From GLM to fourth-corner correlation to double constrained correspondence analysis

Cajo J.F. ter Braak, Biometris, WUR

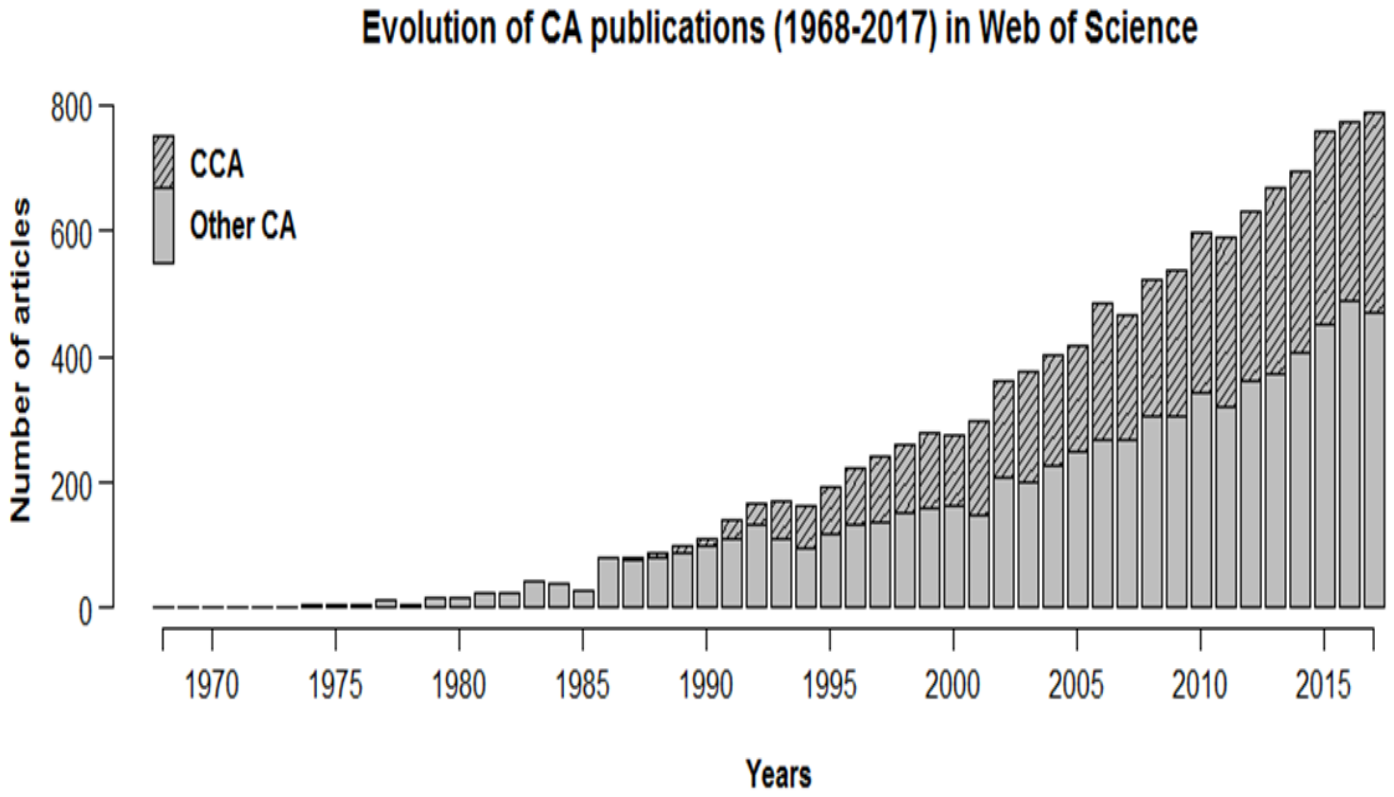
with Petr Šmilauer (České Budějovice),
Stéphane Dray (Lyon) and Pedro Peres-Neto (Montréal)

CARME, 4-6 February 2019, Stellenbosch, South Africa

WAGENINGEN
UNIVERSITY & RESEARCH

Cajo.terbraak@wur.nl

Papers using CA and constrained CA



From: Greenacre 2017 CA in practice. Japanese edition

MVA in ecology

Different schools seem to be fighting one-another...

- **Distance based - Oksanen, Faith, Minchin, M. Anderson,...(Primer)**
  - **CA & Chi-square distance bad; Gaussian model and niche packing unrealistic; Arch effect; trumpet shape ordinations in Detrended CA (DCA)**
  - **Choose your own distance * with NMDS and adonis**
- **PCA/RDA & CA/CCA - Greenacre , Dray, Ter Braak,...(Canoco)**
  - **Eigen value methods, biplots, joint plots and permutation testing**
  - **Simple species-based response models (linear, unimodal)**
- **GLM(M) – Warton (mvabund), Hui (boral)**, Ovaskainen, Yee (VGAM)**
  - **Site-based bootstrap to get the statistics right**

* Which distance is best to use is still an open question!
** Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R

MVA in ecology

Different schools seem to be fighting one-another...

- **Distance based - Oksanen, Faith, Minchin, M. Anderson,...(Primer)**
  - **CA & Chi-square distance bad; Gaussian model and niche packing unrealistic; Arch effect; trumpet shape ordinations in Detrended CA (DCA)**

But:

If A →B then the statement *notA* does not say anything on B, except than one needs another motivation. Example: the mean is the most efficient summary of the expectation of both the normal and the Poisson distribution. Not normal does not imply: do not use the mean.

Neither the chi-square distance nor the Gaussian model are necessary conditions for CA to be useful in ecological data, as we show in the sequel.

WAGENINGEN UNIVERSITY & RESEARCH

See also

ter Braak, C. J. F. and P. Šmilauer. 2015. Topics in constrained and unconstrained ordination. Plant Ecology **216**:683–696. http://edepot.wur.nl/323327

Holy grail of trait-based ecology

Lavorel & Garnier 2002
Funk et al. 2017

With climate and environmental change, can we predict the effects on

- **ecosystems?**
- **constituting species?**

Too many species... Too little is known...

Holy grail: understanding and predicting ecological processes from species traits (aka functional traits)

How to select relevant traits?

- **NB: for a particular set of environmental variables**

→ How to detect trait – environment association?

Issue: it is easy to get false positives

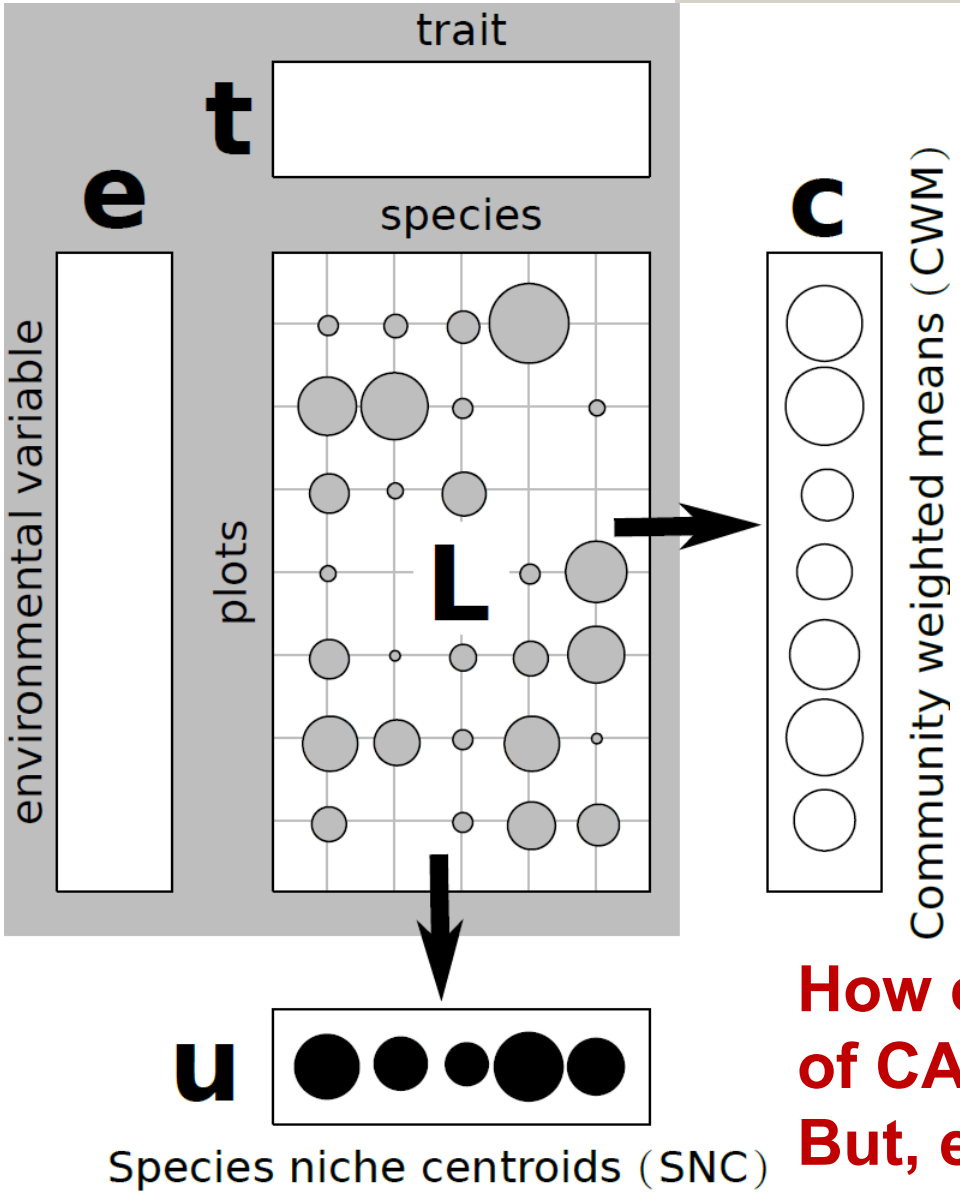
WAGENINGEN
UNIVERSITY & RESEARCH

Let's look at the simplest case first:
1 trait – 1 quant. environmental variable

Simplest case

L=**Y**= abundances, counts, cover or 1/0

L = 'link' between **e** and **t**; **Y** = response



300+ papers in WoS

CWM:

- $c_i = \sum_{j=1}^{S} y_{ij} t_j \,/ y_{i+}$

SNC:

- $u_j = \sum_{i=1}^{n} y_{ij} e_i \,/ y_{+j}$

In matrix notation

- $\mathbf{c} = \mathbf{R}^{-1}\mathbf{Y}\mathbf{t}$

  with R = diag(y_{i+})

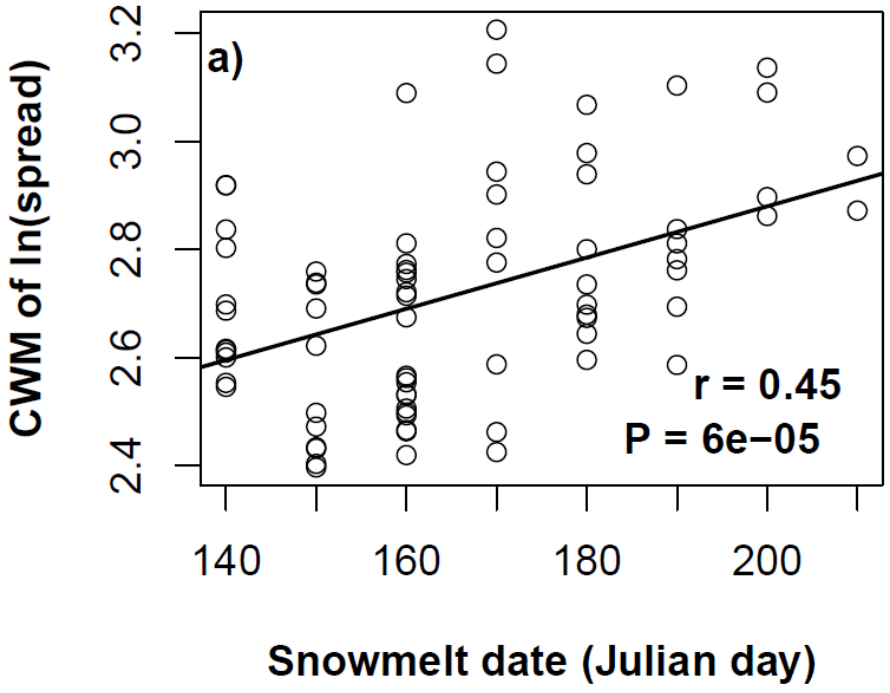
- $\mathbf{u} = \mathbf{K}^{-1}\mathbf{Y}^{T}\mathbf{e}$

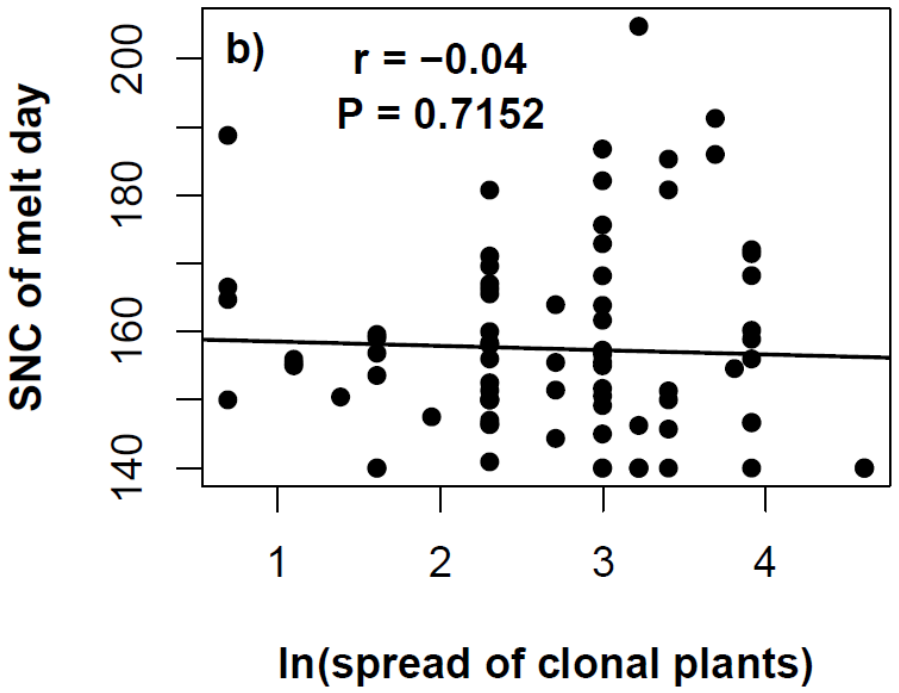
  with K = diag(y_{+j})

How close to transition formulae of CA and reciprocal averaging! But, e and t are given here.

Simple methods devised by ecologists

- Plot CWM against e or (and!) SNC against t





The correlations may have different signs!

Use weighted correlations instead, but these may still yield very different *P*-values

WAGENINGEN
UNIVERSITY & RESEARCH

Issues and alternatives

- What are the issues?
  - How is it possible, these different *P*-values?
- Alternative approaches:
  - Fourth-corner correlation (Legendre et al 1997, Dray & Legendre 2008, ter Braak et al 2012) and RLQ (Dolédec et al. 1996)
  - Model-based via GLM (traitglm in mvabund; Brown et al. 2014, Warton, Shipley & Hastie 2017, ter Braak et al. 2017)
  - Model-based via GLMM (Pollock et al. 2012, Jamil et al. 2013, Miller et al. 2018) Multilevel models: variance components!
- Do these control type I error and have enough power ?
- How to extend to the multi T and E case?
  - select relevant traits and environmental variables

Issues in T – E association

T and E lack a common observation unit:

- **the trait is observed on species,**
- **the environment on sites and**
- **the abundance on species-site combinations.**

How to *define* and *test* correlations between T and E?

Fourth-corner correlation (*f*, FC) (Legendre et al 1997):

$$f = cor_{\mathbf{Y}}(\boldsymbol{t}, \boldsymbol{e})$$

Weighted correlation between e and t (weights = vec(Y)) in the vectorised (inflated) data

Each cell of Y gives a row:

y_{ij} , e_i, t_j $(i = 1,\ldots, n; j = 1,\ldots,m)$

[e_i is repeated m times, t_j is repeated n times; note that this format allows for intra-specific variation]

Issues in T – E association

T and E lack a common observation unit:

- the trait is observed on species,
- the environment on sites and
- the abundance on species-site combinations.

How to *define* and *test* correlations between T and E?

Fourth-corner correlation (*f*, FC) (Legendre et al 1997):

$$f = cor_Y(\boldsymbol{t}, \boldsymbol{e})$$

$$f = \boldsymbol{cor}_{\boldsymbol{Y}}(\mathbf{t}, \mathbf{e}) = \frac{\sum_{i,j} y_{ij} \tilde{t}_j \tilde{e}_i}{\{\sum_j y_{+j} \tilde{t}_j^2 \ \sum_i y_{i+} \tilde{e}_i^2\}^{1/2}}$$

with

$\tilde{t}_j = t_j - \sum_j y_{+j} \, t_j / y_{++}$ and $\tilde{e}_i = \ e_i - \sum_i y_{i+} e_i / y_{++}$

WAGENINGEN
UNIVERSITY & RESEARCH

Statistical testing of t-e association

For t-e association to exist, two links must present

1. **Y↔e**
2. **Y↔t**

Dray & Legendre 2008; ter Braak et al 2012

Using Monte Carlo permutation tests:

Link 1 can be tested by permuting sites (values e_i)

Link 2 can be tested by permuting species (values t_j)

The *P*-values of these test must be combined by taking their maximum.

This is an example of the sequential testing theory*.

NB: the test statistic should be sensitive for the association and insensitive to other aspects of the data. *FC looks ok.*

WAGENINGEN UNIVERSITY & RESEARCH

* Goeman & Solari 2010

Why two tests?

- **Why is a single test not sufficient?**

Many people (including reviewers & myself) would say:

a researcher can only select/manipulate sites → permuting or resampling* sites should do.

However,

if there is a latent trait that is

- ***uncorrelated with t* and *interacts with e***

Then

- **the site-level test has inflated type I error**

Shown by simulation using log-linear (next slides)

and 1d and 2d Gaussian gradient models

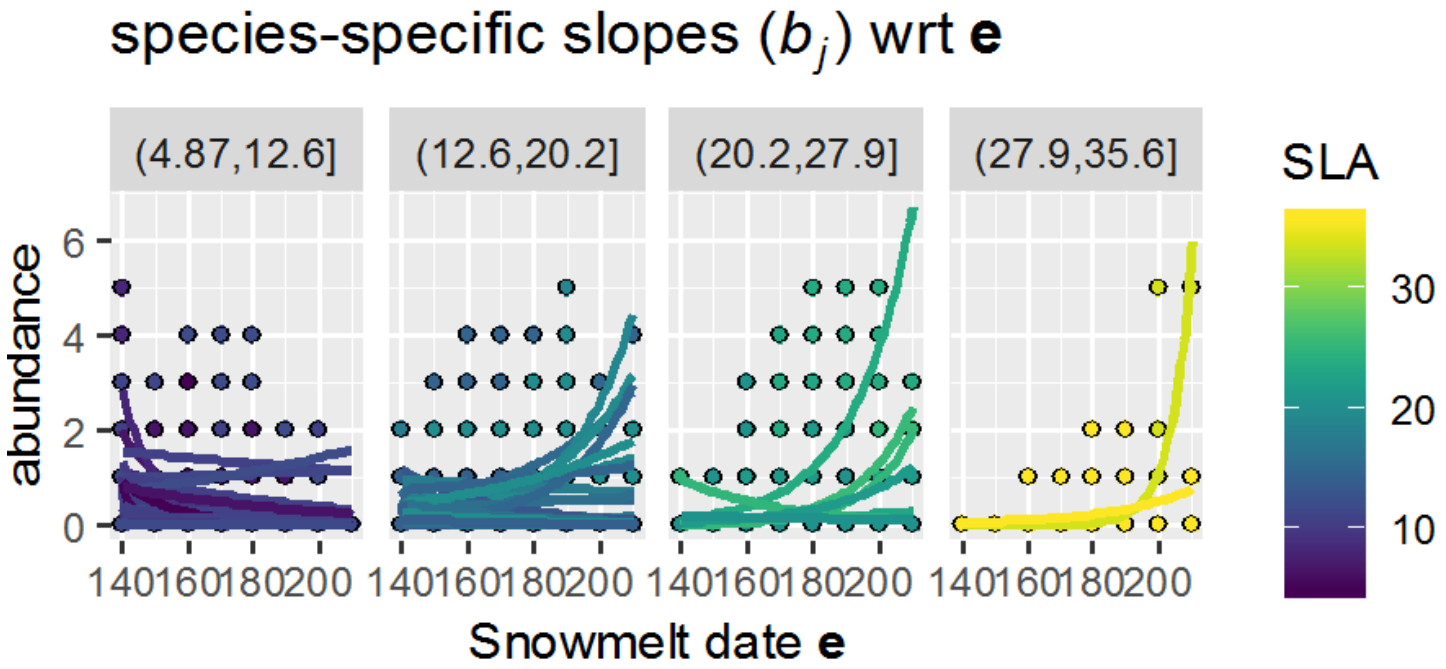
WAGENINGEN
UNIVERSITY & RESEARCH

* e.g. traitglm (anova) in mvabund

Simulation setup (summary)

- **The real world looks like a GLMM model** with negative binomial response, *i.e.*
  - there is (random) **species-specific response** wrt to e

  = there is a **latent uncorrelated trait** interacting with e



Simulation setup (summary)

- The real world looks like a GLMM model with negative binomial response, *i.e.*
  - there is (random) species-specific response wrt to e

  = there is a latent trait interacting with e
- But we analyse using simple GLM models (Poisson-loglinear models)
- Test null hypothesis of no trait-environment relation

H_0: `y ~ site + species`

H_1: `y ~ site + species + trait:env`

Fit by glm to the vectorised data (recall, this allows for intra-specific variation)

Derivation of GLMM model

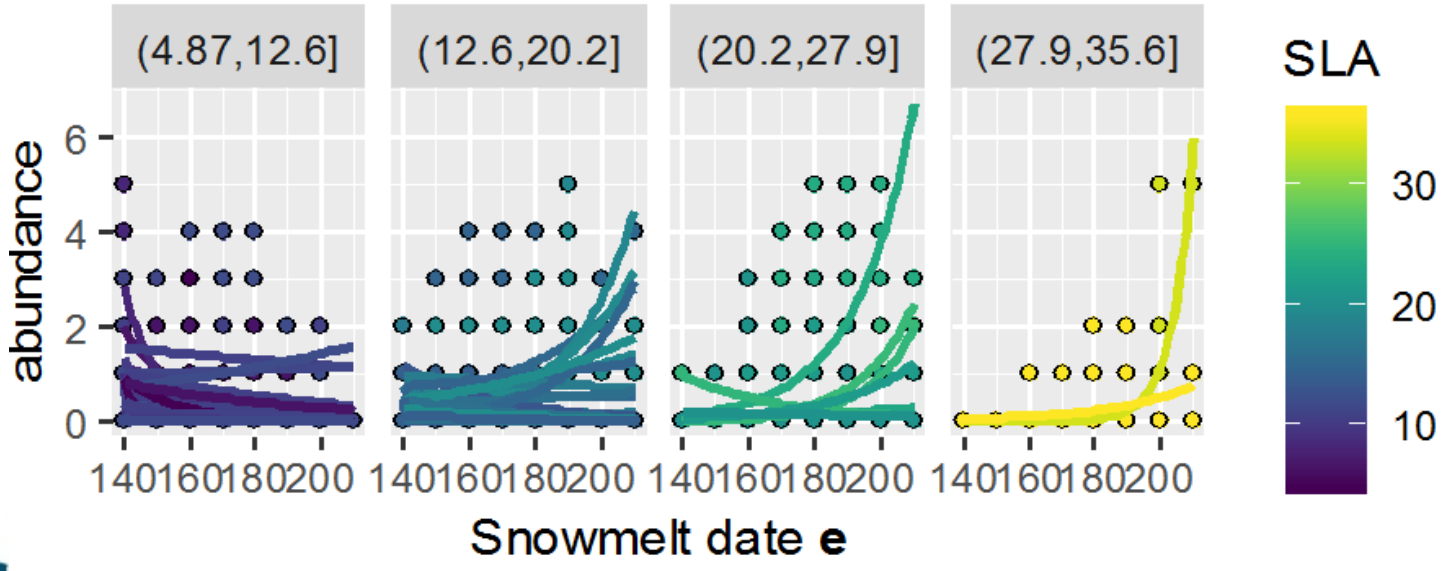
Pollock et al 2012 Jamil et al 2013

Abundance is a count y_{ij}, assumed to follow a distribution with mean specified by

$$\log(\mu_{ij}) = r_i + c_j + b_j e_i$$ **(a model without traits...)** **(1)**

- r_i and c_j row (site) and column (species) main effects (saturated main effects: e$\subseteq \{r_i\}$; t $\subseteq \{c_j\}$)
- b_j a **species-specific slope** with respect to e

species-specific slopes (b_j) wrt **e**



Derivation of GLMM model

Pollock et al 2012 Jamil et al 2013

Abundance is a count y_{ij}, assumed to follow a distribution with mean specified by

$$\log(\mu_{ij}) = r_i + c_j + b_j e_i \quad \text{(a model without traits...)} \quad (1)$$

- r_i and c_j row (site) and column (species) main effects (saturated main effects: e$\subseteq \{r_i\}$)
- b_j a **species-specific slope** with respect to e

Sub-model for the slopes: $b_j \sim N(\beta_0 + \beta_{te} t_j, \sigma_b^2)$ **gives a GLMM model,** `y~ r + c + env +(env|species)`

Insert the trait model:

$$\log(\mu_{ij}) = r_i + c_j + \beta_{te} t_j e_i + \beta_{ze} z_j e_i, \quad (2)$$

with $\beta_{ze} = \sigma_b$ and $z_j \sim N(0,1)$

$z =$ ***latent interacting trait***

Hypotheses:

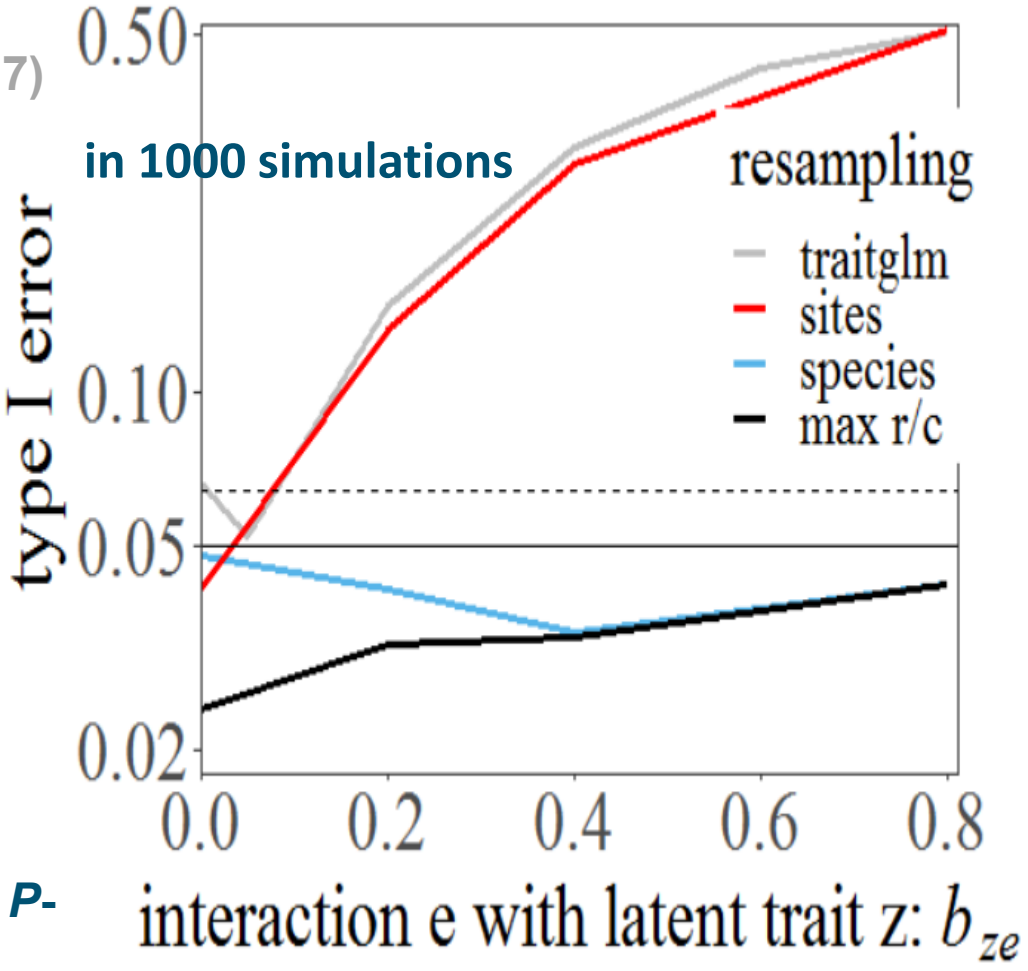
H_0: $\beta_{te} = 0$ **and H_1: with** $\beta_{te} \neq 0$.

WAGENINGEN UNIVERSITY & RESEARCH

False positives if there is a latent interacting trait

ter Braak et al 2017, PeerJ

- **traitglm** (Warton et al. 2017) site-based bootstrap (R package mvabund), negative binomial deviance
- **sites**: site-based, permutation of e, Poisson deviance
- **species**: species-based, permutation of t, Poisson deviance
- **max r/c**: Maximum of the site and species resampling *P*-values



→**variability of species-specific slopes wrt e**

max r/c follows from Goeman's sequential test

False positives if there is latent interacting trait

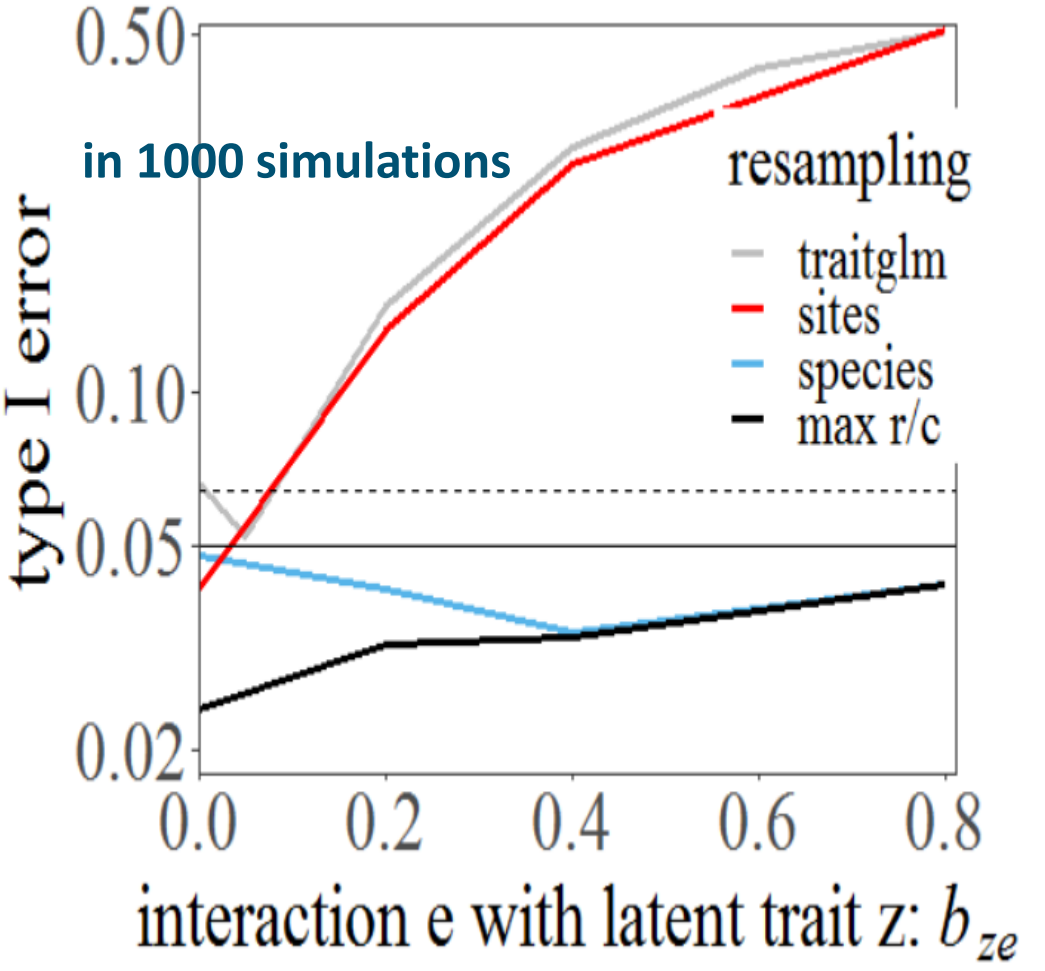
ter Braak et al 2017, PeerJ

Huge type I error rates for

site-level only tests (grey and red lines)

Ok for species-level test here, but not in scenario with a latent environmental variable interacting with the trait

Ok for max test in both scenarios



→variability of species-specific slopes wrt e

WAGENINGEN UNIVERSITY & RESEARCH

Failure of site-level only tests

ter Braak et al 2017, PeerJ, p.13

The issue is not that of *confounding* or *omitted variable*

> *confounding* is due to an *omitted variable* that is highly correlated with variable of interest and the predictor

In trait-environment problems, the failure :

> occurs also if there is an omitted variable that has zero correlation with the predictors

This is due to ignoring species and sites as a random factor, so as to account for

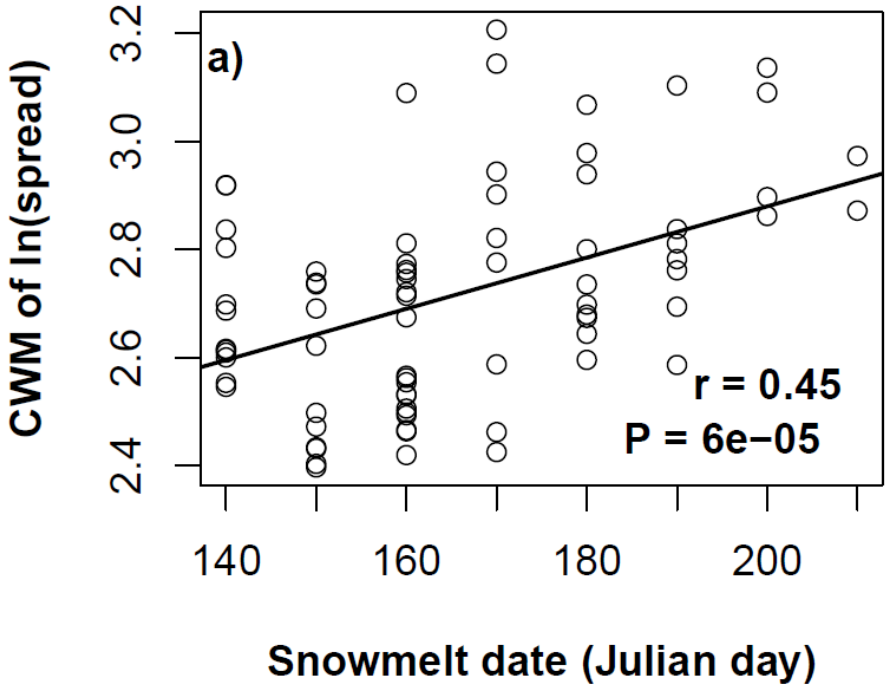
> species-specific response to the environment
>
> site-specific effects of the trait
>
> (both are realistic, important random effects)

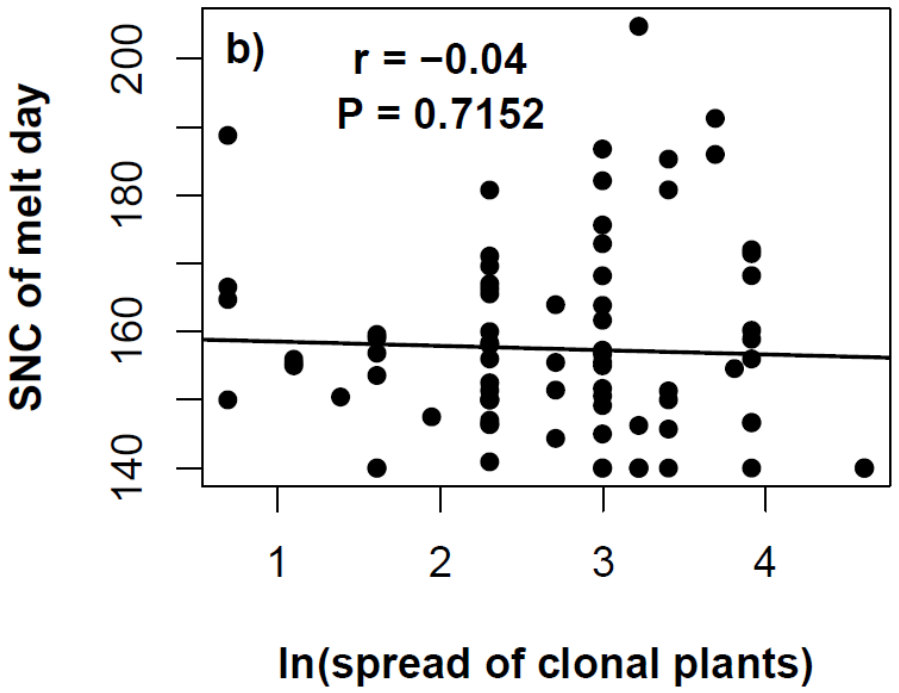
WAGENINGEN UNIVERSITY & RESEARCH

Conclusion: perform a species-level test too and take max *p*-values → max test

Illustrative example (recall)

- Plot CWM against e or (and?) SNC against t





The correlations may have different signs!

Use weighted correlations instead, but these may still yield very different *P*-values

WAGENINGEN
UNIVERSITY & RESEARCH

An illustrative example

ter Braak et al 2017, PeerJ, p.9/10

Shift of Alpine plant traits along a snow-melt gradient.

aravo data in R::ade4 (from Choler 2005)

- **Abundance of 82 plant species in 75 sites**

Association/interaction between

- **Trait: lateral Spread of species and**
- **Environmental variable: Snowmelt date ???**

-GLM test on interaction (site bootstrap) * : $p \approx 0.001$

-4th corner correlation with default resampling: $p \approx 0.36$**

Which one cannot be trusted and why???

WAGENINGEN
UNIVERSITY & RESEARCH

* In R with mvabund::anova.traitglm
** In R with ade4::fourthcorner

An illustrative example

ter Braak et al 2017, PeerJ, p.9/10

Which one cannot be trusted and why???

The slopes wrt snowmelt date are species-specific (GLMM model)

There is a second ('latent') trait (z = SLA*) that has

- **about zero correlation (0.02) with Spread and**
- **interacts with snowmelt date (p_{max} <0.001)**

There is thus no real evidence for Spread ↔Snowmelt date.

WAGENINGEN
UNIVERSITY & RESEARCH

*Specific Leaf Area

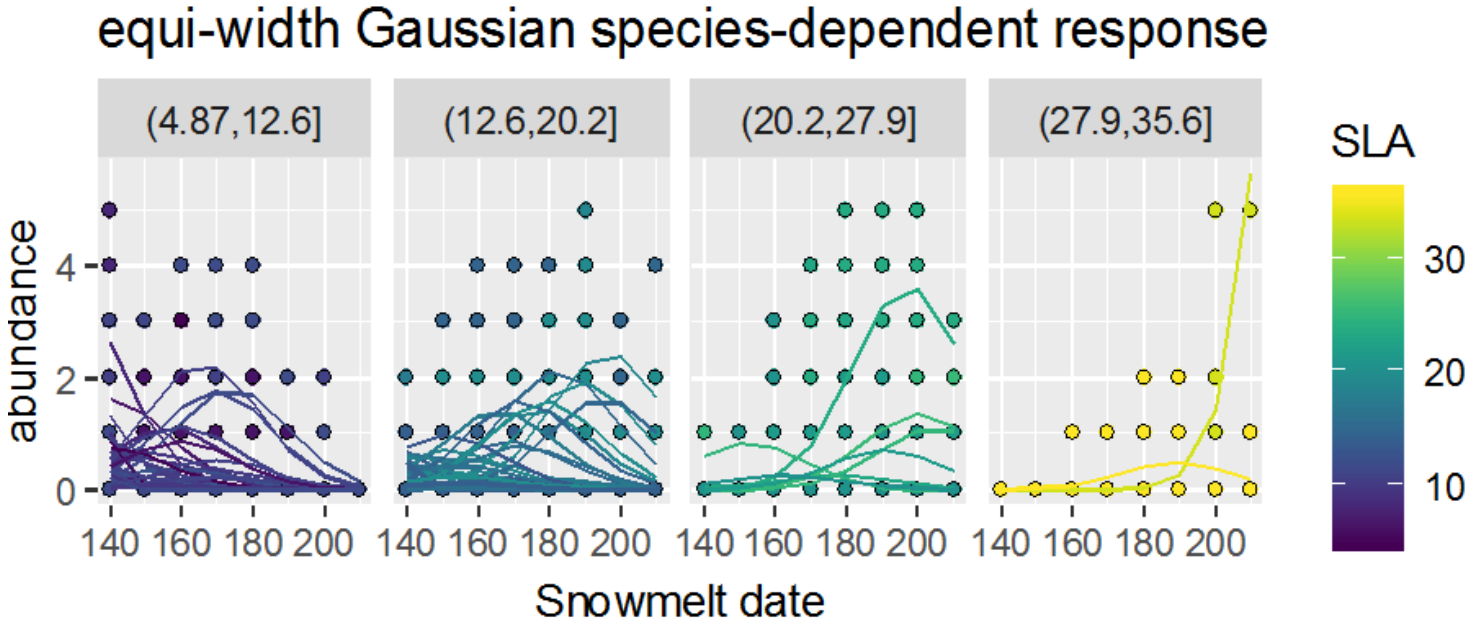
Same story with non-linear main effects....

- The real world looks like a GLMM model

```
y~ poly(env,2) + (1 + env|species)
```

  - there is (random) **species-specific response** wrt to e

= there is a **latent uncorrelated trait** interacting with e



The simulations had quadratic main effects, but that is not the issue here...

GLM deviance versus fourth-corner correlation

- In the simulations, I also investigated a simpler test statistic than deviance:

  the squared fourth-corner correlation

- Surprise, surprise..... (is it?)

- fourth-corner correlation gave similar type I error and power as the GLM deviance!!

- How does this come about? So, is there perhaps a nice property of the fourth-corner correlation that I did not know about?

GLM and fourth corner correlation *f*

ter Braak EEST 2017

GLM model: count y_{ij} follows a Poisson distribution with mean specified by

- $\log(\mu_{ij}) = r_i + c_j + \beta_{te}\, t_j e_i$ **(1)**

$f^2 y_{++}$ = squared fourth corner correlation × y_{++}

= Rao score test statistic

for testing the linear-by-linear interaction H_0: $\beta_{te} = 0$

Asymp. equivalent with LR, much quicker to compute!

Extension to multiple traits and environmental variables:

Score test statistic = y_{++} × inertia of dc-CA

dc-CA = double constrained correspondence analysis

Bacou & Sabatier 1989, Lavorel & Lebreton 1998/9

Böckenholt & Böckenholt 1990, Takane 2013

WAGE UNIVERSITY

Corollary

- T = I_m (= no constraints on columns) gives single constrained correspondence analysis which is canonical correspondence analysis (CCA, ter Braak 1986)
  - Total inertia of CCA × y_{++} = Rao's score test statistic

Used as test statistic in permutation testing since 1990
in Canoco and later in R::vegan

So, we discovered a new property of a much used method!

The result gives a reason for renewed interest in dc-CA

From fourth corner correlation to dc-CA

ter Braak et al EEST 2018

- fourth-corner correlation f between trait t and environmental variable e

$$f = cor_{\mathbf{Y}(\mathbf{t},\mathbf{e})} = \frac{\sum_{i,j} y_{ij} \tilde{t}_j \tilde{e}_i}{\left\{\sum_j y_{+j} \tilde{t}_j^2 \ \sum_i y_{i+} \tilde{e}_i^2\right\}^{1/2}} \quad (1)$$

with

$$\tilde{t}_j = t_j - \sum_j y_{+j} \, t_j / y_{++} \text{ and } \tilde{e}_i = e_i - \sum_i y_{i+} e_i / y_{++} \quad (2)$$

- **Definition:**

dc-CA is a method that finds linear combinations of traits and of environmental variables that maximize their fourth corner correlation

WAGENINGEN
UNIVERSITY & RESEARCH

Derivation of dc-CA

Assume traits and environmental variables are centered

$$\mathbf{1}_n^T \mathbf{R}\mathbf{E} = \mathbf{0}_p \text{ and } \mathbf{1}_m^T \mathbf{K}\mathbf{T} = \mathbf{0}_q$$

with R = diag($\{y_{i+}\}$)and K= diag($\{y_{+j}\}$).

The definition of dc-CA leads to the following maximization problem:

$max_{\mathbf{b},\mathbf{c}}\ \mathbf{x}^T\mathbf{Y}\mathbf{u}$ **with** $\mathbf{x} = \mathbf{E}\mathbf{b}$, $\mathbf{u} = \mathbf{T}\mathbf{c}$, $\mathbf{x}^T\mathbf{R}\mathbf{x} = 1$ **and** $\mathbf{u}^T\mathbf{K}\mathbf{u} = 1$ **(3)**

or

$max_{\mathbf{b},\mathbf{c}}\ \mathbf{b}^T\mathbf{E}^T\mathbf{Y}\mathbf{T}\mathbf{c}$ **subject to** $\mathbf{b}^T\mathbf{E}^T\mathbf{R}\mathbf{E}\mathbf{b} = 1$ **and** $\mathbf{c}^T\mathbf{T}^T\mathbf{K}\mathbf{T}\mathbf{c} = 1$. **(4)**

Lagrange multiplier method leads to

$$\lambda_b\mathbf{b} = (\mathbf{E}^T\mathbf{R}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{Y}\mathbf{T}\mathbf{c} \quad (6)$$

$$\lambda_c\mathbf{c} = (\mathbf{T}^T\mathbf{K}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{Y}^T\mathbf{E}\mathbf{b} \quad (7)$$

$$\rightarrow \lambda(\mathbf{E}^T\mathbf{R}\mathbf{E})\mathbf{b} = \mathbf{E}^T\mathbf{Y}\mathbf{T}(\mathbf{T}^T\mathbf{K}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{Y}^T\mathbf{E}\mathbf{b} \quad (8)$$

→dc-CA is weighted canonical correlation

Transition formulae of dc-CA

1. $\lambda^{\alpha} u_k^* = \sum_i y_{ik} x_i / y_{+k}$ or in matrix notation, $\lambda^{\alpha}\mathbf{u}^* = \mathbf{K}^{-1}\mathbf{Y}^T\mathbf{x}$
2. $\mathbf{c} = (\mathbf{T}^T\mathbf{K}\mathbf{T})^{-1}\mathbf{T}^T\mathbf{K}\mathbf{u}^*$
3. $\mathbf{u} = \mathbf{T}\mathbf{c}$
4. $\lambda^{1-\alpha} x_i^* = \sum_k y_{ik} u_k / y_{i+}$ or in matrix notation, $\lambda^{1-\alpha}\mathbf{x}^* = \mathbf{R}^{-1}\mathbf{Y}\mathbf{u}$
5. $\mathbf{b} = (\mathbf{E}^T\mathbf{R}\mathbf{E})^{-1}\mathbf{E}^T\mathbf{R}\mathbf{x}^*$
6. $\mathbf{x} = \mathbf{E}\mathbf{b}$

λ = eigenvalue, c and b are canonical weights, $\alpha \epsilon [0,1]$ user-defined.

Two sets of row scores $\{x_i\}$ and $\{x_i^*\}$ & columns scores, $\{u_k\}$ and $\{u_k^*\}$

1&4 $\rightarrow$CA with $\{u_k^* = u_k\}$ and $\{x_i^* = x_i\}$ or $\{\mathbf{E} = \mathbf{I}_n \ , \ \mathbf{T} = \mathbf{I}_m\}$

1,4,5&6 $\rightarrow$CCA with $\{u_k^* = u_k\}$ or $\mathbf{T} = \mathbf{I}_m$

WAGENINGEN
UNIVERSITY & RESEARCH

iterative algorithm based on this:
power algorithm, slow but can be accelerated

And is this all a surprise? Hmm...

- T = I_m, E = I_n gives (unconstrained) correspondence analysis (CA)
  - Total inertia of CA × y_{++} = $y_{++} \sum_a \lambda_a = \chi^2$
  - which is a Rao score test statistic on row-column independence
- T = **t**, E = **e** gives the simplest case of dc-CA with

$$\lambda_1 = [cor_Y(\mathbf{e}, \mathbf{t})]^2 = f^2$$

Recall an original definition of CA (Hirshfield 1935*, Fisher 1940)**

- CA finds a **latent e*** and **latent t*** such that

$$\lambda_1 = max_{\{x,u\}}[cor_Y(x, u)]^2 = [cor_Y(\mathbf{e}^*, \mathbf{t}^*)]^2 = \max f^2$$

with $\mathbf{e}^*, \mathbf{t}^*$ row- and column scores of CA

→the maximum attainable squared fourth-corner correlation is thus the first CA-eigenvalue!

WAGENINGEN UNIVERSITY & RESEARCH

* Side-effect of simultaneous linear regressions
** Side-effect of discriminant analysis

History of correspondence analysis (CA)

- **CA:** Hirschfield 1935, Fisher 1940, Guttman 1941, Benzecri 1969, Hill 1974, Greenacre, 1984, Gifi 1990 and many others..
- **Single constrained CA (CCA):** ter Braak 1986/7, Chessel, Lebreton et al 1987/8, with a precursor: Green 1971!
- **Double constrained CA:** Bacou & Sabatier 1989, Lavorel & Lebreton 1998/9, Böckenholt & Böckenholt 1990, Takane 2013

Many different rationales! Relations to PCA, contingency tables, analysis of variance, log-linear models, unfolding, gradient analysis, Gaussian response models,...

- All are special cases of canonical correlation analysis (or of discriminant analysis, except dc-CA)
- But... it is nontrivial to do the computing via a program for canonical correlation analysis ...so Algorithms for...

Algorithm based on a SVD

Similar to canonical correlation. Define

$$\mathbf{D} = (\mathbf{E}^T\mathbf{R}\mathbf{E})^{-1/2}\mathbf{E}^T\mathbf{Y}\mathbf{T}(\mathbf{T}^T\mathbf{K}\mathbf{T})^{-1/2}$$

SVD of $\mathbf{D}$:

$$\mathbf{D} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathrm{T}}$$

with $\mathbf{P}$ and $\mathbf{Q}$ orthonormal matrices and $\boldsymbol{\Delta}$ a diagonal matrix with singular values in decreasing order.

Then the singular values are the maximized fourth corner correlations of the dc-CA axes and the columns of

$$\mathbf{B} = (\mathbf{E}^T\mathbf{R}\mathbf{E})^{-1/2}\mathbf{P}\boldsymbol{\Delta}^{\alpha} \text{ and } \mathbf{C} = (\mathbf{T}^T\mathbf{K}\mathbf{T})^{-1/2}\mathbf{Q}\boldsymbol{\Delta}^{\alpha-1}$$

satisfy the transition formulae.

$\mathbf{X} = \mathbf{E}\mathbf{B}$ and $\mathbf{U} = \mathbf{T}\mathbf{C}$, are R- and K-orthogonal.

The scaling factor $\boldsymbol{\Delta}^{\alpha}$ ensures that $\mathbf{X}^T\mathbf{R}\,\mathbf{X} = \boldsymbol{\Lambda}^{\alpha}$ and $\mathbf{U}^T\mathbf{K}\,\mathbf{U} = \boldsymbol{\Lambda}^{1-\alpha}$, where $\boldsymbol{\Lambda} = \boldsymbol{\Delta}^2$

WAGENINGEN UNIVERSITY & RESEARCH

$tr(\mathbf{D}^T\mathbf{D}) = \sum_a \lambda_a$ is the Rao score test statistic/y_{++}

Comparison with dc-PCA

Douglas Carrol et al 1980, two-way CANDELINC

A weighted dc-PCA can be obtained from an SVD of

$$\mathbf{D}_{\mathrm{dc-pca}} = (\mathbf{E}^T\mathbf{RE})^{-1/2}\mathbf{E}^T\mathbf{RYKT}(\mathbf{T}^T\mathbf{KT})^{-1/2}$$

Compare:

$$\mathbf{D}_{\mathbf{dc-ca}} = (\mathbf{E}^T\mathbf{RE})^{-1/2}\mathbf{E}^T\mathbf{YT}(\mathbf{T}^T\mathbf{KT})^{-1/2}$$

→ dc-CA is a weighted dc-PCA of the contingency ratios

$$y_{++}\mathbf{R}^{-1}\mathbf{YK}^{-1}$$

with weight matrices with $\mathbf{R}$ = diag($\{y_{i+}\}$) and $\mathbf{K}$= diag($\{y_{+j}\}$).

All very similar... dc-CA is a natural method for count-like data

WAGENINGEN
UNIVERSITY & RESEARCH

Comparison with RLQ (1) (the current standard in ecology)

Dolédec et al EEST 1996

An RLQ can be obtained from an SVD of

$$\mathbf{D_{rlq}} = \mathbf{E}^T\mathbf{YT}$$ where E and T are R- and K-standardized

Compare:

$$\mathbf{D_{dc-ca}} = (\mathbf{E}^T\mathbf{RE})^{-1/2}\mathbf{E}^T\mathbf{YT}(\mathbf{T}^T\mathbf{KT})^{-1/2}$$

→ dc-CA uses the correlations among traits & among environmental variables, whereas RLQ does not

→ RLQ is more robust to near-collinearity than dc-CA,

dc-CA needs regularization or variable selection to counter this

Another way of saying similar things:

→ dc-CA is based on correlation (based on regression)

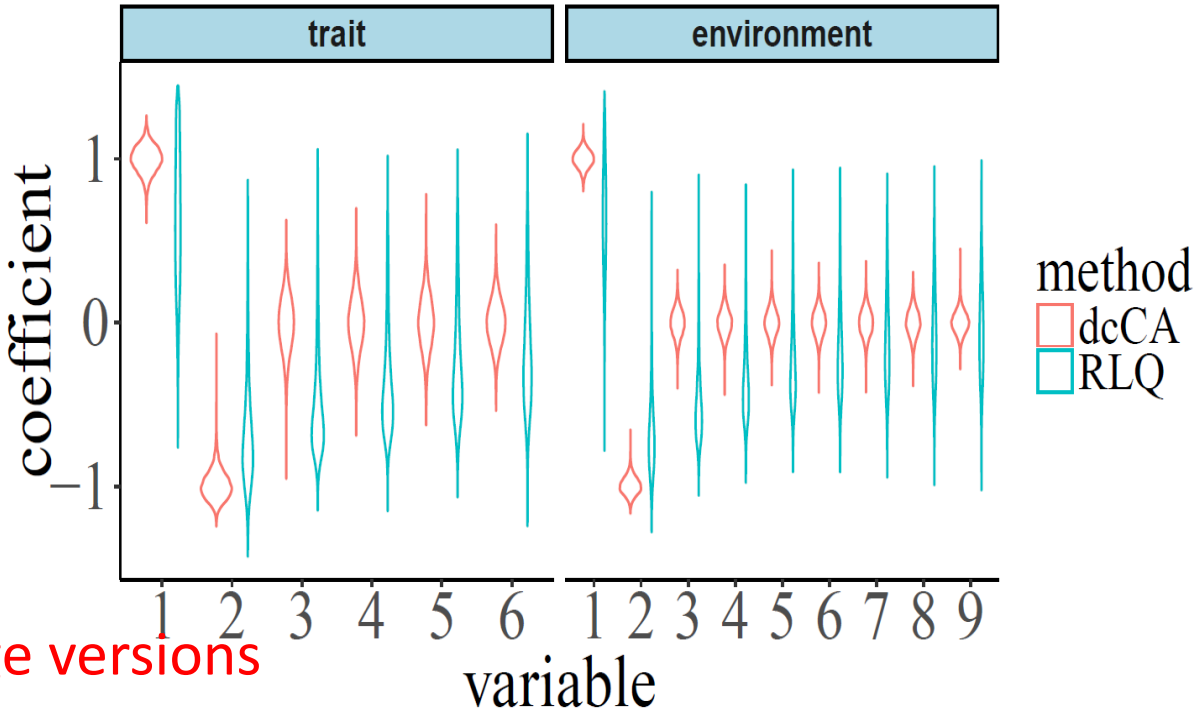
→ RLQ is based on covariance (based on coinertia analysis, a tiny bit like PLS)

Comparison with RLQ (2)

Because its regression base:

- **dc-CA can reveal trait and environment dimensions that remain hidden in RLQ**
  - **if trait and/or env. vars. are moderately correlated**
- **A simulation study, 10,000 simulated data sets with:**
  - ***n*=*m* = 100**
  - **6 traits, 9 environmental variables ~ $AR_1(0.7)$**
  - **One latent dimension defined by a contrast of the first two traits and the first two environmental variables; a second dimension unrelated to E,T.**
  - **So: 4 of the traits and 7 of the env. vars are noise**

dc-CA reveals the contrast, RLQ does not



PLS and ridge versions of dc-CA: useful bridges between dc-CA and RLQ.

Alternative: variable selection in dc-CA.

	ρ_1		ρ_2		λ_1/λ_2	
	dc-CA	RLQ	dc-CA	RLQ	dc-CA	RLQ
2.5%	0.15	0.03	0.04	0.02	8.0	0.2
50%	**0.21**	**0.08**	**0.05**	**0.03**	**18.2**	**7.4**
97.5%	0.28	0.15	0.06	0.08	39.2	45.6

WAGENINGEN UNIVERSITY & RESEARCH

Algorithm based on combining CCA and RDA

… gives insight in relations with another existing method, called CWM-RDA (combine two tables (Y & T), then use a two-table method):

1. Combine Y with T in a single table of trait means per site $\mathbf{M} = \mathbf{R}^{-1}\mathbf{YT}$ = a table with CWMs.
2. Analyse $\mathbf{M} \sim \mathbf{E}$ by redundancy analysis (RDA)

This is essentially an SVD of

$$\mathbf{D}_{\text{cwm-rda}} = (\mathbf{E}^T\mathbf{E})^{-1/2}\mathbf{E}^T\mathbf{M} = (\mathbf{E}^T\mathbf{E})^{-1/2}\mathbf{E}^T(\mathbf{R}^{-1}\mathbf{YT})$$

- Lacks R-weighing and trait covariances
- **Obtain dc-CA by adding R&K-weighing and a prior ortho-normalization of T**
- **Can be done by first performing a CCA and then a weighted RDA on its scores …**

Useful in Canoco as it has
testing and selection of variables
for (weighted) RDA

WAGENINGEN
UNIVERSITY & RESEARCH

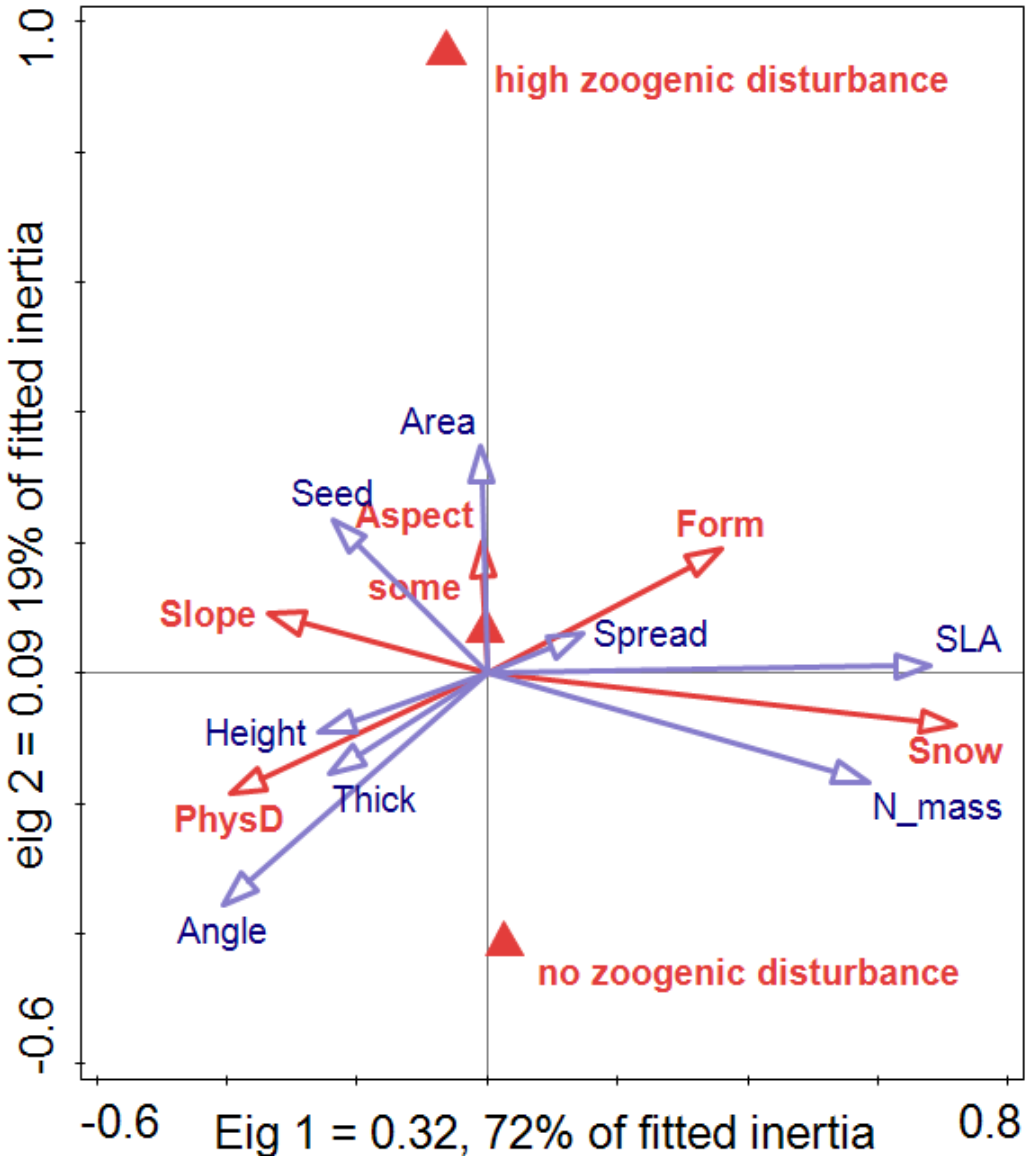
Illustrative example: snowmelt (aravo data)

- Biplot of fourth-corner correlations
- Describes 91% of fitted inertia
- Snow is highly associated with SLA; Spread is not.

- NB:

Eigenvalues

$$\lambda_{dc-CA} \leq \lambda_{CCA} \leq \lambda_{CA}$$



WAGENINGEN
UNIVERSITY & RESEARCH

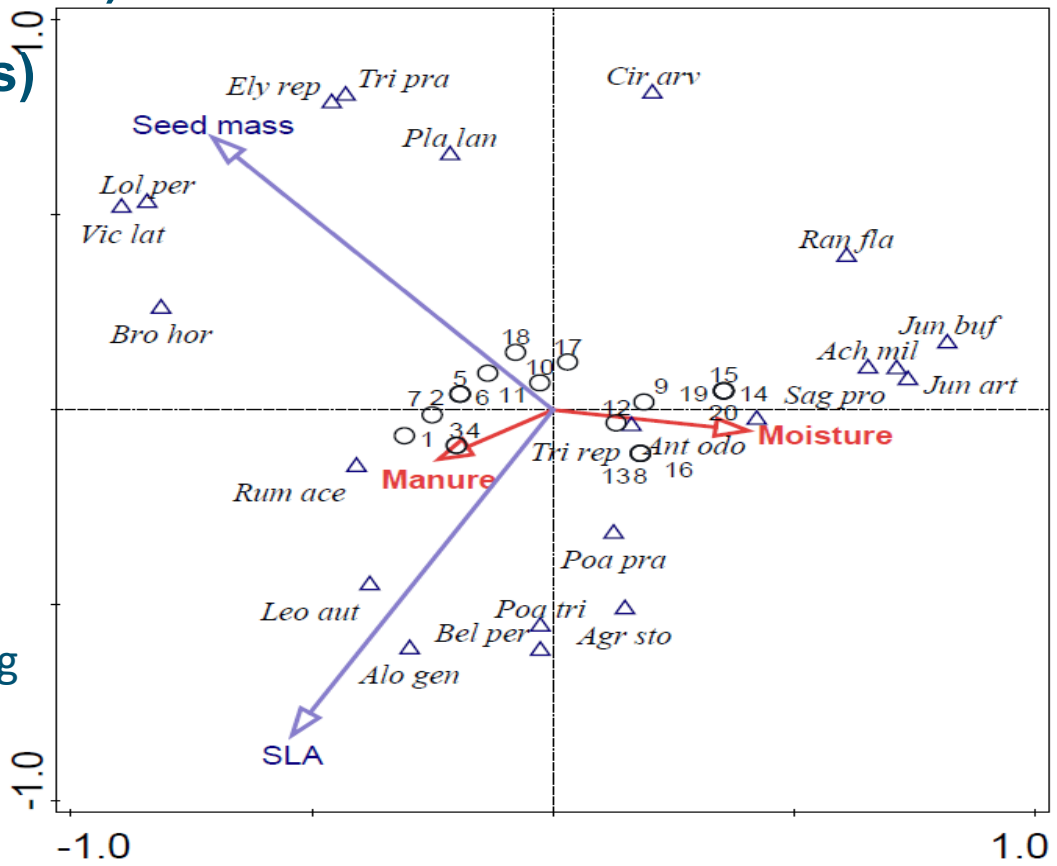
Quadriplot of dc-CA: example

5 out of 6 pairs are weighted least-squares biplots of:

1. **Fourth-corner correlations: $\mathbf{E}^T\mathbf{YT}$**
2. **E means per species (SNCs)**
3. **T means per site (CWMs)**
4. **Contingency ratios**
5. **Trait data* T**

Dune meadow data:
n= 20, *m* = 28
two traits
two environmental variables

* In column-metric preserving scaling and with fixed species points



WAGENINGEN
UNIVERSITY & RESEARCH

Comparison of fourth-corner and GLMM models

Miller, Damschen & Ives 2018 Method in Ecology and Evolution

Method	Advantages	Disadvantages
Community-weighted mean regression **CWM only**	1. None	1. Inflated type I error rates yield unreliable results
Weighted correlations **FC CWM/SNC**	1. Give correct type I error	1. The primary information given is only *p*-values
	2. Simple and fast	2. Limited to simple analyses
Model-based approaches **GLMM**	1. Can give highest power	1. Complicated and computationally intensive
	2. Give most information about the data	2. Require diagnostics and possibly parametric bootstraps
	3. Flexible for multivariate analyses	

Whittaker Revisit data (Siskiyou Mountains)

Example data in Miller, Damschen & Ives 2018

Is there association between

- **Functional trait: leaf Carbon-to-Nitrogen ratio (C:N) and**
- **Topographic Moisture Gradient (TMG)?**

Miller et al. found little evidence using the 'good' methods

	P-value
Fourth corner (FC)	**0.059**
GLMM 1 (Wald)	0.47
GLMM 2* (Wald)	**0.088**
GLMM 2 (boot)	0.012
traitglm (mvabund)	0.27

Unweighted	**P-value**
CWM/SNC	**0.034**
Peres-Neto: transform Y	
FC $Y^{0.25}$	**0.031**

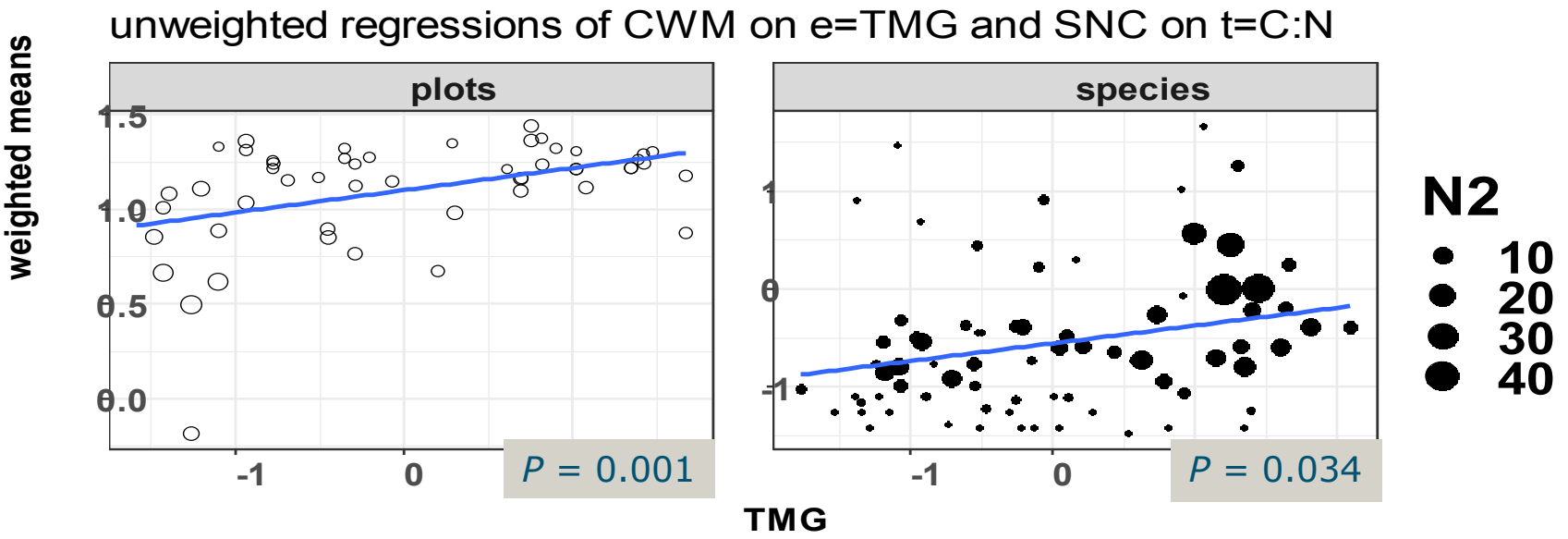
But is there a principled way to choose the transformation?

* Jamil et al. 2013

WAGENINGEN UNIVERSITY & RESEARCH

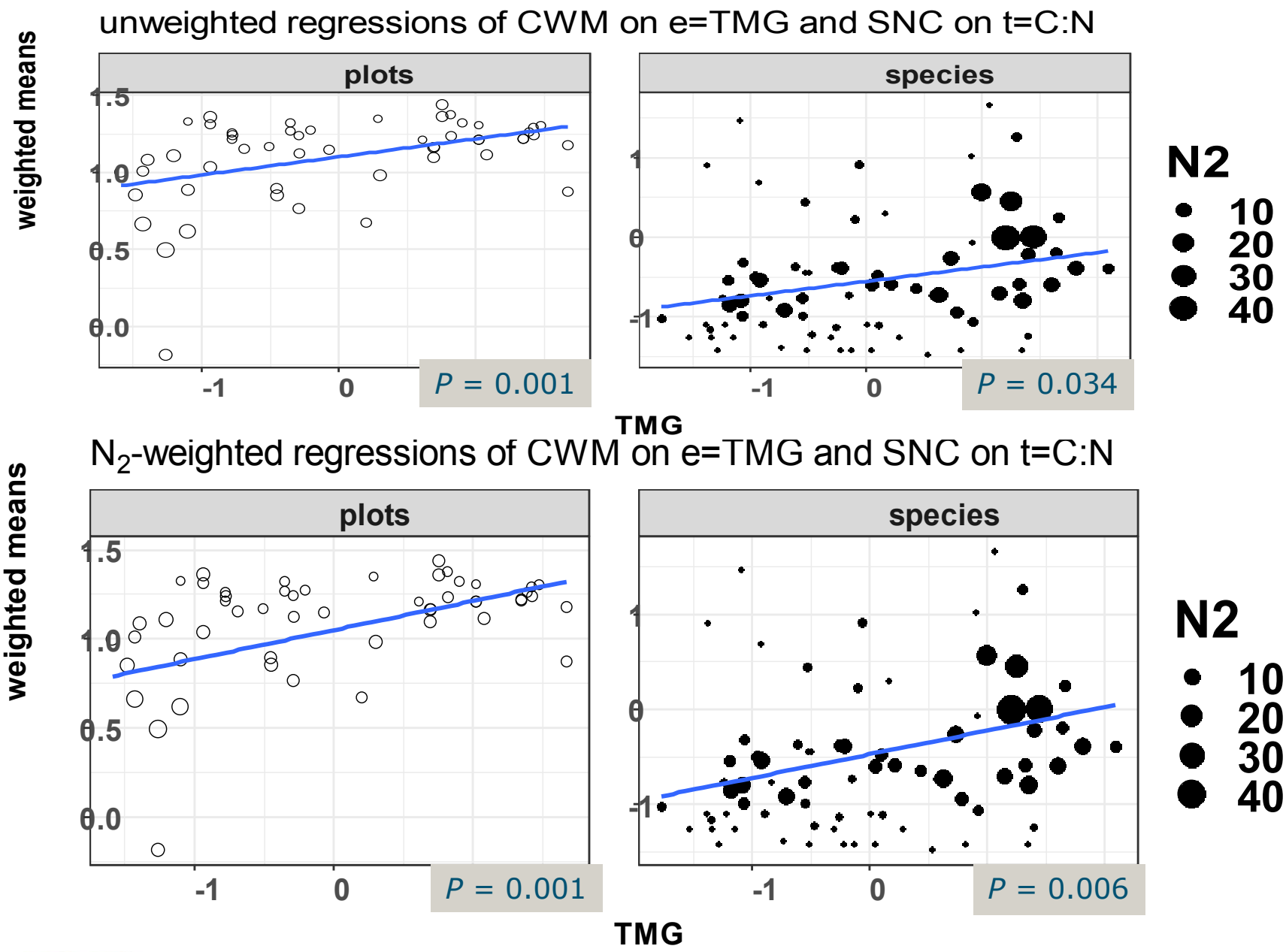
Weighting in fourth-corner & CWM/SNC

- **Permutation testing of FC is the same as**
  1. **Site permutation testing CWM ↔ e (weight = y_{i+})**
  2. **Species permutation testing SNC ↔ t (weight = y_{+j})**
- **Why not try unweighted regressions? ter Braak et al 2018**



N_2 = effective number of occurrences (Hill 1973)

From unweighted to N_2-weighted CWM/SNC



Better GLMM model

- **GLMM 2** is asymmetric in species and sites

```
y ~ trait*env +(1+env|species)+ (1|site)
```

- It does account for
  - species-specific response wrt e

  = latent trait interacting with e
- But not for
  - site-specific effects of the trait

  = latent environmental variable interacting with t

GLMM 3 is symmetric in species and sites

```
y ~ trait*env +(1+env|species)+ (1+trait|site)
                    .... + trait^2 + env^2
```

to allow for simple unimodal response

WAGENINGEN
UNIVERSITY & RESEARCH

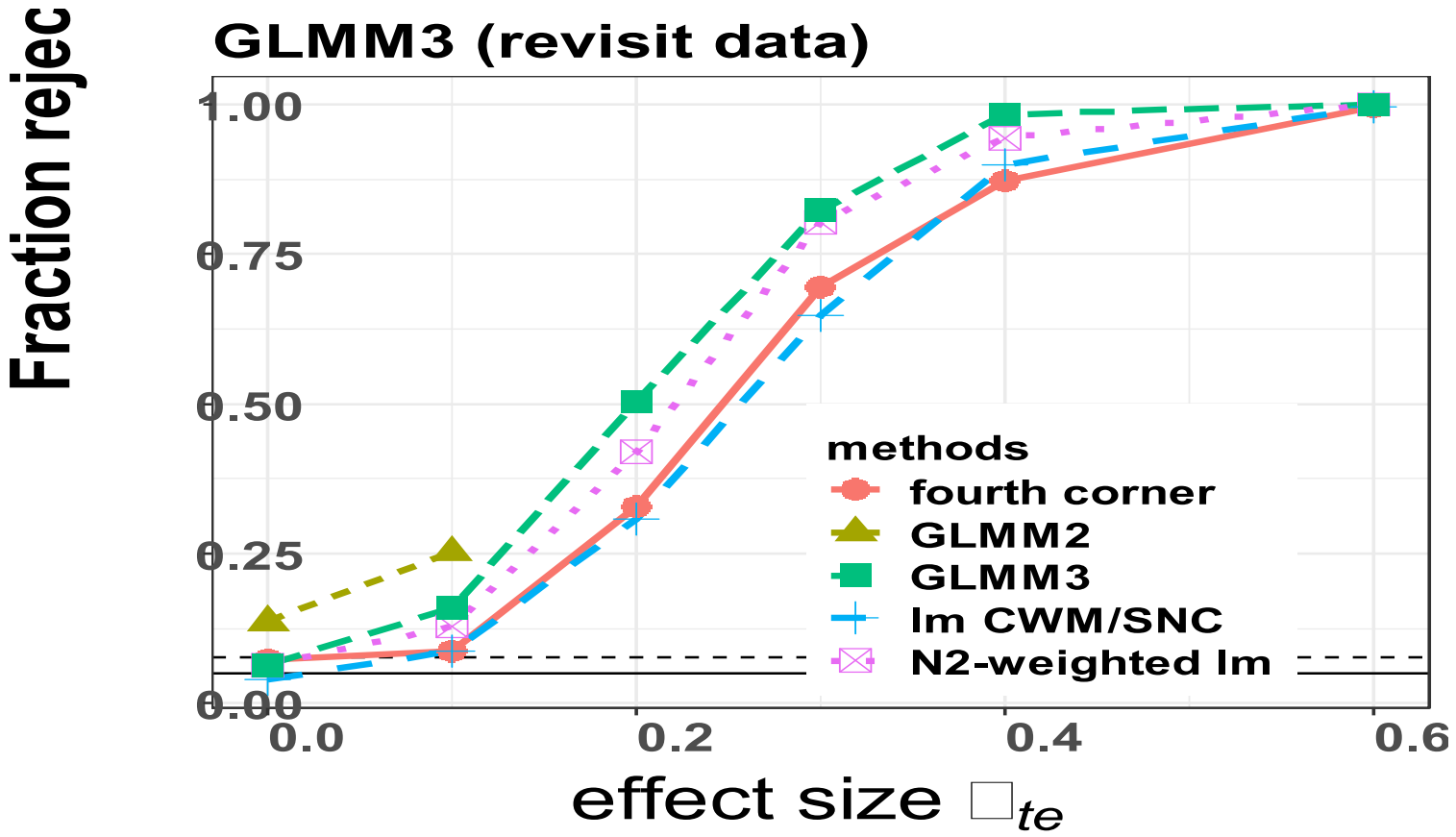
Whittaker Revisit data: evidence for interaction

		P-value
GLMM2		0.088
FC	$Y^{0.25}$	0.031
N2-weighted C/S		0.006
GLMM3		0.014

- Based on the new model, the conclusion changes from weak evidence to strong evidence for TMG-C:N interaction

Power simulations

- Simulation from fitted GLMM3 model:



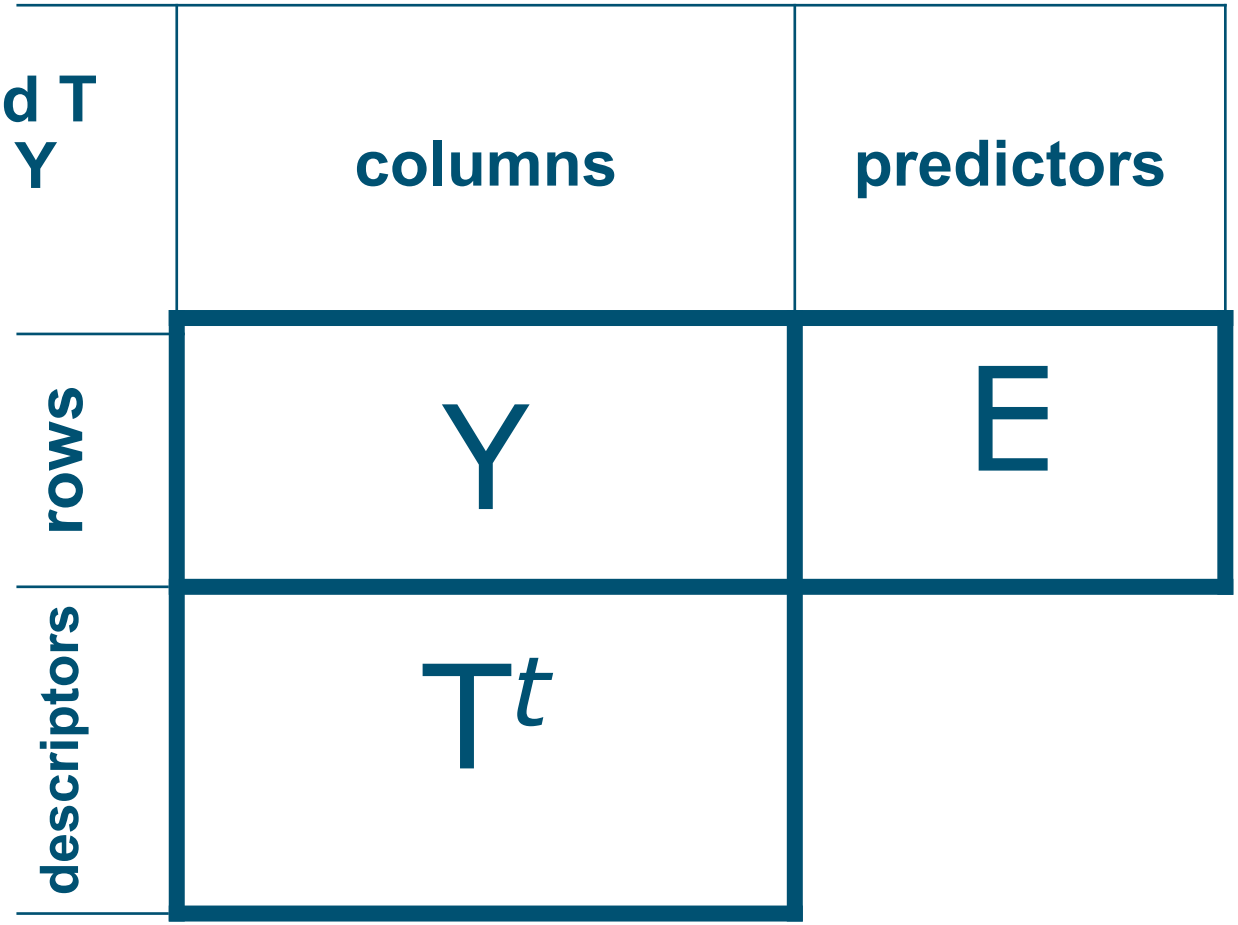
- Power and type I error control of GLMM3 and N_2-weighted CWM/SNC similar

Concluding remarks

- Statistical issues
  - Sites, species and abundance values are random
  - Needs a GLMM that is symmetric in sites and species
  - Simpler models (GLM, fourth-corner, dc-CA) need
    - combination of site- and species-resampling as "the noise in the rows is likely different from that in the columns"
- Fourth-corner and dc-CA
  - provide Rao score test statistics of the simple GLM models that are useful in resampling
- dc-CA allows easy testing and variable selection scheme
  - combining site- and species-analyses (Canoco 5.10)

L-shaped data (Γ–shaped data): not only in ecology

- Central matrix (Y≥0,) with associated descriptors for rows and columns (E and T)
- Descriptors E and T as predictors for Y



Examples of central table Y

Data on:

- Abundance of species in sites in ecology
  - Which traits (T) of species determine in which type of environments (E, sites by variables) they prosper
  - Trait-based ecology, trait-environment relationships
- Preference of consumers for products
  - Which consumer characteristics and product features can predict the preference
  - consumer segments, niche markets, niche products
- Supervisory board memberships of firms
  - Which person characteristics determine which type of firm they supervise?

Some references

Jamil, T., W. A. Ozinga, M. Kleyer, and C. J. F. ter Braak. 2013. Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. Journal of Vegetation Science 24:988-1000. http://dx.doi.org/10.1111/j.1654-1103.2012.12036.x

Peres-Neto, P. R., S. Dray, and C. J. F. ter Braak. 2017. Linking trait variation to the environment: critical issues with community-weighted mean correlation resolved by the fourth-corner approach. Ecography 40:806-816. http://dx.doi.org/10.1111/ecog.02302

ter Braak, C. J. F. 2017. Fourth-corner correlation is a score test statistic in a log-linear trait–environment model that is useful in permutation testing. Environmental and Ecological Statistics 24:219-242. http://dx.doi.org/10.1007/s10651-017-0368-0

ter Braak, C. J. F., P. Peres-Neto, and S. Dray. 2017. A critical issue in model-based inference for studying trait-based community assembly and a solution. PeerJ 5:e2885. https://doi.org/10.7717/peerj.2885

ter Braak, C. J. F., P. Šmilauer, and S. Dray. 2018. Algorithms and biplots for double constrained correspondence analysis. Environmental and Ecological Statistics. https://doi.org/10.1007/s10651-017-0395-x or http://rdcu.be/ETPh

ter Braak, C.J.F. (2018) New robust weighted averaging- and model-based methods for assessing trait-environment relationships. PeerJ Preprints, 6, e27439v27431 https://t.co/Ln6gOMCUFZ.

WAGENINGEN
UNIVERSITY & RESEARCH

Thank you!

See also

www.Canoco.com

www.Canoco5.com

WAGENINGEN
UNIVERSITY & RESEARCH