

## Canonical correspondence analysis and weighted averaging

## Types of ordinations:

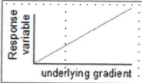
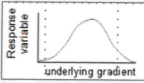
We talked about two types of ordinations and about two types of response models:

What is the main difference between the two ordination types?

What is the main difference between the two response models?

# Ordination types:

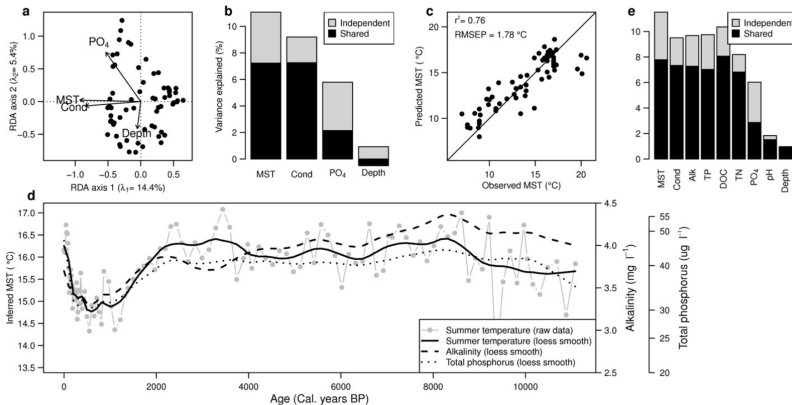
## Classification of ordination methods

| Response model                            |          | Linear  | Unimodal  | Complex (Unimodal, linear, skewed)         |
|---|----------|---|---|--|
|   |          |  |  |  |
| Distance measure                          |          | Euclidean   | Chi-square  | Many to choose from                        |
| Role of explanatory variables in analysis | Indirect | Principal Components Analysis (PCA)   | Correspondence Analysis (CA) & Detrended Correspondence Analysis (DCA)            | non-Metric Multidimensional Scaling (nMDS) |
|   | Direct   | Redundancy Analysis (RDA)   | Canonical Correspondence Analysis (CCA)   | Non-Euclidean RDA                          |

# Application of ordinations and reconstructions

*S. Juggins / Quaternary Science Reviews 64 (2013) 20–32*

23

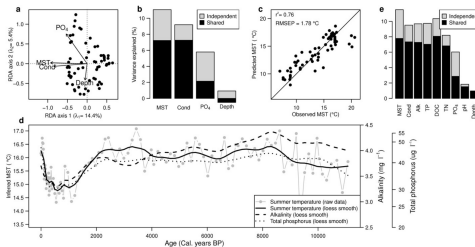


Juggins (2013)

# Application of ordinations and reconstructions

S. Juggins / Quaternary Science Reviews 64 (2013) 20–32

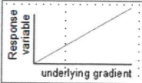
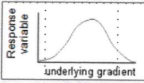
23



- Constrained ordination
- Independent and shared variance
- Transfer function
- Reconstructions
- Variable selection

# Ordination types:

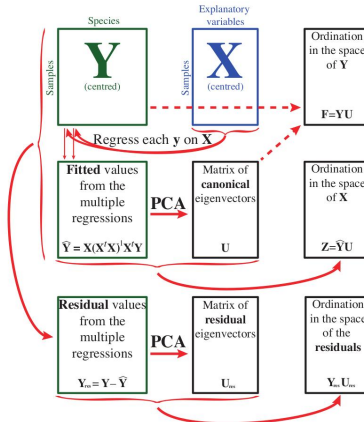
## Classification of ordination methods

| Response model                            |          | Linear  | Unimodal  | Complex (Unimodal, linear, skewed)         |
|---|----------|---|---|--|
|   |          |  |  |  |
| Distance measure                          |          | Euclidean   | Chi-square  | Many to choose from                        |
| Role of explanatory variables in analysis | Indirect | Principal Components Analysis (PCA)   | Correspondence Analysis (CA) & Detrended Correspondence Analysis (DCA)            | non-Metric Multidimensional Scaling (nMDS) |
|   | Direct   | Redundancy Analysis (RDA)   | Canonical Correspondence Analysis (CCA)   | Non-Euclidean RDA                          |

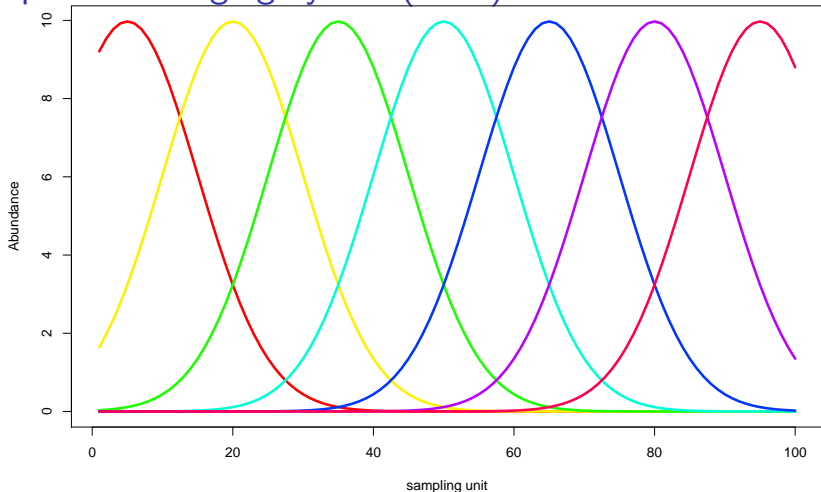
# Canonical correspondence analysis

Same principle as redundancy analysis:

- ▶ **regression:** weighted regression
- ▶ **ordination:** correspondence analysis



## Reciprocal averaging by Hill (1973):



*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

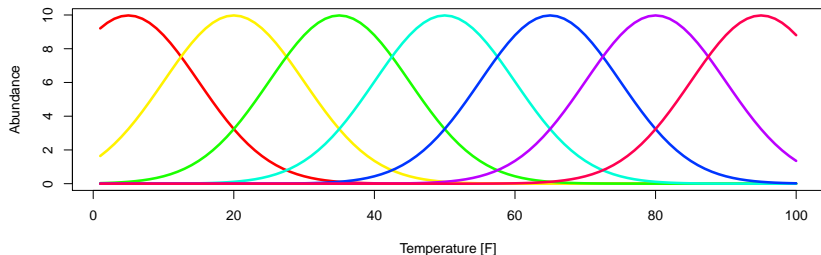
| ...      | species 1 | species 2 | ... | species i | $\Sigma$ |
|----------|-----------|-----------|-----|-----------|----------|
| site 1   | $p_{11}$  | $p_{12}$  | ... | $p_{1i}$  | $p_{1+}$ |
| site 2   | $p_{21}$  | $p_{22}$  | ... | $p_{2i}$  | $p_{2+}$ |
| ...      | ...       | ...       | ... | ...       | ...      |
| site M   | $p_{M1}$  | $p_{M2}$  | ... | $p_{Mi}$  | $p_{M+}$ |
| $\Sigma$ | $p_{+1}$  | $p_{+2}$  | ... | $p_{+i}$  | $p_{++}$ |

**Site scores:**  $x$

**Species scores:**  $y$

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

# Canonical correspondence analysis



| ...      | species 1 | species 2 | ... | species i | $\Sigma$ | Env. var. |
|----------|-----------|-----------|-----|-----------|----------|-----------|
| site 1   | $p_{11}$  | $p_{12}$  | ... | $p_{1i}$  | $p_{1+}$ | $env1_1$  |
| site 2   | $p_{21}$  | $p_{22}$  | ... | $p_{2i}$  | $p_{2+}$ | $env1_2$  |
| ...      | ...       | ...       | ... | ...       | ...      | ...       |
| site M   | $p_{M1}$  | $p_{M2}$  | ... | $p_{Mi}$  | $p_{M+}$ | $env1_M$  |
| $\Sigma$ | $p_{+1}$  | $p_{+2}$  | ... | $p_{+i}$  | $p_{++}$ |           |

**Environmental variable:** env

## Canonical correspondence analysis

| ...      | species 1 | species 2 | ... | species i | $\Sigma$ | Env. var. |
|----------|-----------|-----------|-----|-----------|----------|-----------|
| site 1   | $p_{11}$  | $p_{12}$  | ... | $p_{1i}$  | $p_{1+}$ | $env1_1$  |
| site 2   | $p_{21}$  | $p_{22}$  | ... | $p_{2i}$  | $p_{2+}$ | $env1_2$  |
| ...      | ...       | ...       | ... | ...       | ...      | ...       |
| site M   | $p_{M1}$  | $p_{M2}$  | ... | $p_{Mi}$  | $p_{M+}$ | $env1_M$  |
| $\Sigma$ | $p_{+1}$  | $p_{+2}$  | ... | $p_{+i}$  | $p_{++}$ |           |

**Environmental variable:** env

Start with environmental variable to calculate species scores

**Species score: environmental preference of a species**

$$scores_{species} = \frac{p_{11}}{p_{+1}} env1_1 + \frac{p_{21}}{p_{+1}} env1_2 + \dots + \frac{p_{M1}}{p_{+1}} env1_M$$

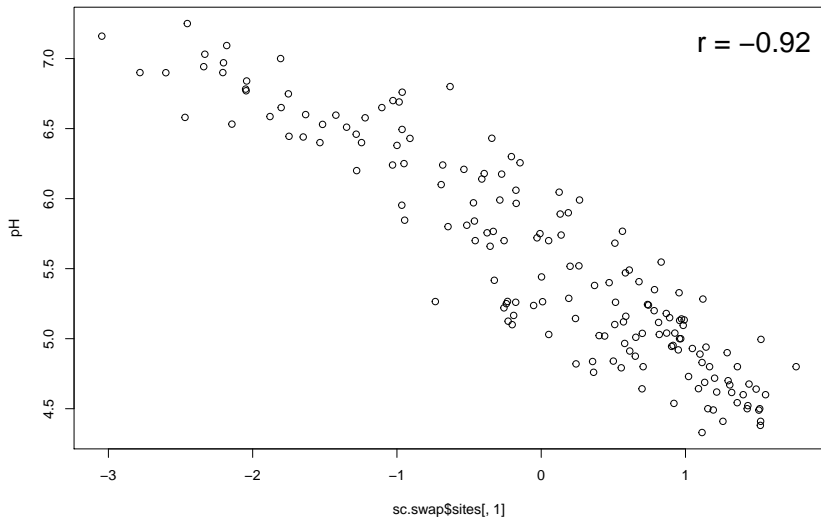
## Canonical correspondence analysis: site scores

| ...      | species 1 | species 2 | ... | species i | $\Sigma$ | Env. var. |
|----------|-----------|-----------|-----|-----------|----------|-----------|
| site 1   | $p_{11}$  | $p_{12}$  | ... | $p_{1i}$  | $p_{1+}$ | $env1_1$  |
| site 2   | $p_{21}$  | $p_{22}$  | ... | $p_{2i}$  | $p_{2+}$ | $env1_2$  |
| ...      | ...       | ...       | ... | ...       | ...      | ...       |
| site M   | $p_{M1}$  | $p_{M2}$  | ... | $p_{Mi}$  | $p_{M+}$ | $env1_M$  |
| $\Sigma$ | $p_{+1}$  | $p_{+2}$  | ... | $p_{+i}$  | $p_{++}$ |           |

**Site scores: environmental conditions at a site as determined by the environmental preferences of species found at that site**

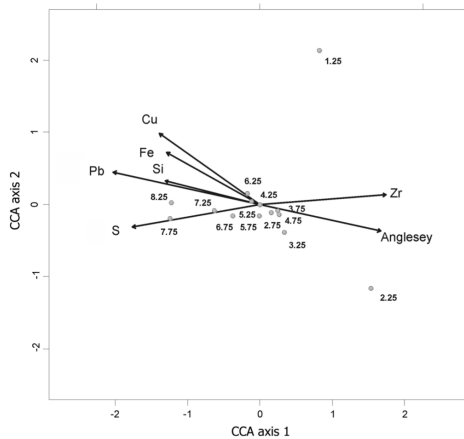
$$scores_{site} = \frac{p_{11}}{p_{1+}} scores_{spec,1} + \frac{p_{12}}{p_{1+}} scores_{spec,2} + \dots + \frac{p_{1N}}{p_{1+}} scores_{spec,N}$$

## Compare site scores to environmental variable; Diatoms and pH

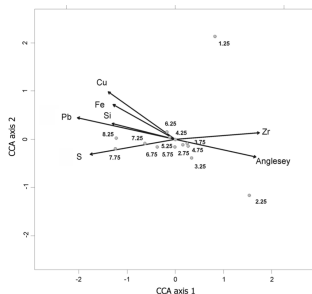




## CCA: bi or triplot



# Variable selection:



## Issues:

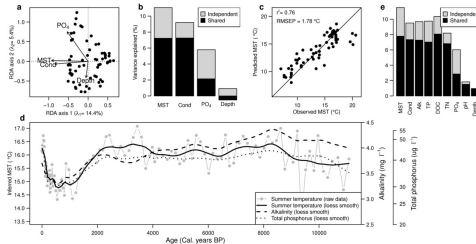
- ▶ many, highly correlated variables
- ▶ 7 variables for 14 samples
  - ▶ which variables are really important?
  - ▶ how do we select them?



# Variable selection:

S. Juggins / Quaternary Science Review 64 (2013) 20–32

23



## Different selection and inclusion criteria

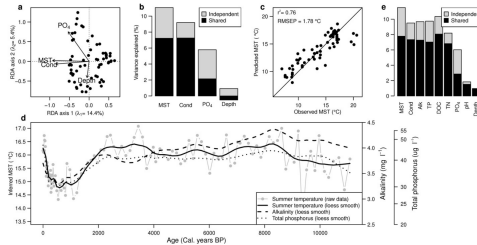
- ▶ forward selection
- ▶ backward elimination
- ▶ Reviewed in Borcard et al. (2011, see github repository)
- ▶ **Always use your ecological knowledge**

# Correlated environmental variables:

Part of the variance explained is shared between the variables

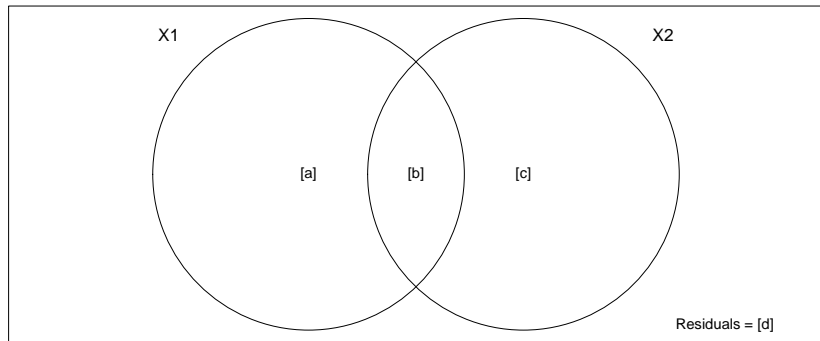
S. Juggins / Quaternary Science Reviews 64 (2013) 20–32

23



## Correlated environmental variables:

Part of the variance explained is shared between variables

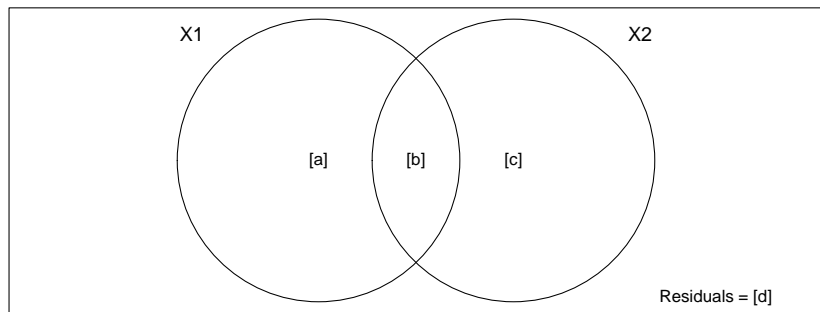


**a:** variance exclusively explained by X1

**c:** variance exclusively explained by X2

**b:** variance shared between X1 and X2

## Variance partitioning:



Step1: Remove influence of one environmental variable (**X1**) on species assemblage

- fit model with env variable you want to condition on (**[a]** + **[b]**) and then take residuals

Step2: Constrain residuals using second environmental variable (**X2**) (**[c]**)

## Variance partitioning: RDA

**varpart** in **vegan** package

| ##                 | Df | R.squared | Adj.R.squared | Testable |
|--------------------|----|-----------|---------------|----------|
| ## [a+b] = X1      | 1  | 0.2289742 | 0.2280408     | TRUE     |
| ## [b+c] = X2      | 1  | 0.1482025 | 0.1471713     | TRUE     |
| ## [a+b+c] = X1+X2 | 2  | 0.2895756 | 0.2878534     | TRUE     |

| ##                 | Df | R.squared | Adj.R.squared | Testable |
|--------------------|----|-----------|---------------|----------|
| ## [a] = X1 X2     | 1  | NA        | 0.1406821     | TRUE     |
| ## [b]             | 0  | NA        | 0.0873587     | FALSE    |
| ## [c] = X2 X1     | 1  | NA        | 0.0598126     | TRUE     |
| ## [d] = Residuals | NA | NA        | 0.7121466     | FALSE    |

## Variance partitioning: CCA

**cca** in **vegan**

```
cca.cond.tjul <- cca(sqrt(arctic.pollen) ~ taug + Condition(tjul),  
data= arctic.env)
```

```
cca.cond.taug <- cca(sqrt(arctic.pollen) ~ tjul + Condition(taug),  
data= arctic.env)
```

**Conditional:** [a] + [b] or [b]+[c]

**Constrained:** [c] or [a]

**Unconstrained:** residuals [d]

**Conditional + Constrained** = [a] + [b] + [c]

No adjusted  $R^2$

## Variance partitioning:

Transform data following Legendre and Gallagher (2001) and then run rda

**decostand()** in vegan

**varpart**

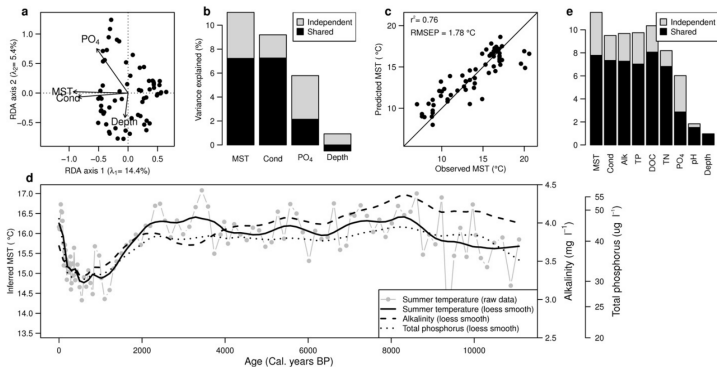




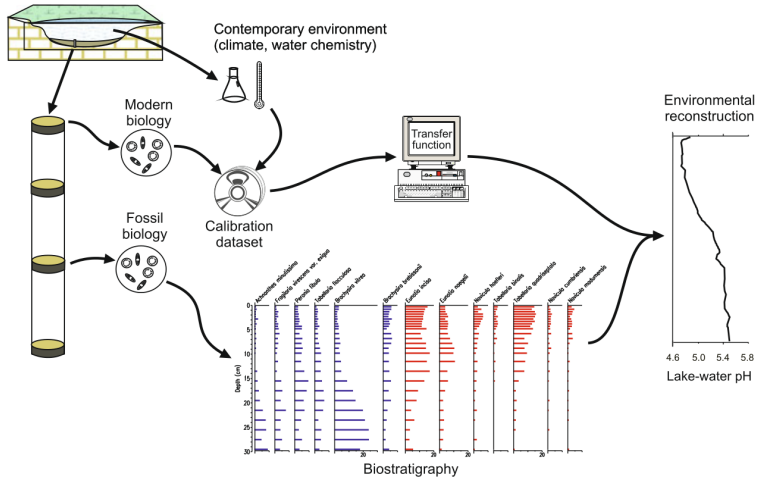
# Environmental reconstructions

*S. Juggins / Quaternary Science Reviews 64 (2013) 20–32*

23



# Environmental reconstructions

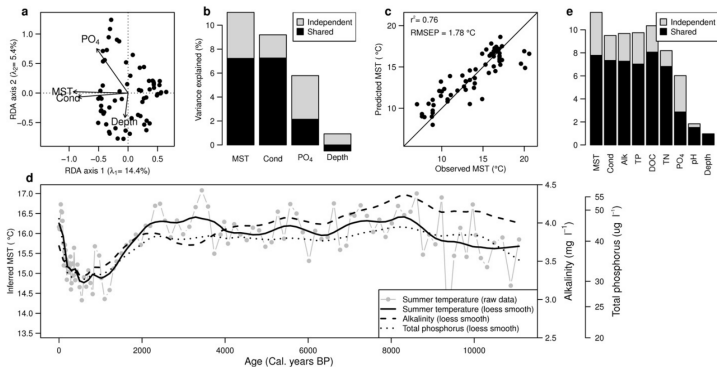


Juggins and Birks (2012)

# Environmental reconstructions

*S. Juggins / Quaternary Science Reviews 64 (2013) 20–32*

23



# Modern analogue technique vs weighted averaging

**MAT:** local solution (use  $k$  closest analogues and average environment of these analogues)

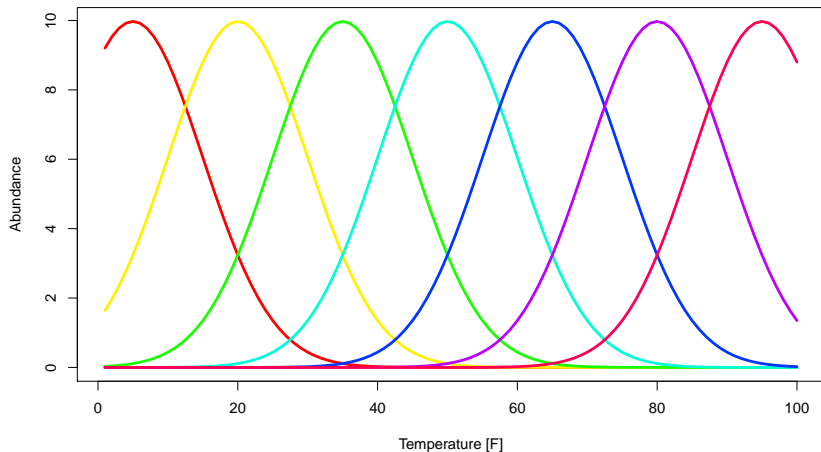
$k$ : number of analogues (usually a few)

**WA:** global solution estimating optimal environmental conditions for species/taxa

From optimal conditions for species estimate past environment based on species composition

Have we encountered this procedure before?

## WA: unimodal response



Estimate species optimum

Using species optima, estimate environmental conditions at a site

## WA and CCA: Optima and species scores

| ...      | species 1 | species 2 | ... | species i | $\Sigma$ | Env. var. |
|----------|-----------|-----------|-----|-----------|----------|-----------|
| site 1   | $p_{11}$  | $p_{12}$  | ... | $p_{1i}$  | $p_{1+}$ | $env_1$   |
| site 2   | $p_{21}$  | $p_{22}$  | ... | $p_{2i}$  | $p_{2+}$ | $env_2$   |
| ...      | ...       | ...       | ... | ...       | ...      | ...       |
| site M   | $p_{M1}$  | $p_{M2}$  | ... | $p_{Mi}$  | $p_{M+}$ | $env_M$   |
| $\Sigma$ | $p_{+1}$  | $p_{+2}$  | ... | $p_{+i}$  | $p_{++}$ |           |

**Optimum:** environmental preference of a species

**Environmental variable:** env

$$Optimum_{species1} = \frac{p_{11}}{p_{+1}} env_1 + \frac{p_{21}}{p_{+1}} env_2 + \dots + \frac{p_{M1}}{p_{+1}} env_M$$

**Optimum = CCA species scores**

## WA and CCA: Predicted value and site scores

| ...      | species 1 | species 2 | ... | species N | $\Sigma$ | Env. var. |
|----------|-----------|-----------|-----|-----------|----------|-----------|
| site 1   | $p_{11}$  | $p_{12}$  | ... | $p_{1N}$  | $p_{1+}$ | $env1_1$  |
| site 2   | $p_{21}$  | $p_{22}$  | ... | $p_{2N}$  | $p_{2+}$ | $env1_2$  |
| ...      | ...       | ...       | ... | ...       | ...      | ...       |
| site M   | $p_{M1}$  | $p_{M2}$  | ... | $p_{MN}$  | $p_{M+}$ | $env1_M$  |
| $\Sigma$ | $p_{+1}$  | $p_{+2}$  | ... | $p_{+N}$  | $p_{++}$ |           |

**Predicted values:** environmental conditions at a site as determined by the environmental preferences of species found at that site

$$Pred_{site} = \frac{p_{11}}{p_{1+}} Opt_{spec1} + \frac{p_{12}}{p_{1+}} Opt_{spec2} + \dots + \frac{p_{1N}}{p_{1+}} Opt_{specN}$$

**Predicted values = CCA site scores**

# Deshrinking

$\max(\text{Optima}) \leq \max(\text{env})$

$\min(\text{Optima}) \geq \min(\text{env})$

$\max(\text{predicted value}) \leq \max(\text{Optima})$

$\min(\text{predicted value}) \geq \min(\text{Optima})$

Reconstructed values have less variance than environmental variables

**Deshrinking:** increase variance of reconstructed values

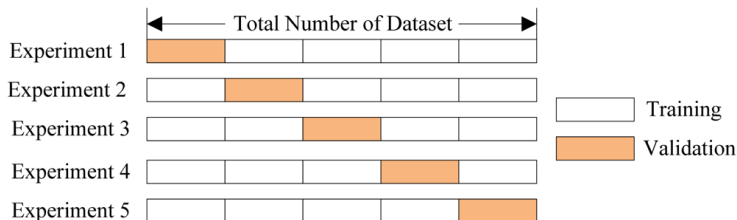
- ▶ inverse
- ▶ classical
- ▶ monotonic (predicted values no longer equal to CCA site scores but still monotonically related)



# Cross-validation

CCA and apparent WA: method knows environmental conditions it is supposed to estimate

Should test model on (hopefully) independent data: **calibration** and **validation**



# Cross-validation

- ▶ Divide data set in two parts
  - ▶ construct model (calibrate) on part one
  - ▶ test model (validate) on part two

## **Leave-one-out:**

- ▶ Use all except one sample to calibrate model, predict omitted sample (validate)

## **k-fold:**

- ▶ Divide data set into k-parts (e.g.  $k = 10$ )
- ▶ Use k-1 parts to calibrate and one part to validate
- ▶ repeat k times

# Outlook

- ▶ Weighted averaging partial least squares (WA-PLS)
- ▶ Validate predictions/reconstructions
- ▶ Cautionary notes