Hiearchically modelling species responses

with functional traits and phylogeny

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline

- Fourth corner
- Phylogenetic

Questions so far?

The model so far

So far, we have studied species' environment responses (niches) with the model:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + \mathbf{z}_i^\top \mathbf{u}_j \tag{1}$$

where:

- $\mathbf{x}_i$ are the covariates
- $\boldsymbol{\beta}_j$ are the fixed effects
- $\mathbf{z}_i$ are also covariates
- $\mathbf{u}_j$ are random effects

This helps us understand **what environmental conditions species prefer**.

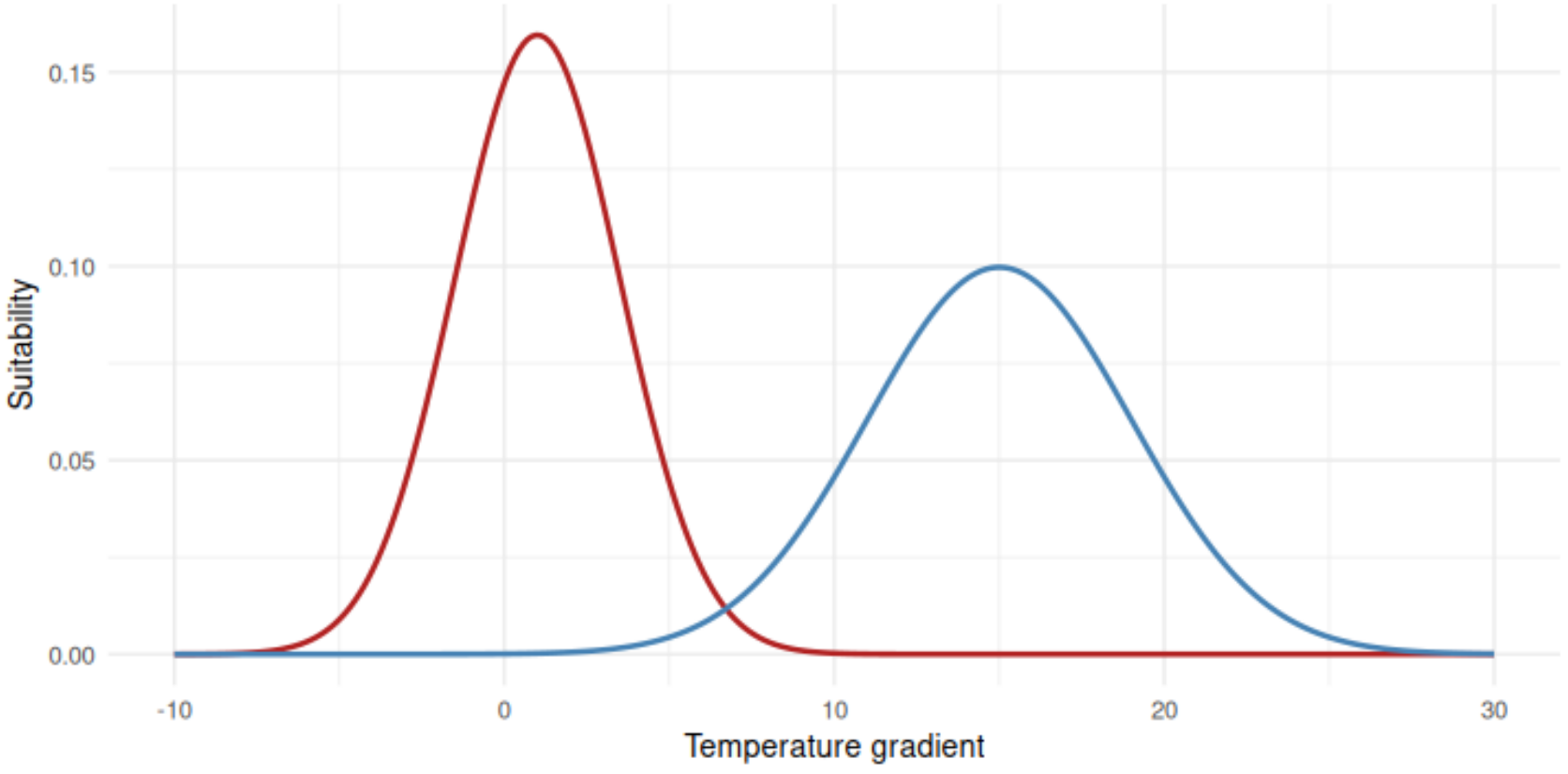
The more interesting question

Taking it a step further, we might want to answer instead:

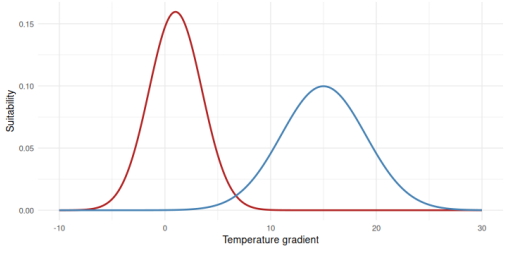
Why do species prefer certain environmental conditions?

But what makes species prefer certain environments?

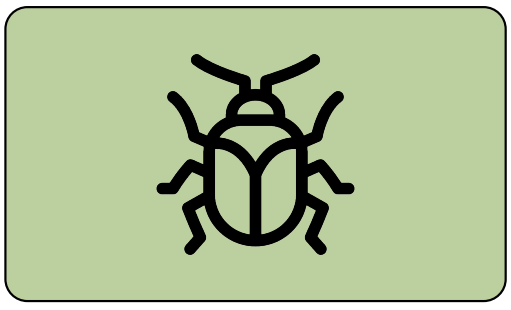
Trait-environment relationship

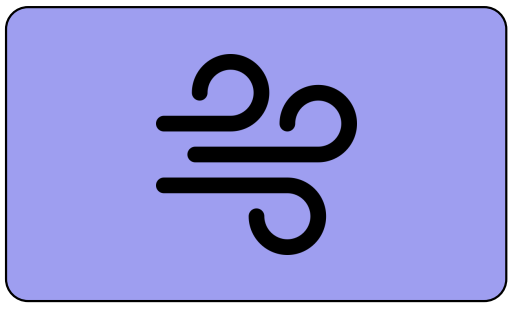


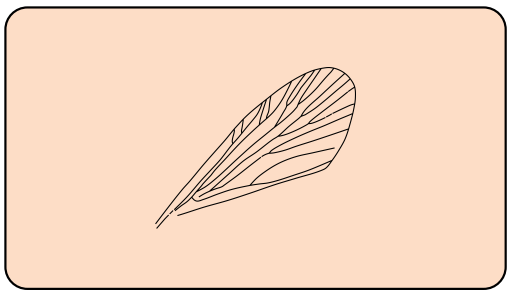
Trait-environment relationship



Fourth corner analysis

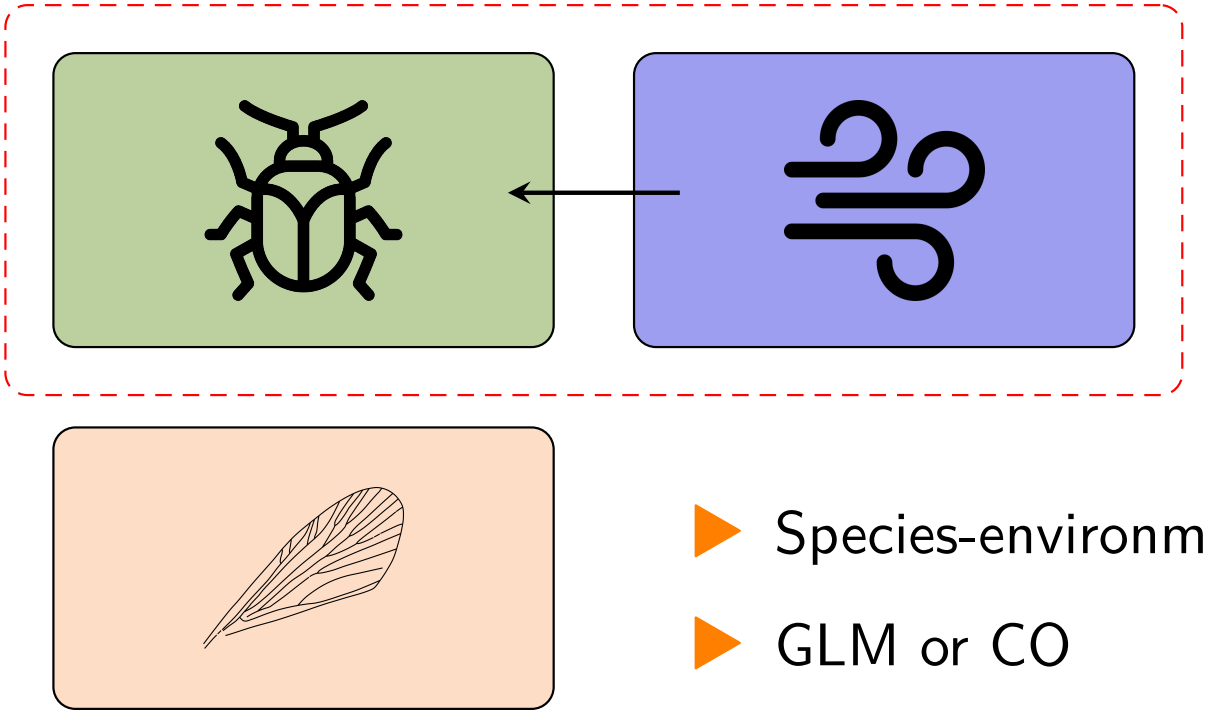






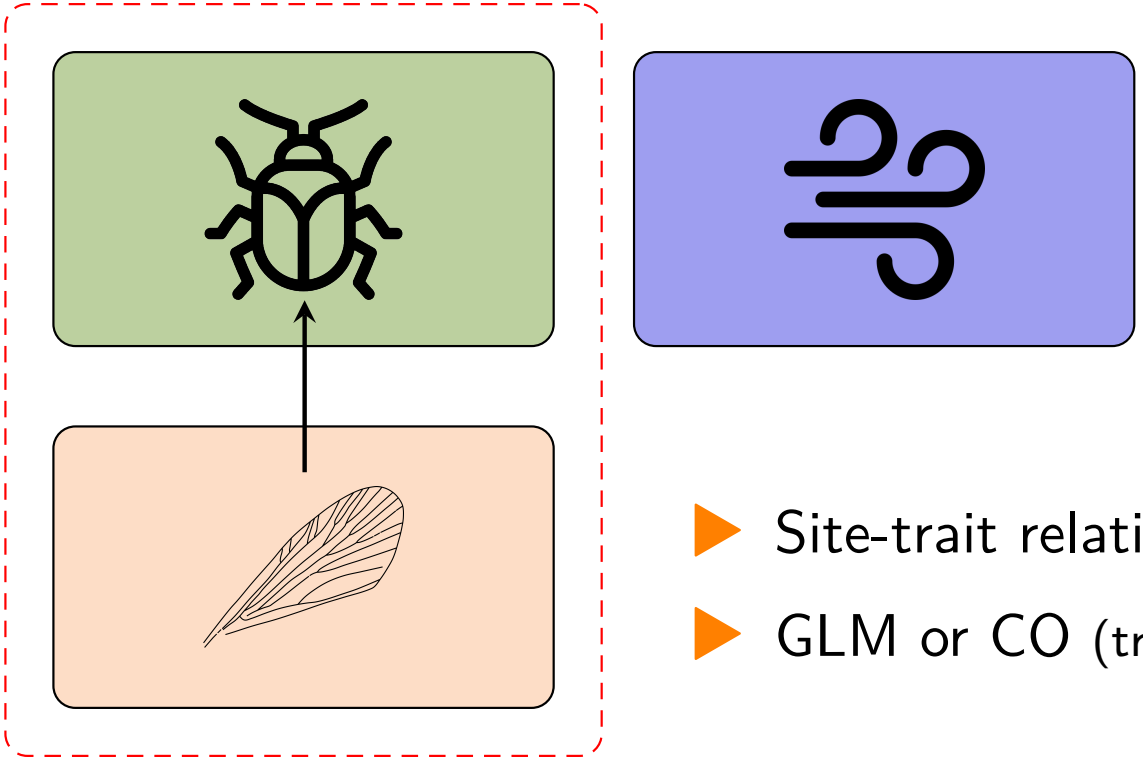
- **Y**: community data
- **X**: environmental variables
- **TR**: species traits

Fourth corner analysis



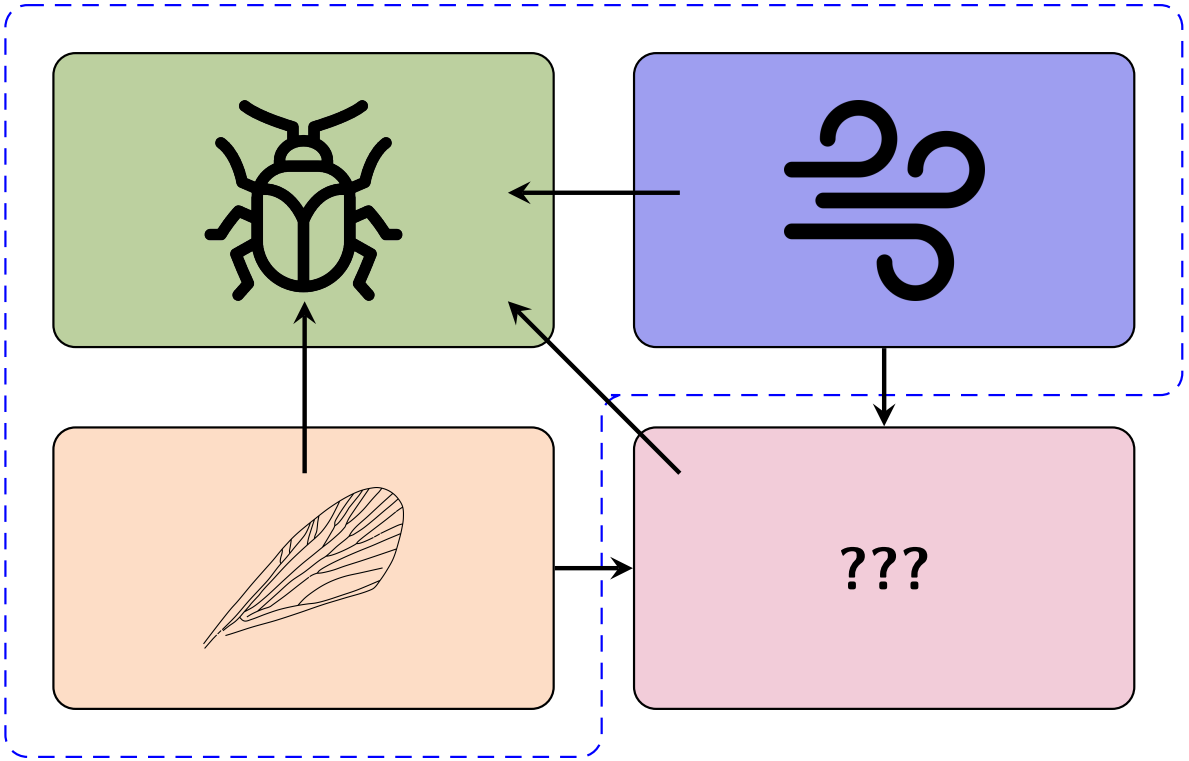
- Species-environment relationship
- GLM or CO

Fourth corner analysis



- Site-trait relationship
- GLM or CO (transpose the data)

Fourth corner analysis



History of the 4^{th} corner

also referred to as RLQ-analysis

For a long time, 4th-corner analysis was a recurring problem

- Traditionally solved by ordination Legendre et al. 1997, Dolédec et al. 1996)
- Environment-trait effects are tested with permutation (Dray and Legendre 2008, ter Braak et al. 2012)
- Statistical models can tackle the problem (Jamil et al. 2012, Brown et al. 2014)
- Dray et al. (2014) think the two complementary
- ter Braak et al. (2018) introduce a doubly constrained ordination method
- GLM-based methods have inflated error (ter Braak et al. 2017)
- Ovaskainen et al. (20177), Niku et al. (2021) combine with random effects (JSDM)

Often used trait-based analysis

1. RLQ Doledec et al. (1996)
2. CWM + RDA
3. Double constrained ordination
   Lebreton et al. (1988), ter Braak et al. (2018)
4. Fourth corner LVMs Ovaskainen at al. (2017), Niku et al. (2021)

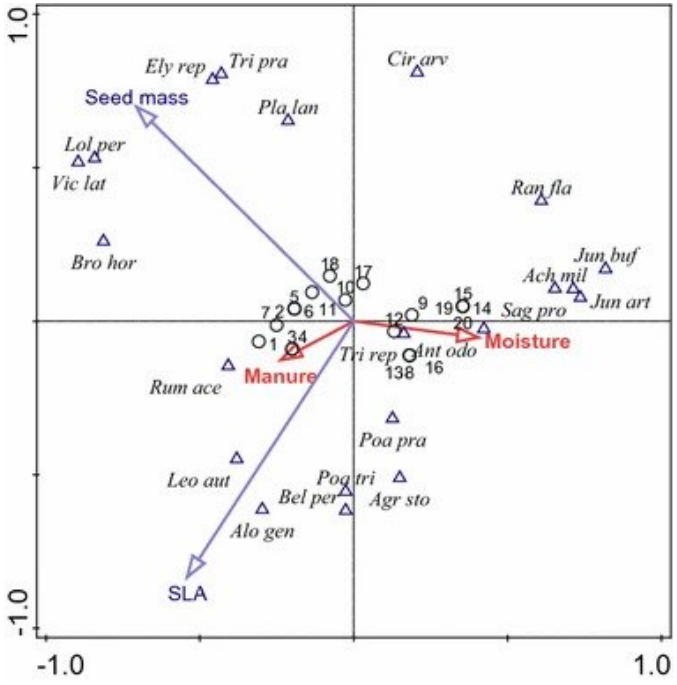


Figure 1: Quadriplot
ter Braak et al. (2018)

Intraspecific trait variation: snowshoe hare example

Intraspecific variation is ignored. Is there such a thing as "species traits"?

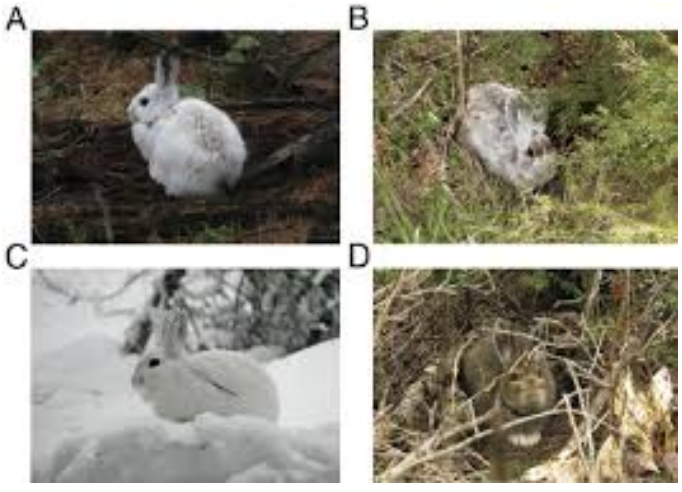


Figure 2: Mills et al. 2013

Intraspecific trait variation: snowshoe hare example

Intraspecific variation is ignored. Is there such a thing as "species traits"?

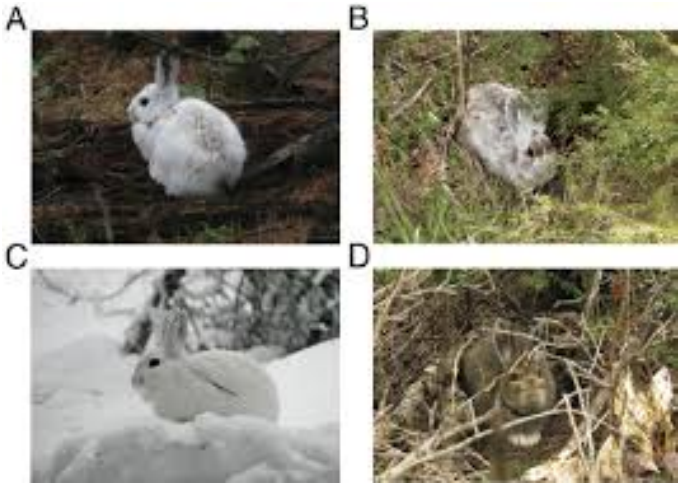


Figure 2: Mills et al. 2013

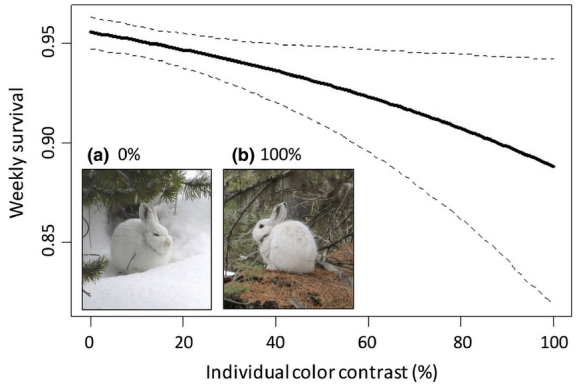


Figure 3: Zimova et al. 2016

Example: trait-environment GLM (no REs)

Traits can enter the model "just" as an interaction. That is the model we will explore here, first:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^\top \boldsymbol{\beta}_x + \mathbf{x}_i^\top \mathbf{B}_{x,tr} \mathbf{tr}_j \tag{2}$$

- $\boldsymbol{\beta}_x$ are the species-common effects
- $\mathbf{B}_{x,tr}$ are the environment-trait interaction coefficients

Example: alpine plants in France

- Data by Choler 2005
- Occurrence of 92 species at 75 5 by 5 plots
- 6 environmental variables: aspect, slope, microscale landform, disturbance level (physical and trampling/burrowing), and mean Julian snowmelt date
- 7 traits: height, spread, angle, area, thick, sla, N_mass, seed
- In the jSDM package

Example: Fitting a 4^{th}-corner VGLM

First recall, the VGLM:

```
model1 <- gllvm(Y, X,
                formula = ~Aspect + Slope,
                family = "binomial", num.lv = 0)
```

The model with traits:

```
model2 <- gllvm(Y, X, TR,
                formula = ~Aspect + Slope + (Aspect + Slope) : (Height + S
                family = "binomial", num.lv = 0)
```

Example: species responses

Compared to the (without traits) VGLM, species responses:

- have a common component: $\boldsymbol{\beta}_x$ (recall from yesterday)
- have a trait component: $\mathbf{tr}_j^\top \mathbf{B}_{x,tr}$

so we can write: $\boldsymbol{\beta}_j = \boldsymbol{\beta}_x + \mathbf{tr}_j^\top \mathbf{B}_{x,tr}$

Extracting the coefficients:

```
B <- (coef(model2, "B"))
Bx <- B[1:2]
Bxtr <- matrix(B[-c(1:2)], ncol = 2, byrow =  TRUE)
beta <- Bx + model2$TR%*%Bxtr
```

You can also get $\hat{\mathbf{B}}_{x,tr}$ via `gllvm::getFourthCorner(model2)`

Example: examining statistical significance

```
##
## Call:
## gllvm(y = Y, X = X, TR = TR, formula = ~Aspect + Slope + (Aspect +
##     Slope):(Height + SLA), family = "binomial", num.lv = 0)
##
## Family:  binomial
##
## AIC:  4391.6 AICc:  4394.192 BIC:  4974.37 LL:  -2108.8 df:  87
##
## Informed LVs:  0
## Constrained LVs:  0
## Unconstrained LVs:  0
##
## Formula:  ~Aspect+Slope+Aspect:Height+Aspect:SLA+Slope:Height+Slope:SLA
## LV formula:  ~ 0
## Row effect:  ~ 1
##
## Coefficients predictors:
##                Estimate Std. Error z value Pr(>|z|)
## Aspect        -0.025933   0.021351  -1.215   0.2245
## Slope          0.098612   0.021295   4.631 3.64e-06 ***
## Aspect:Height  0.006624   0.022010   0.301   0.7634
## Aspect:SLA     0.045142   0.023059   1.958   0.0503 .
## Slope:Height   0.040864   0.021589   1.893   0.0584 .
## Slope:SLA     -0.094442   0.023480  -4.022 5.76e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: fourth-corner interpretation

```
##      Aspect       Slope
## -0.02593288  0.09861163
```

On average Aspect (north, south, west, flat) decreases the probability occurrence and Slope increases it.

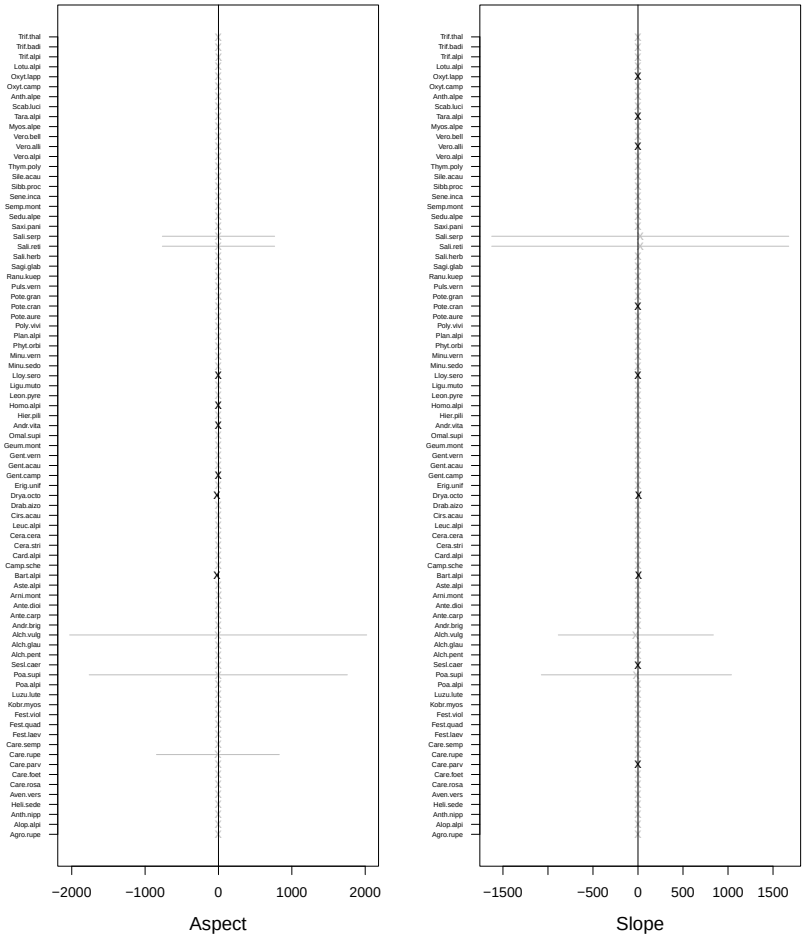
```
##                Height         SLA
## Aspect  0.006624329  0.04514243
## Slope   0.040864227 -0.09444214
```

High plants with thin leaves occur more in flat areas, and small plants with thick leaves least.
High plants with thick leaves occur more in steep places, while small plants with thin leavs occur more in flat places.

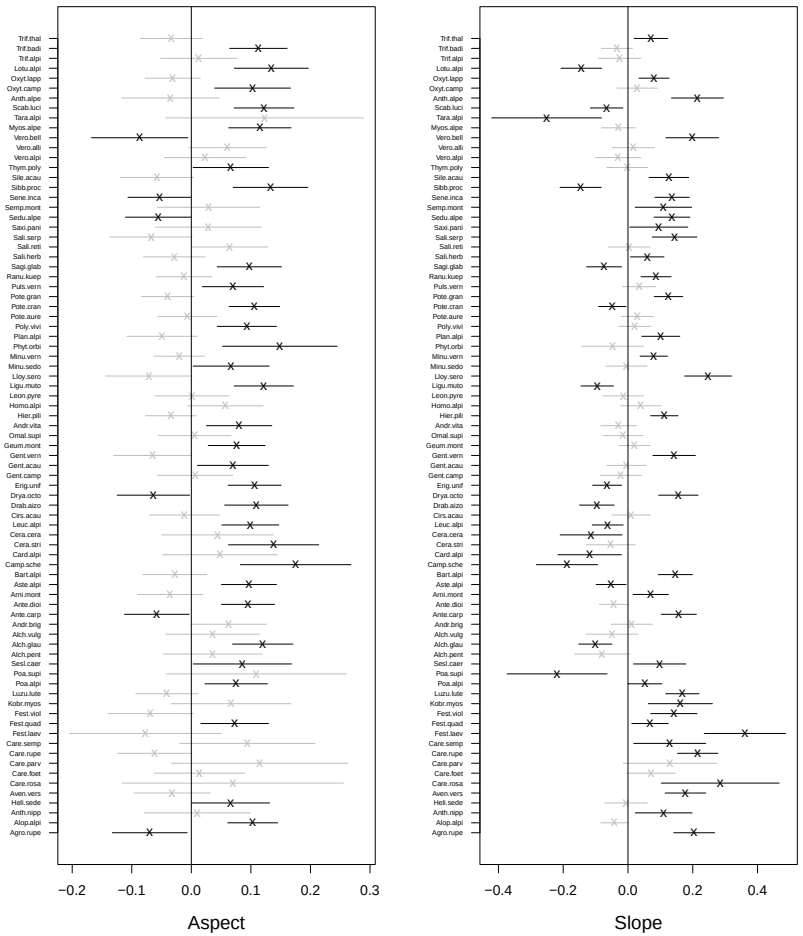
(but there is only enough evidence for slope and slope:SLA) to conclude an effect; thicker leaves in steep places

Example: species responses results (coefplot)

No traits



With traits



Example: comparison

```
anova(model1, model2)
```

```
## Warning in anova.gllvm(model1, model2): This test was not designed f
value should be treated as approximate.
```

```
## Model  1 :  ~ Aspect + Slope + (Aspect + Slope):(Height + SLA)
## Model  2 :  ~ Aspect + Slope
```

```
##   Resid.Df        D Df.diff     P.value
## 1     5907   0.0000       0
## 2     5751 234.8945     156 4.51301e-05
```

There are a lot more "free" parameters in the VGLM. The second model is either "constrained", or alternative; we provide is (much) more information.

Hierarchical responses

The model we implemented has one main limitation: it assumes species environmental responses are fully determined by the traits. But what if we measure the wrong traits?

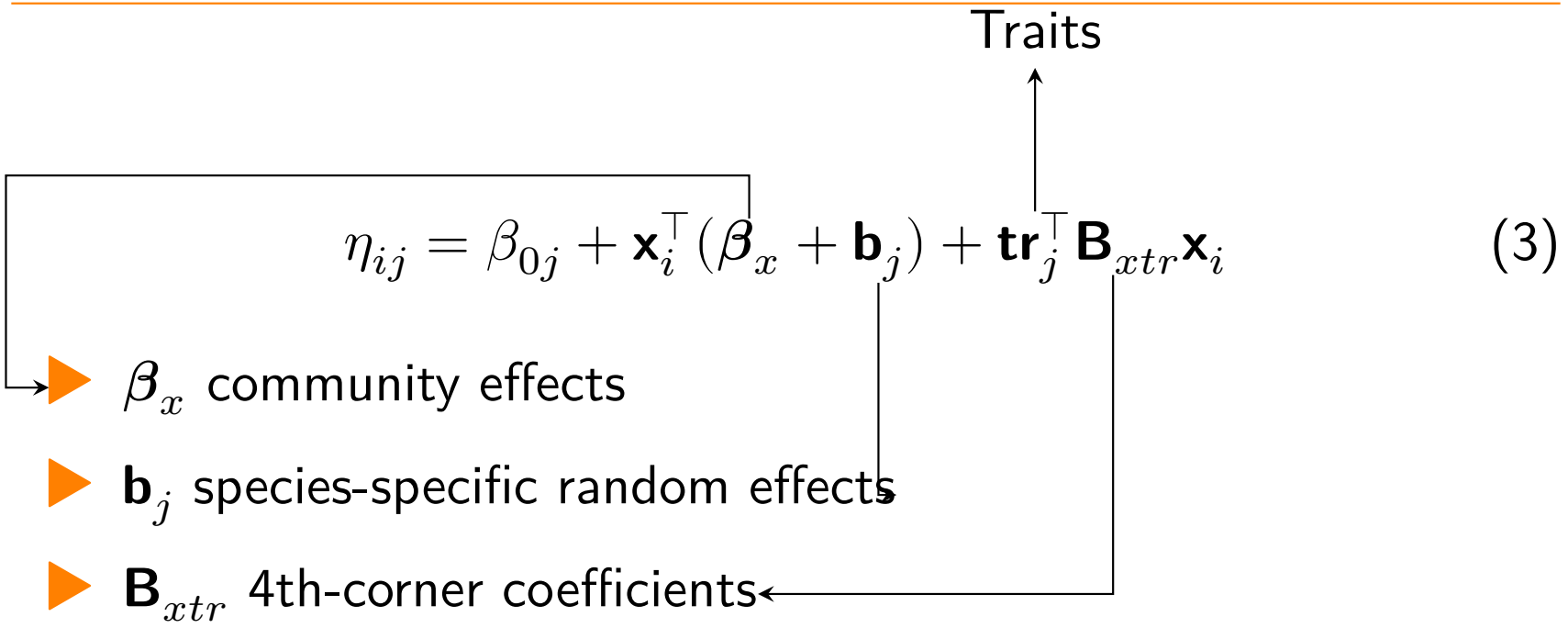
We can extend our model to incorporate "residual" information on species responses:

$$\boldsymbol{\beta}_j = \boldsymbol{\beta}_x + \mathbf{tr}_j^\top \mathbf{B}_{x,tr} + \mathbf{b}_j$$

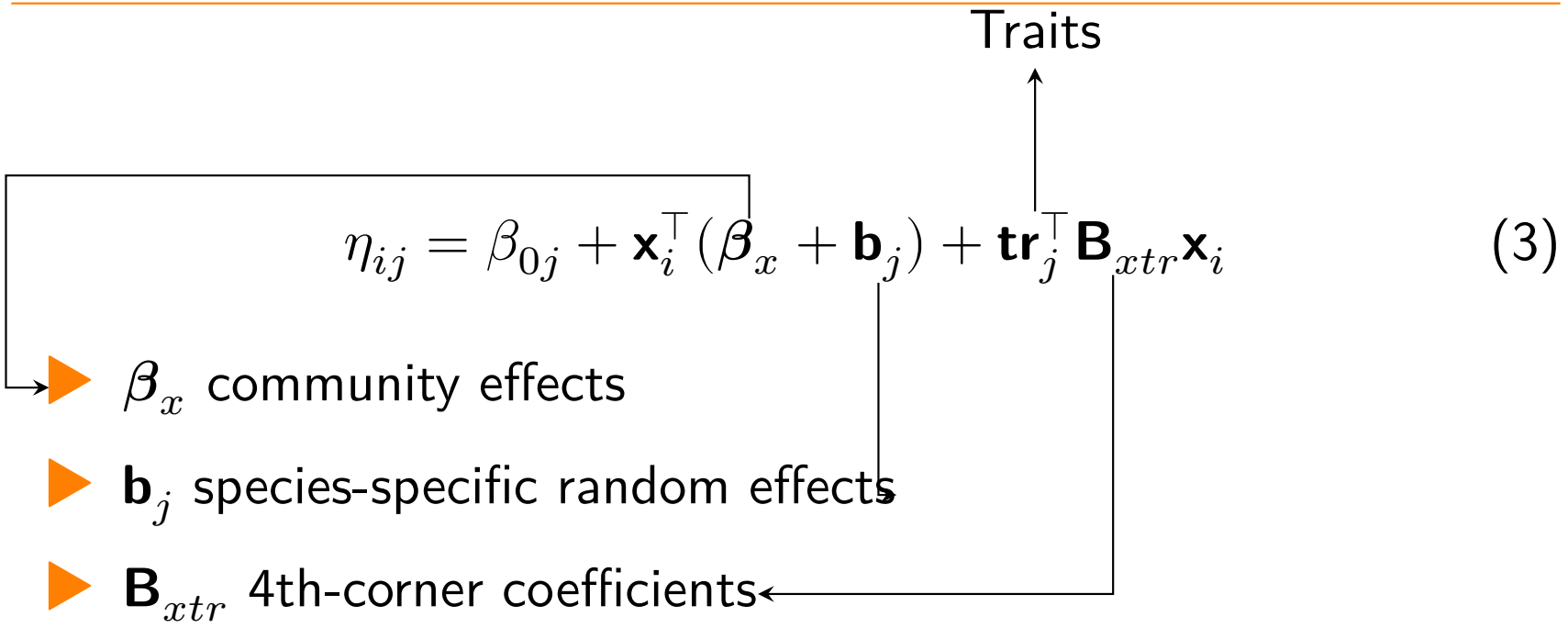
with $\mathbf{b}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_r)$ so that $\boldsymbol{\beta}_j \sim \mathcal{N}(\boldsymbol{\beta}_x + \mathbf{tr}_j^\top \mathbf{B}_{x,tr})$

this is much more similar to our VGLM, except that we now have a VGLMM with traits.

The full fourth corner model



The full fourth corner model



So, if $\boldsymbol{\Sigma} \approx \mathbf{0}$, there is no (excess) species-specific variation in responses.

Testing trait-environment interactions

A lot of attention has gone to testing environment-trait interactions.

Niku et al. (2021) concluded that omitting $\mathbf{b}_j$ leads to inflated Type I error (over optimistic conclusions w.r.t traits).

Testing trait-environment interactions

A lot of attention has gone to testing environment-trait interactions.

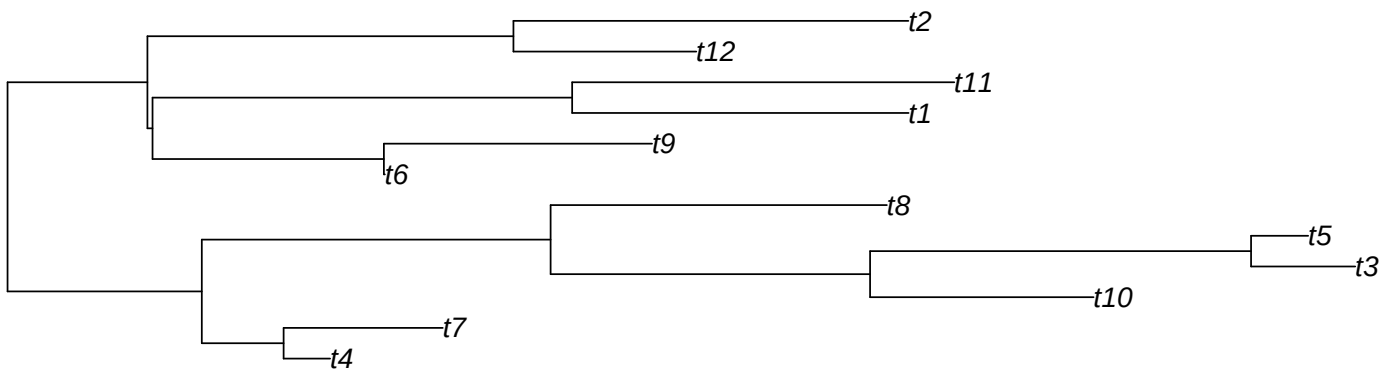
Niku et al. (2021) concluded that omitting $\mathbf{b}_j$ leads to inflated Type I error (over optimistic conclusions w.r.t traits).

Main issue: the model becomes **much** slower.

Trait evolution

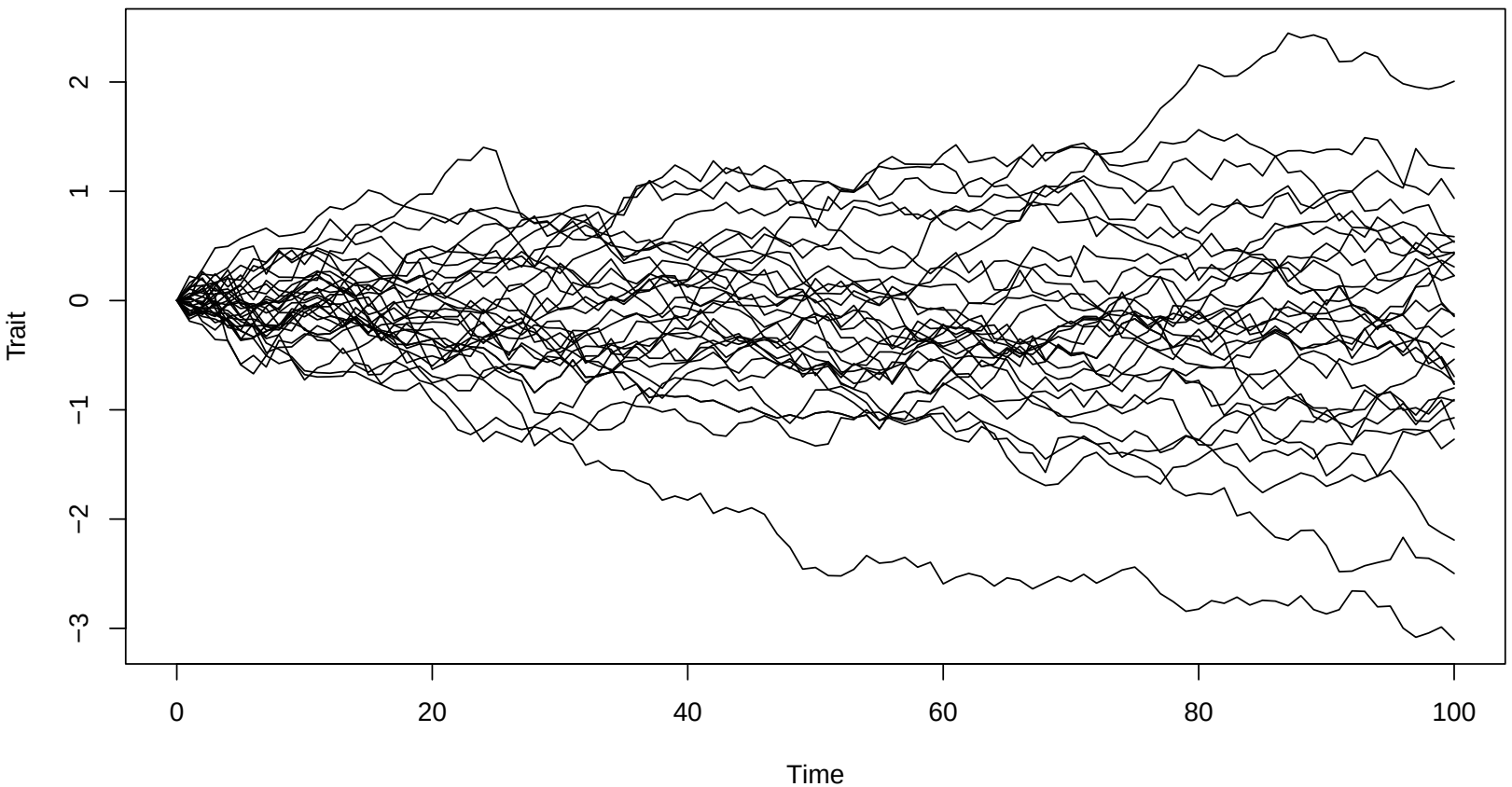
Traits develop by selection: you develop a trait if it increases your survival

- Related species might be similar
- Their evolutionary divergence is more recent, so less time to develop different characteristics
- Also in the environment, we might expect to see them in the same place



Brownian motion

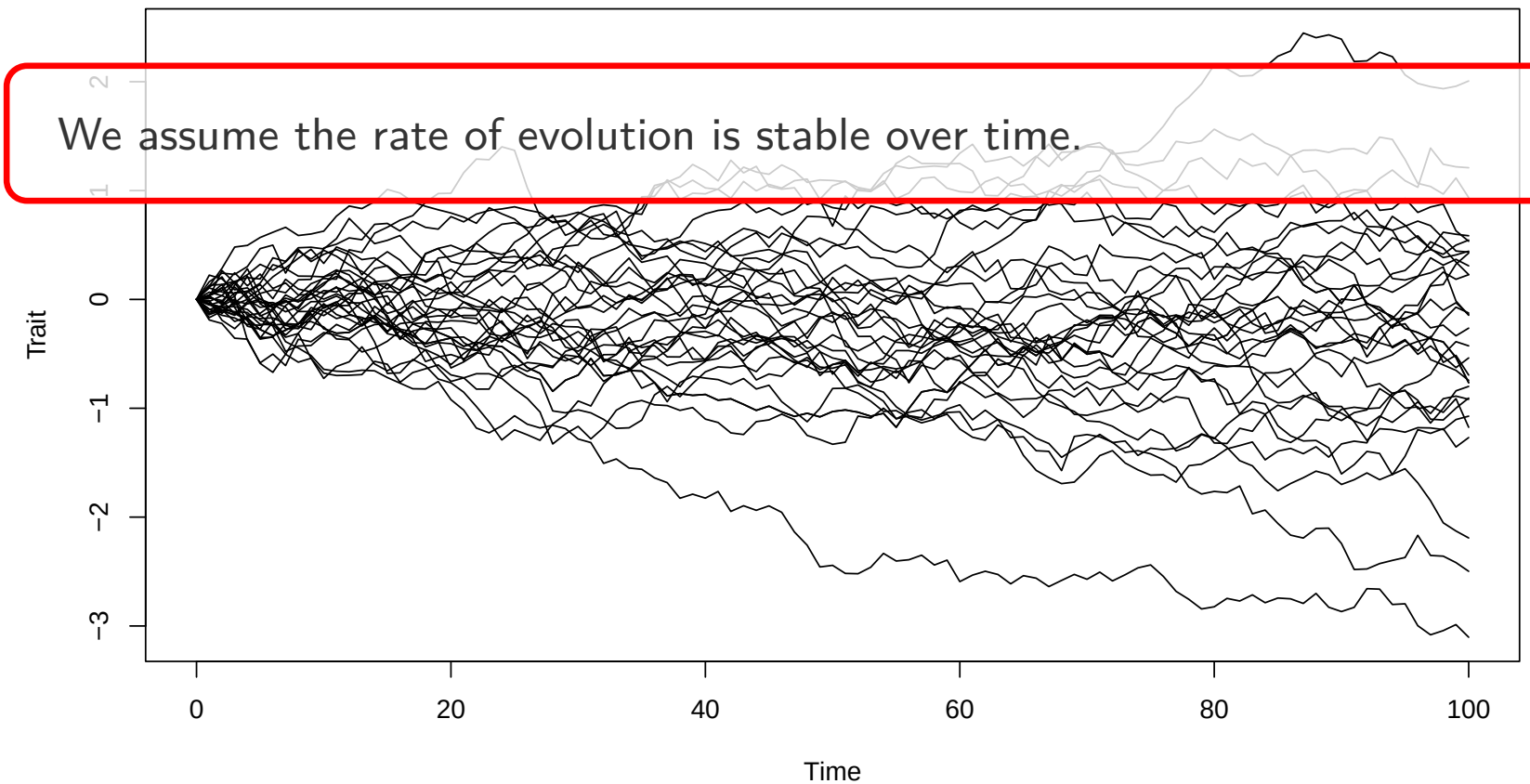
At time t our species j has the trait $tr^t_{j,k}$ given by the equation: $tr^t_{j,k} = tr^{t-1}_{j,k} + \epsilon^t_k \sim \mathcal{N}(0, \sigma^2_k)$ with covariance of the tips proportional to the shared branch length.



Brownian motion

At time t our species j has the trait $tr^t_{j,k}$ given by the equation: $tr^t_{j,k} = tr^{t-1}_{j,k} + \epsilon^t_k \sim \mathcal{N}(0, \sigma^2_k)$ with covariance of the tips proportional to the shared branch length.

We assume the rate of evolution is stable over time.



Competitive exclusion

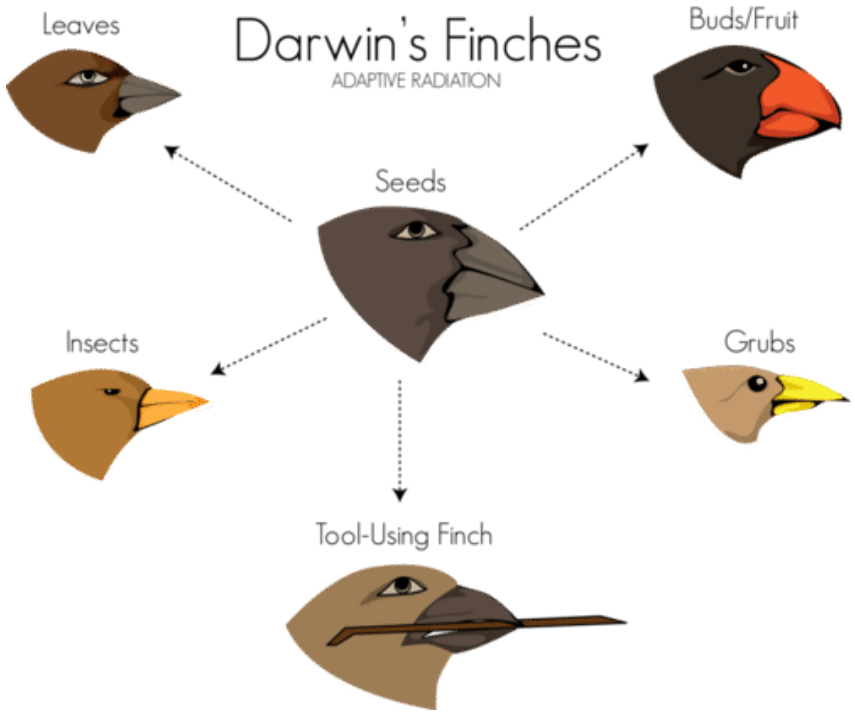


Figure 4: flexbooks.ck12.org, Christopher AuYeung

Species can (stably) co-occur, if they segregate their resource use.

Phylogenetic random effects

- In the 4th corner model $\mathbf{b}_j$ cab be structured by Phylogeny
- More closely related species have similar responses to the environment

The Phylogeny provides more information and makes for more accurate estimation

(and we can predict for species without data)

Phylogenetic random effects

Here I will omit traits for brevity. So our model is:

$$\boldsymbol{\eta} = \mathbf{1}\beta_{0j}^{\top} + \mathbf{XB} \tag{4}$$

- Now, $\mathbf{B}$ are the species random slopes for covariates
- In the simplest case we assume $\mathbf{B} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_m \otimes \boldsymbol{\Sigma}_r)$
- $\boldsymbol{\Sigma}_r$ covariance matrix of random effects ("traits")
- $\boldsymbol{\Sigma}_m$ correlation matrix due to phylogeny ("tips")

We assume that all our random effects are structured by the Phylogeny

Phylogenetic random effects

$$\mathbf{\Sigma}_m = \mathbf{C}\rho + (1-\rho)\mathbf{I} \tag{5}$$

1) $\mathbf{C}$ is a correlation matrix due to the Phylogeny (`ape::vcv(., corr = TRUE)`)
2) $0 \leq \rho \leq 1$ is Pagel's λ: the Phylogenetic signal parameter

Phylogenetic random effects

$$\mathbf{\Sigma}_m = \mathbf{C}\rho + (1-\rho)\mathbf{I} \tag{5}$$

1) $\mathbf{C}$ is a correlation matrix due to the Phylogeny (`ape::vcv(., corr = TRUE)`)
2) $0 \leq \rho \leq 1$ is Pagel's λ: the Phylogenetic signal parameter

This model only generates positive species associations.

Phylogenetic signal

- 1: Fully phylogenetically structured responses
- 0: Normal ("iid") random effects

When it is 0, it does not mean there is nothing going on.

Absence of phylogenetic signal:

- Scale mismatch
- Evolution moves very fast
- Too little information
- Traits are phylogenetically structured
- There are other (flexible) terms in the model
- Model misspecification

Presence of phylogenetic signal:

- Related species have similar "traits" (environmental response)
- Occupy similar environments

Model limitation

This phylogenetic model assumes traits evolve following the Brownian motion model of evolution. This can only generate positive associations.

But, competitive exclusion tells us that species evolve to differentiate resource.

- Similar species can (stably) co-occur if they utilize a different resource
- Similar species that utilize the same resource should not (stably) co-occur

The latter results in negative correlations, but no corresponding model for trait evolution has been developed

- unless species do not stably co-occur and/or evolution is still ongoing

Example with fungi data (Abrego 2021)

Received: 1 November 2021 | Accepted: 20 December 2021

DOI: 10.1111/1365-2745.13839

Journal of Ecology | BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Traits and phylogenies modulate the environmental responses of wood-inhabiting fungal communities across spatial scales

Nerea Abrego[1,2] | Claus Bässler[3,4] | Morten Christensen[5] | Jacob Heilmann-Clausen[6]

Example with fungi data

- 215 species (after cleaning)
- 1666 sites
- 19 covariates of various kinds

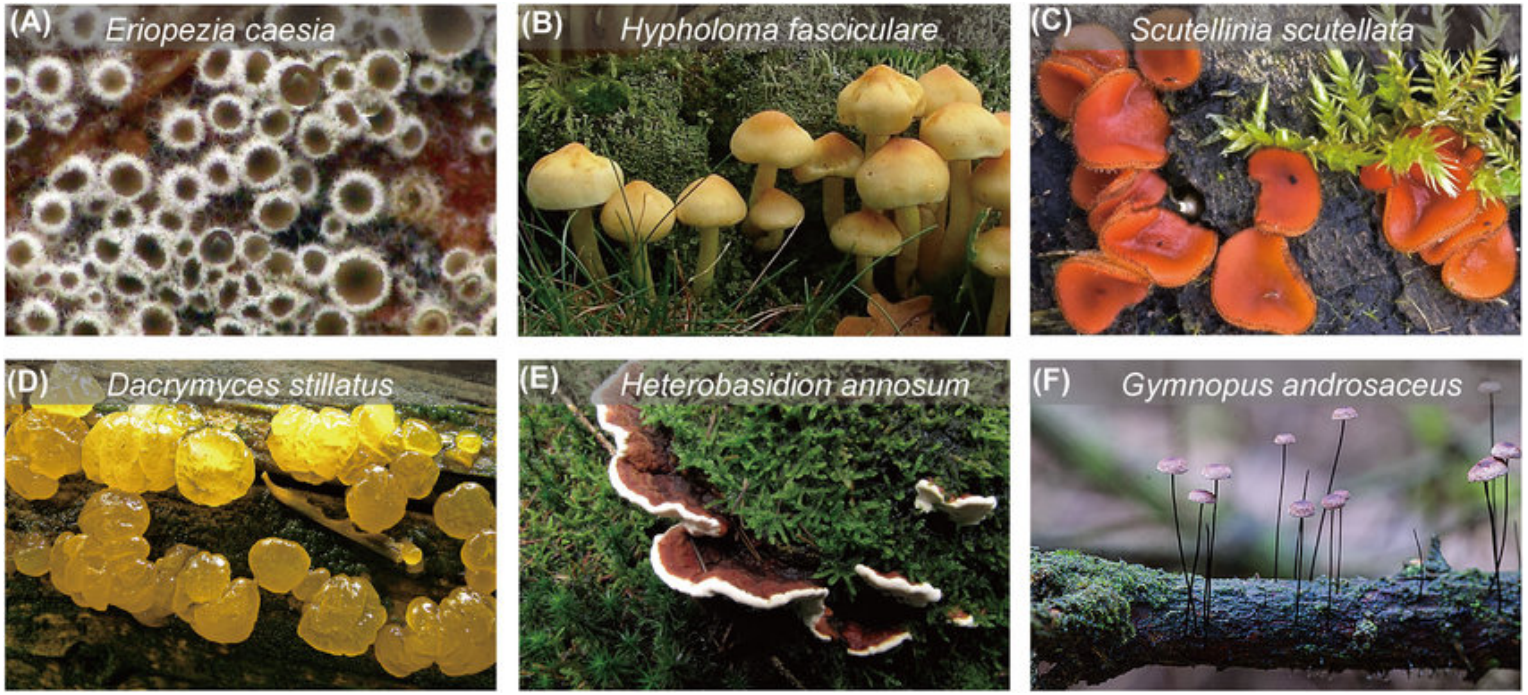
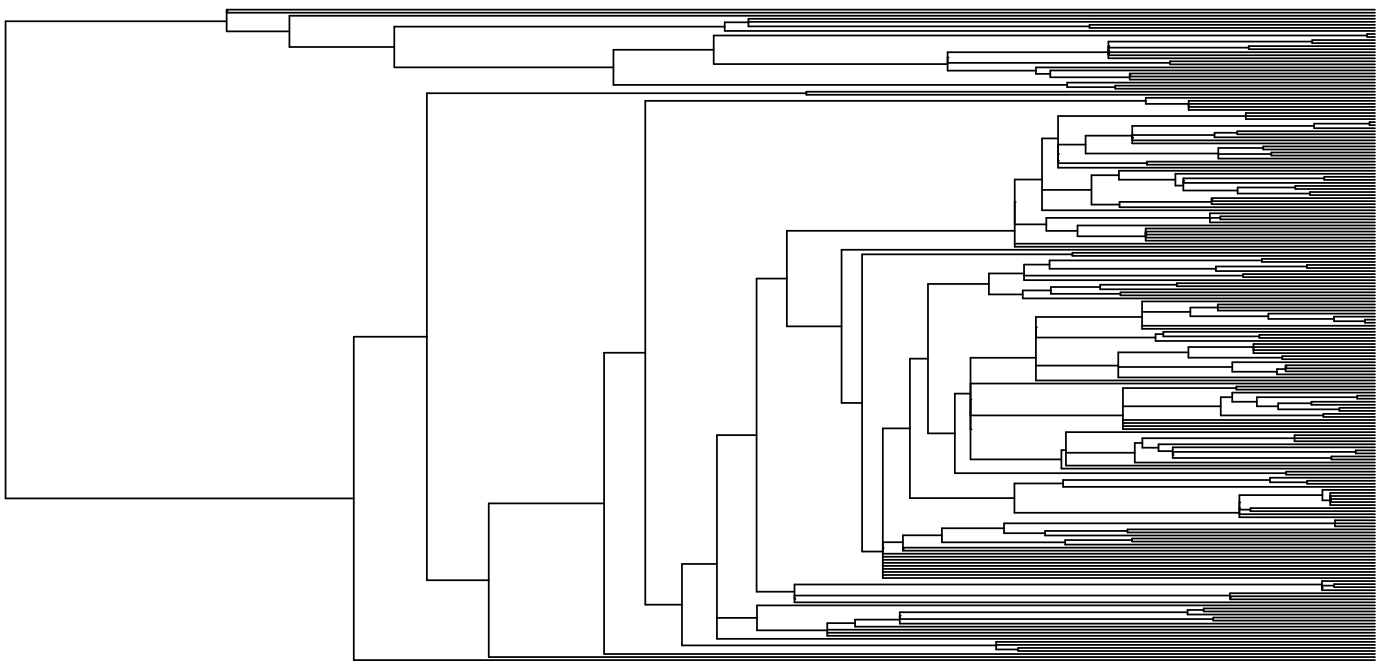


Figure 5: image from Yang et al. 2021

Example with fungi data



Example with fungi data

Phylogenetic models in `gllvm` use a **nearest neighbour approximation**

- We need to set the number of tips to consider on the tree
- The ordering of species matters!

```
covMat <- ape::vcv(tree)
e <- eigen(covMat)
distMat <- ape::cophenetic.phylo(tree)
ord <- gllvm:::findOrder(covMat = covMat, distMat = distMat,
species <- colnames(covMat)[ord]
Y <- Y[, species]
covMat <- covMat[species, species]
distMat <- distMat[species, species]
```

Ordering species

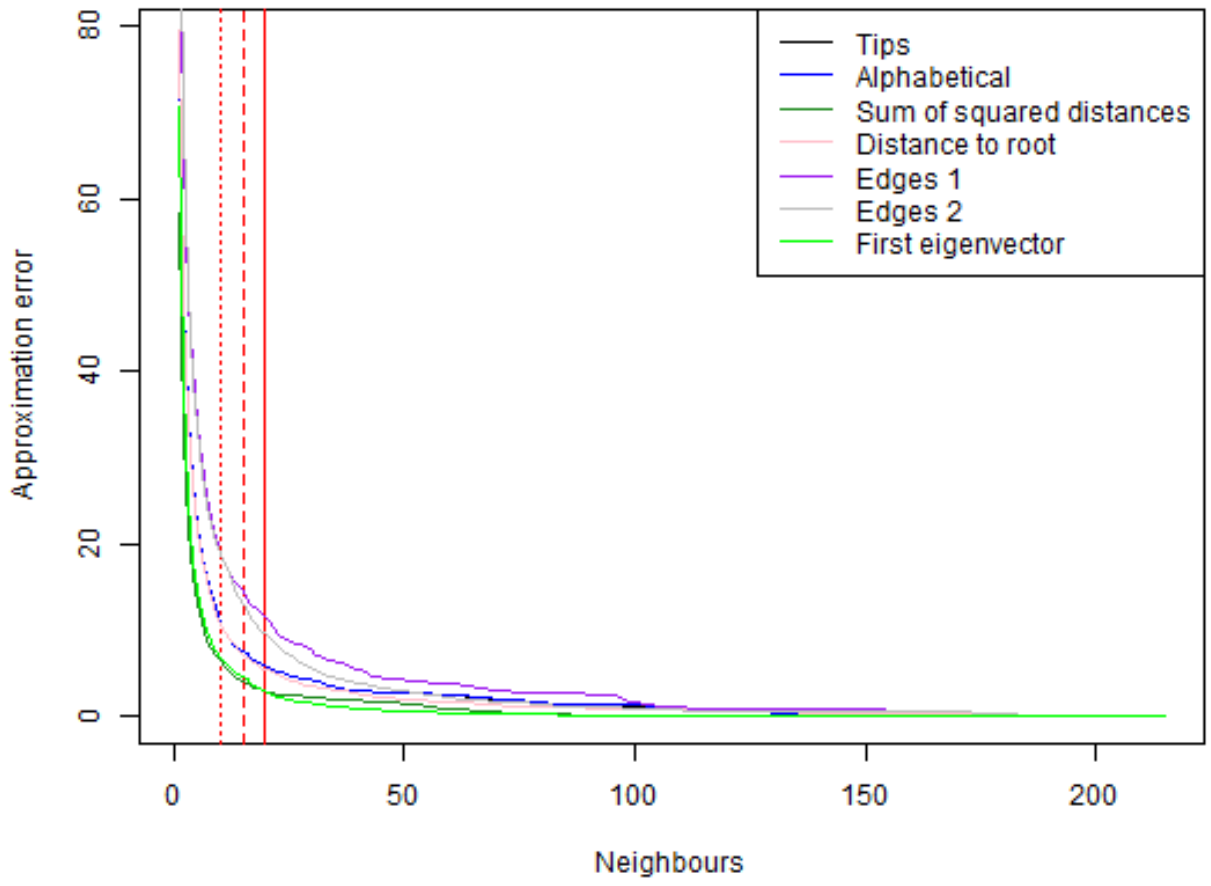


Figure 6: See vignette 7

Example with fungi data

```
TMB::openmp(parallel::detectCores()-1, autopar = TRUE, DLL = "gllvm")
model3 <- gllvm::gllvm(y = Y, X=X, family = "binomial", num.lv = 0, beta0com = TRUE,
      row.eff = ~(1 | REGION/RESERVE), studyDesign = X[,c("REGION","RESERVE")],
      formula = ~(DBH.CM+AVERDP+I(AVERDP^2)+CONNECT10+TEMPR+PRECIP+log.AREA|1),
      colMat = list(covMat, dist = distMat), nn.colMat = 15, max.iter = 10e3, optim.method = "L-BFGS-B")
```

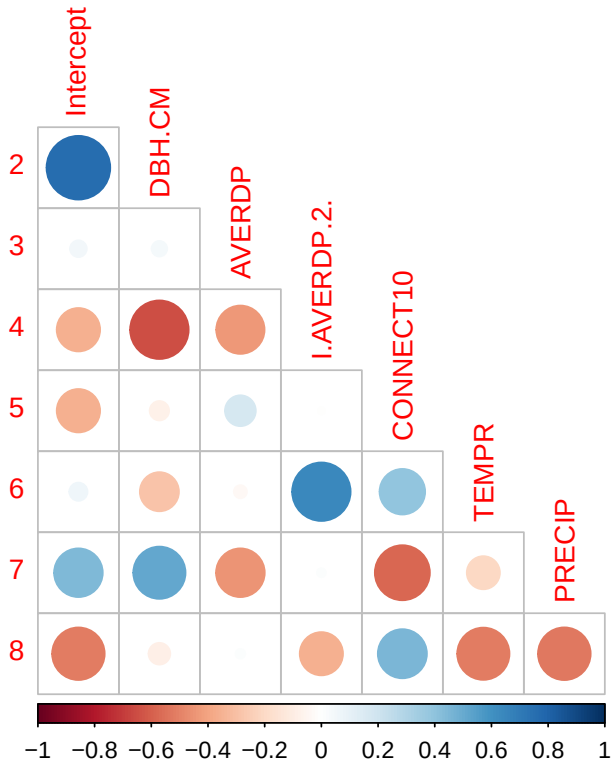
This takes a while to fit, but is really incredibly fast (it is just a complex model)

Example with fungi data

```
summary(model3)
```

```
##
## Call:
## gllvm::gllvm(y = Y, X = X, formula = ~(DBH.CM + AVERDP + I(AVERDP^2) +
##     CONNECT10 + TEMPR + PRECIP + log.AREA | 1), family = "binomial",
##     num.lv = 0, studyDesign = X[, c("REGION", "RESERVE")], colMat = list(covMat,
##         dist = distMat), row.eff = ~(1 | REGION/RESERVE), beta0com = TRUE,
##     nn.colMat = 15, max.iter = 10000, optim.method = "L-BFGS-B")
##
## Family:  binomial
##
## AIC:  103171.8 AICc:  103171.8 BIC:  103678.9 LL:  -51539 df:  47
##
## Informed LVs:  0
## Constrained LVs:  0
## Unconstrained LVs:  0
##
## Formula:  ~(DBH.CM + AVERDP + I(AVERDP^2) + CONNECT10 + TEMPR + PRECIP + log.AREA | 1)
## LV formula:  ~ 0
## Row effect:  ~(1 | REGION/RESERVE)
##
## Random effects:
##  Name        Signal Variance Std.Dev Corr
##  Intercept   0.6037 1.0495   1.0244
##  DBH.CM      0.6037 0.0051   0.0715   0.7642
##  AVERDP      0.6037 0.1796   0.4238   0.0529  0.0458
##  I.AVERDP.2. 0.6037 0.0066   0.0815  -0.3550 -0.6454 -0.4397
##  CONNECT10   0.6037 0.0401   0.2003  -0.3544 -0.0711  0.1790 -0.0091
##  TEMPR       0.6037 0.0689   0.2625   0.0626 -0.2879 -0.0321  0.6438  0.3917
##  PRECIP      0.6037 0.0440   0.2098   0.4461  0.5139 -0.4465  0.0148 -0.5701
##  log.AREA    0.6037 0.0140   0.1184  -0.5196 -0.0889  0.0173 -0.3518  0.4538
##
##
##
##
##
```

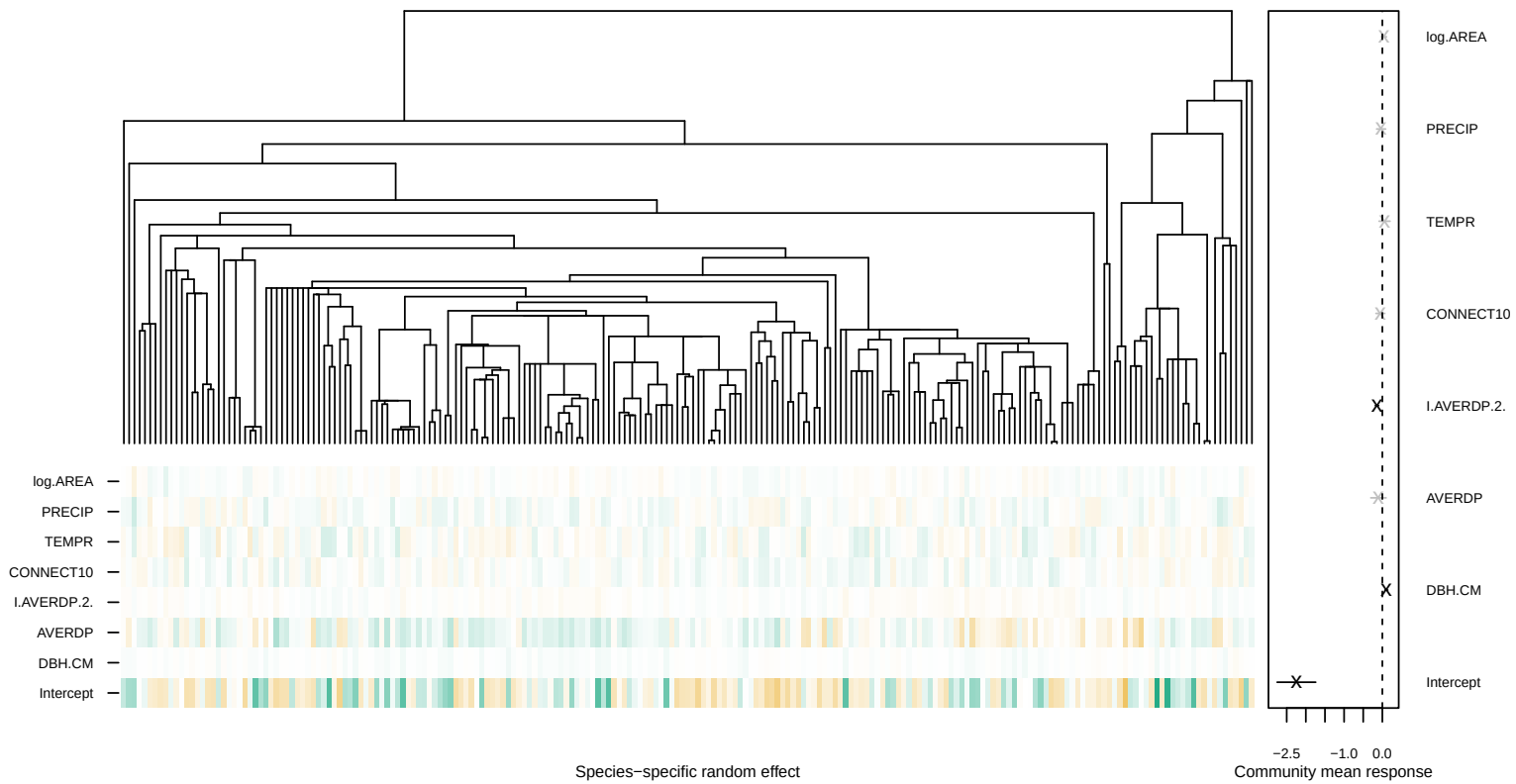
Example with fungi data



Correlated effects: trait syndromes. Fungi with a particular preference in temperature range, might also have a preference for less precipitation.

Example with fungi data

```
gllvm::phyloplot(model3, tree)
```



Conclusion:

There is phylogenetic structuring; species environmental responses are more similar if they have a shared evolutionary history.

Of course, this might be covariate dependent (`colMat.rho.struct = "term"`)

Summary

- JSDMs is a framework for analysing species co-occurrence data
- Focussed on prediction, but also suitable for inference
- We can also fit models with non-binary data (e.g., counts or biomass)
- The GLLVM framework is used here to implement JSDM efficiently
- We can incorporate random effects
- Phylogenetically structure species' effects
- Above all: we incorporate correlation of species