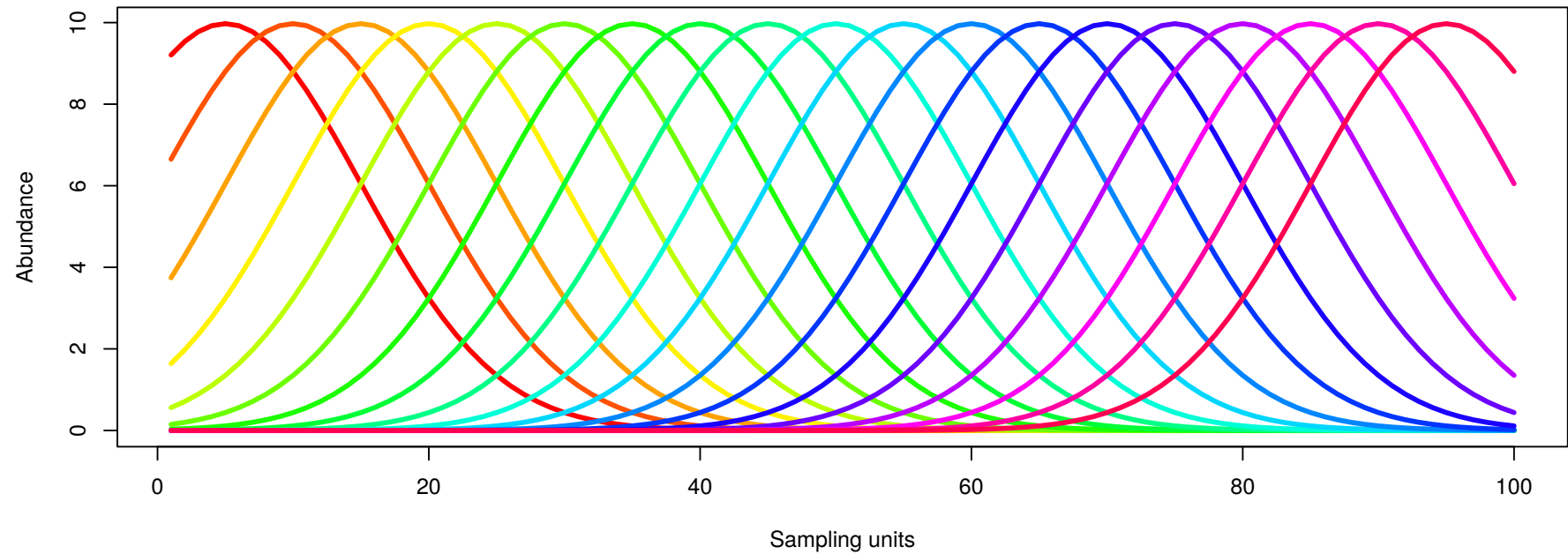
Long environmental gradient:

- ▶ species at either end of the gradient do not co-occur
- ▶ correlation between those species  $r = 0$
- ▶ angle between species scores:  $90^\circ$  and not  $180^\circ$

Ninety degree angle in 3 dimensions:

- ▶ lower (upper) end of the gradient:
  - ▶ negative (positive) on axis 1
  - ▶ positive on axis 2
  - ▶ negative (positive) on axis 3

# CA Arch effect



# CA arch effect - PCA horse shoe effect

**Arch effect:** only interpret axis one

**Horse shoe effect:** find a recycling bin...

Alternatives: Detrended correspondence analysis

- ▶ axis 1 is virtually unchanged from CA
- ▶ not possible to interpret axis 2

Non metric multidimensional scaling (NMDS)

Principal curves (flexible response model instead of WA)

In our crafted example:  $DCA > CA > NMDS >>>>>>>>$

Principal curve (unimodal responses favor DCA and CA)

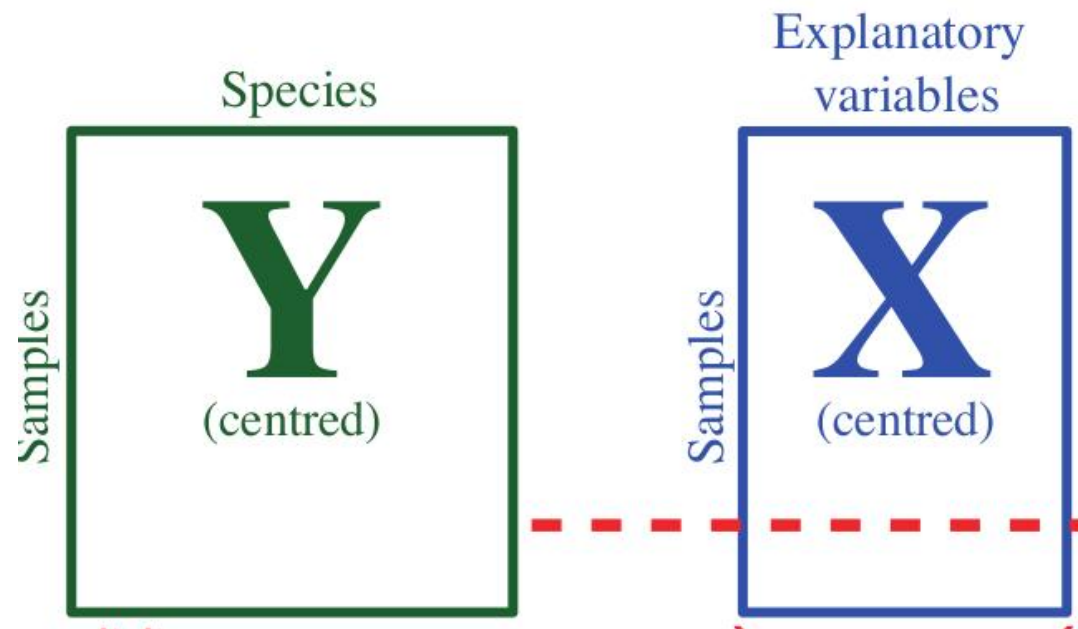
# Constrained ordination

Goal: explain dependent variables (usually species data) as function of independent variables (usually environmental variables)

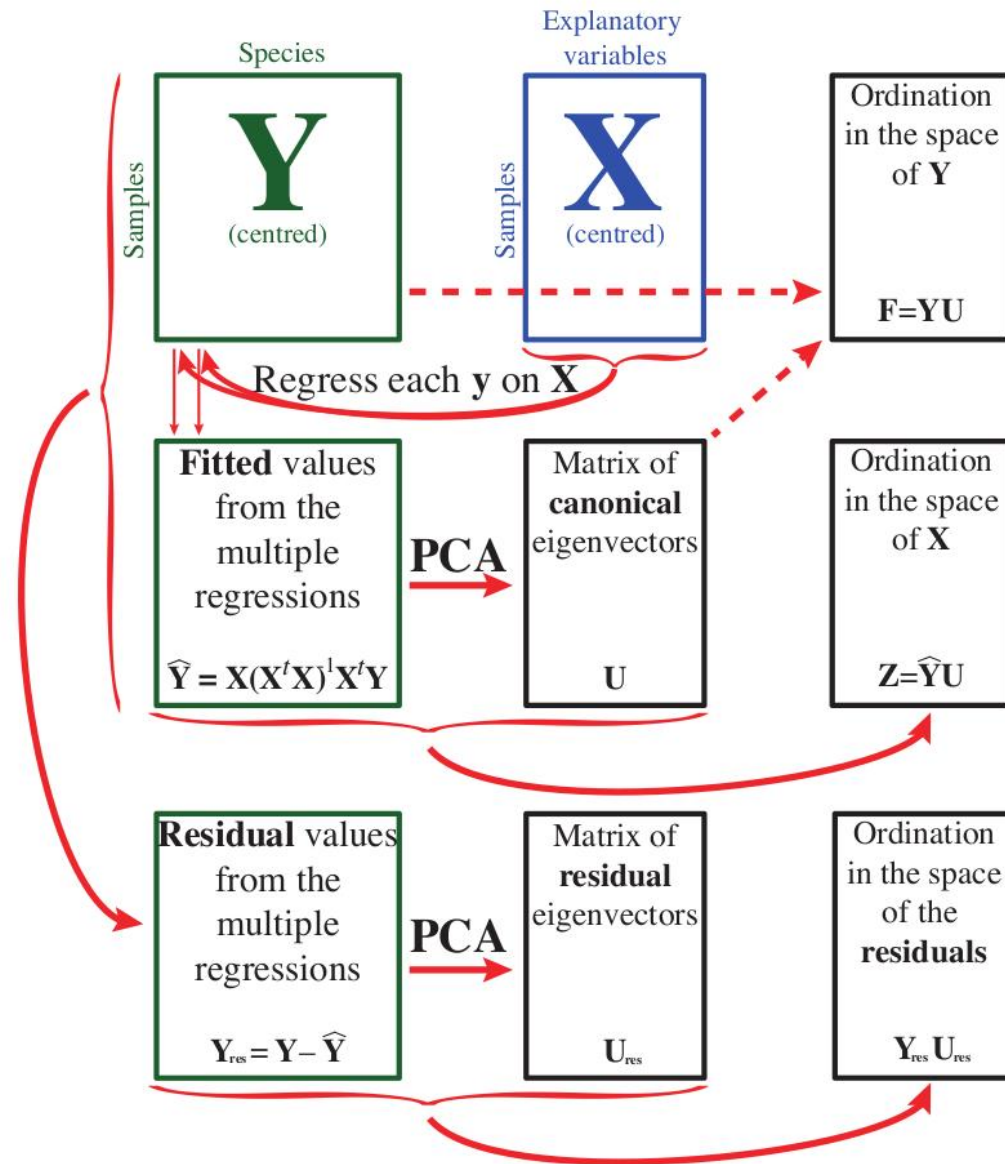
PCA and CA: axes chosen freely

Constrained ordination:

- ▶ ordination axes are linear combinations of independent variables
- ▶ allows assessment of influence of environmental variables on species assemblages



# Redundancy analysis



# Redundancy analysis

Two datasets:

- ▶ explanatory data set (e.g. environmental data): **X**
- ▶ response data set (e.g. species assemblages): **Y**

Two steps:

- ▶ Regression
- ▶ Ordination

# Redundancy analysis: One environmental variable

## Regression

Fitted values

$$\hat{y}_1 = a + bx$$

$$\hat{y}_2 = a + bx$$

...

$$\hat{y}_i = a + bx$$

$$\text{Reminder: } b = \frac{s_{xy}^2}{s_x} = r_{xy} s_y$$

Residuals

$$y_{1,res} = y_1 - \hat{y}_1$$

$$y_{2,res} = y_2 - \hat{y}_2$$

...

$$y_{i,res} = y_i - \hat{y}_i$$

# Redundancy analysis: Ordination

One constraining variable

PCA on  $\hat{\mathbf{Y}}$ : RDA axis 1

PCA on  $\mathbf{Y}_{\text{res}}$ : PCA axes 1 to (n-1)

RDA site scores: rescaled environmental variable

RDA species scores: rescaled regression coefficients



# Redundancy analysis

**Total variance:** Sum of variances of individual variables in  $y$

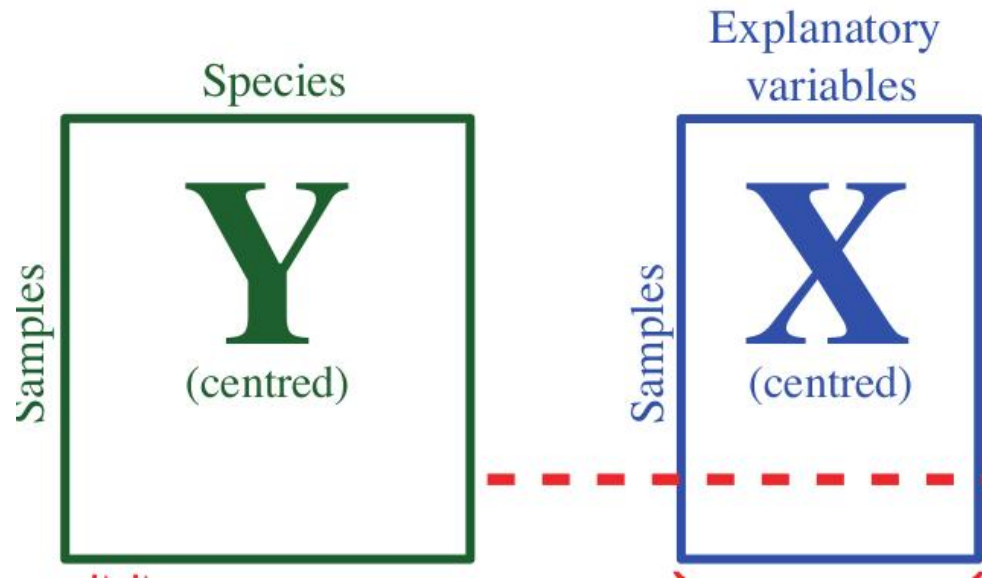
**Variance explained by RDA:** Variance of all variables in  $\hat{Y}$

**Variance explained by PCA:** Variance of all variables in  $Y_{res}$

**Quantity of interest:**

- ▶ Variance explained by RDA axis 1 divided by variance explained by PCA 1
- ▶ Is the first constrained gradient more important than the first unconstrained gradient?
- ▶ Ideally:  $\frac{\lambda_{RDA1}}{\lambda_{PCA1}} \gg 1$

# Redundancy Analysis: Regression and Ordination



fitted values  $\hat{y}_1 = a + b_1x_1 + b_2x_2 + \dots b_lx_l$

residuals  $y_{1,res} = y_1 - \hat{y}_1$

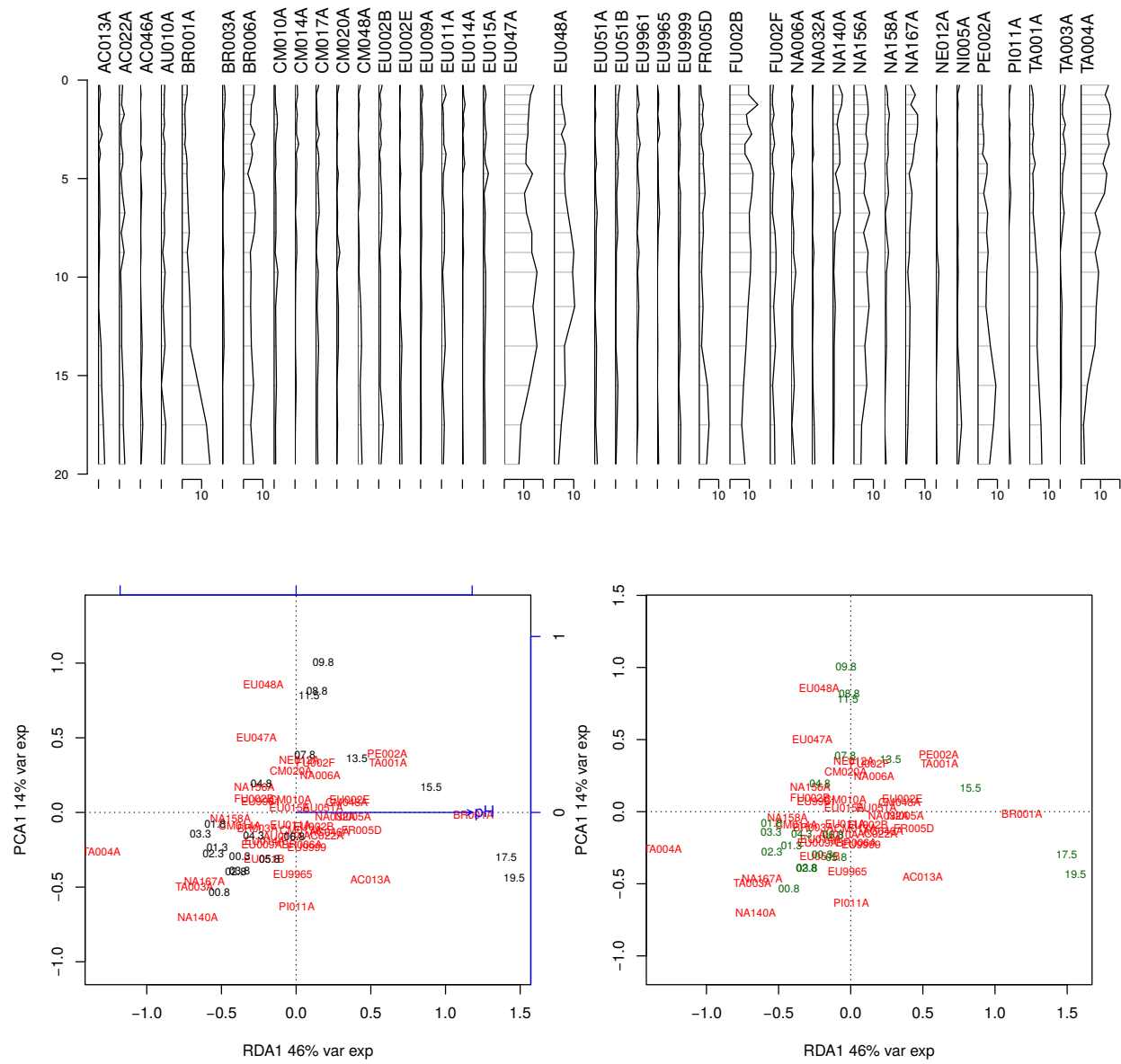
matrix of fitted values  $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$

matrix of residuals  $\mathbf{Y}_{res} = \mathbf{Y} - \hat{\mathbf{Y}}$

PCA on  $\hat{\mathbf{Y}}$ : RDA axes 1 to  $l$

PCA on  $\mathbf{Y}_{res}$ : PCA axes 1 to  $(n-l)$

# RDA example: Diatoms Round Loch of Glenhead and pH



## How good is the representation:

Species (optima) should be widely dispersed: Measured by between species variance  $SS_B$

Species responses or tolerances should be narrow: measured by within-site variance of species optima  $SS_W$

Total variance is  $SS_T = SS_B + SS_W$

If  $SS_B$  is high relative to  $SS_W$  (i.e. good separation of species) scaling is optimal

Eigenvalue ( $\lambda$ ) =  $\frac{SS_B}{SS_T}$

$\lambda$  lies between 0 and 1

“Variance” explained =  $\frac{\lambda_j}{\sum_{i=1}^K \lambda_i}$