Joint Species Distribution Modeling

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline

- Attitude
- Species associations
- Residual correlations concurrent ordination
- Fourth corner LVMs
- Phylogenetic models

Questions so far?

Background

If you have presence-absence data of a species, you fit a Species Distribution Model of the form:

$$y_{ij} = \alpha + \mathbf{x}_i^\top \boldsymbol{\beta} \tag{1}$$

- $\mathbf{x}_i$ is usually a bioclimatic variable
- Then you want to predict where a species may occur
- Potentially based on future climate scenarios
- But what if you have data on multiple species?
  - Fitting multiple SDMs requires *independence*

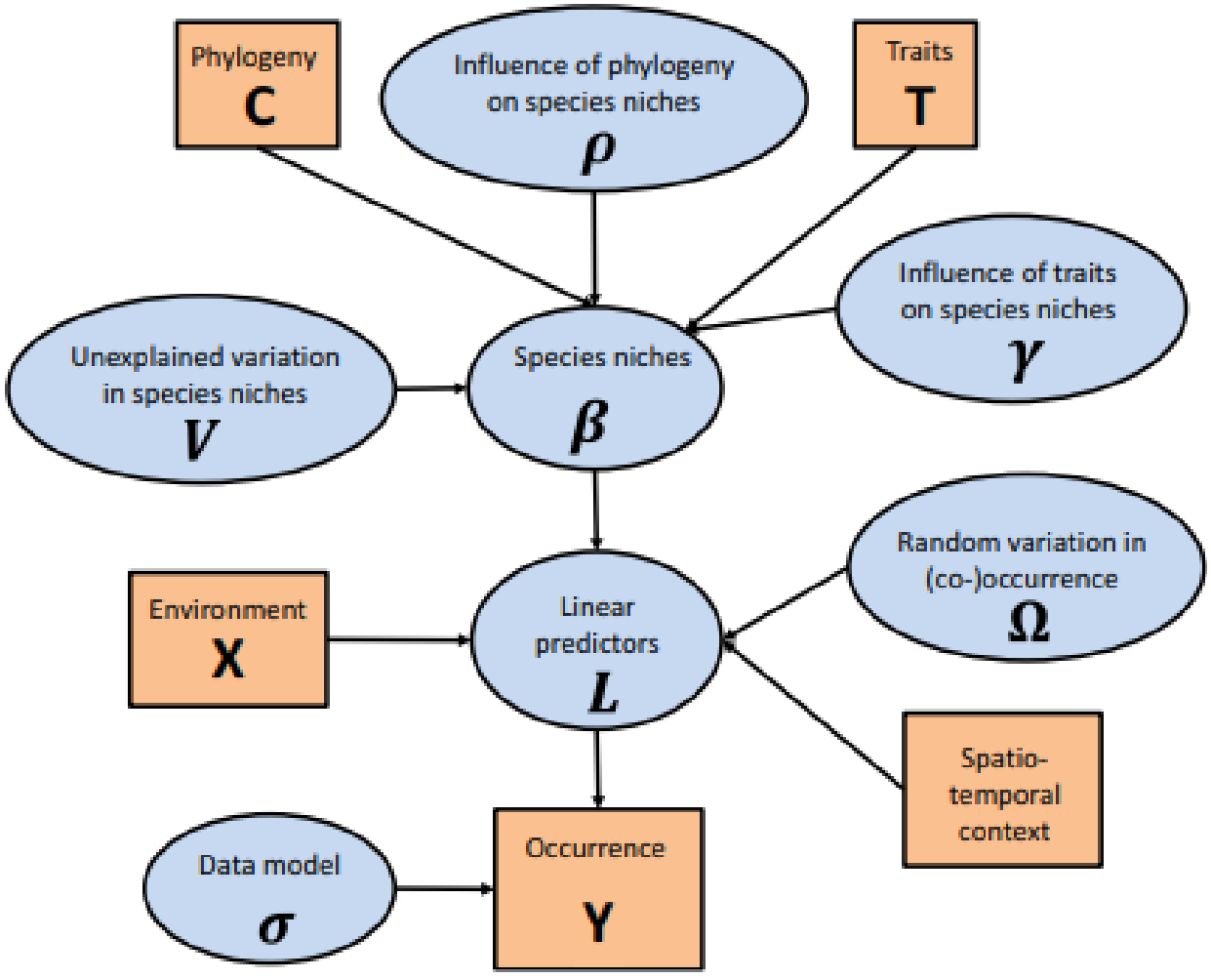
Joint Species Distribution Modeling

A decade ago, Joint Species Distribution Models (JSDM) were introduced to model data of multiple species

- Pollock et al. (2015): co-occurrence of frogs and trees
- Clark et al. (2015): co-occurrence of trees

Both of these are multivariate probit models

Joint Species Distribution Modeling



Joint Species Distribution Modeling

The idea of JSDMs was to incorporate *species associations*

- Species may co-occur due to biotic interactions
- Due to similar environmental preferences
- Or because they have a similar history

Either how, it results in correlations between responses

JSDM: the model

$$\eta_{ij} = \beta_{0j} + \ldots + \epsilon_{ij} \tag{2}$$

- $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$
- $\boldsymbol{\Sigma}$ is the matrix of *species associations*
- So we expect a positive values of species co-occur, and negative if they do not
- Problem: its size grows very quickly

Species associations

- Difficult to estimate: there are usually too many parameters
- Can only fir this way when there are (much) more sites than species
- The number of pairwise associations grows quadratically
  - 2 with 2 species, 6 for 4 species, 45 for 10 species, 4950 for 100

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & sp_{12} & \cdots & sp_{1j} \\ sp_{21} & 1 & \cdots & sp_{2j} \\ \vdots & & \ddots & \vdots \\ sp_{j1} & sp_{j2} & \cdots & 1 \end{bmatrix} \tag{3}$$

This very quickly becomes an issue for fitting models

Ordination to the rescue

GLLVMs were introduced as a technical solution to this problem

- The species loadings represent correlation of species, for random effect LVs
- $\epsilon_{ij} \overset{d}{\approx} \mathbf{z}_i^\top \boldsymbol{\gamma}_j$
- The square of species loadings forms the species association matrix: $\boldsymbol{\Sigma} \approx \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top$

"The factor analytic solution" because factor analysis (Spearman, 1904) is the precursor of GLLVMs

Ordination vs. JSDMs

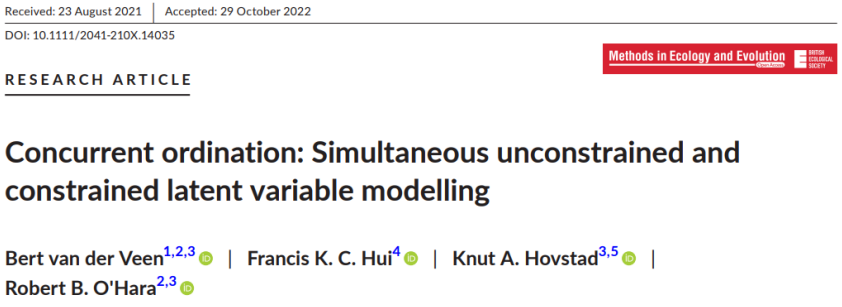
O'Hara and van der Veen (submitted) discuss some differences. Generally:

	Ordination	JSDM
Focus	Species relationships	Distributions
Goal	Inference	Prediction
Data type	Usually quantitative	Binary
Scale	Local	Regional
Covariates	Environmental	Bioclimatic
Presentation	Ordination diagram	Correlation plot/map
Audience	Community ecologists	Macro ecologists

Ordination vs. JSDMs

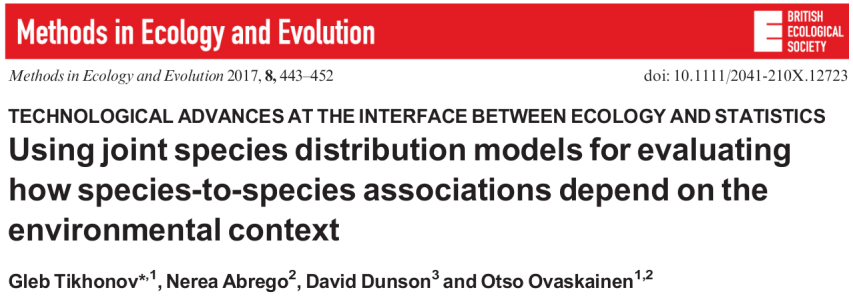
The models formulated are usually different:

Ordination: formulate models based on LVs
JSDM: formulate models based on species associations


Received: 23 August 2021 | Accepted: 29 October 2022
DOI: 10.1111/2041-210X.14035
Methods in Ecology and Evolution
RESEARCH ARTICLE

Concurrent ordination: Simultaneous unconstrained and constrained latent variable modelling

Bert van der Veen[1,2,3] | Francis K. C. Hui[4] | Knut A. Hovstad[3,5] | Robert B. O'Hara[2,3]


Methods in Ecology and Evolution
BRITISH ECOLOGICAL SOCIETY
Methods in Ecology and Evolution 2017, **8**, 443–452 doi: 10.1111/2041-210X.12723
TECHNOLOGICAL ADVANCES AT THE INTERFACE BETWEEN ECOLOGY AND STATISTICS

Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context

Gleb Tikhonov*[,1], Nerea Abrego[2], David Dunson[3] and Otso Ovaskainen[1,2]

Ordination and species associations

In `gllvm` we can also extract species associations: one more tool for examining our results
With the `getResidualCov` and `getResidualCor`, or `getEnvironmentalCov` and `getEnvironmentalCor` functions

Example with Eucalyptus data (Pollock et al. 2015)

```
Y <- read.csv("../data/eucalyptY.csv")[,-1]
X <- read.csv("../data/eucalyptX.csv")[,-1]
X <- data.frame(lapply(X, function(x)if(is.numeric(x)){scale(
```

- 20 species
- 458 sites
- Soil covariates and a few bioclimatic

Example with Eucalyptus data

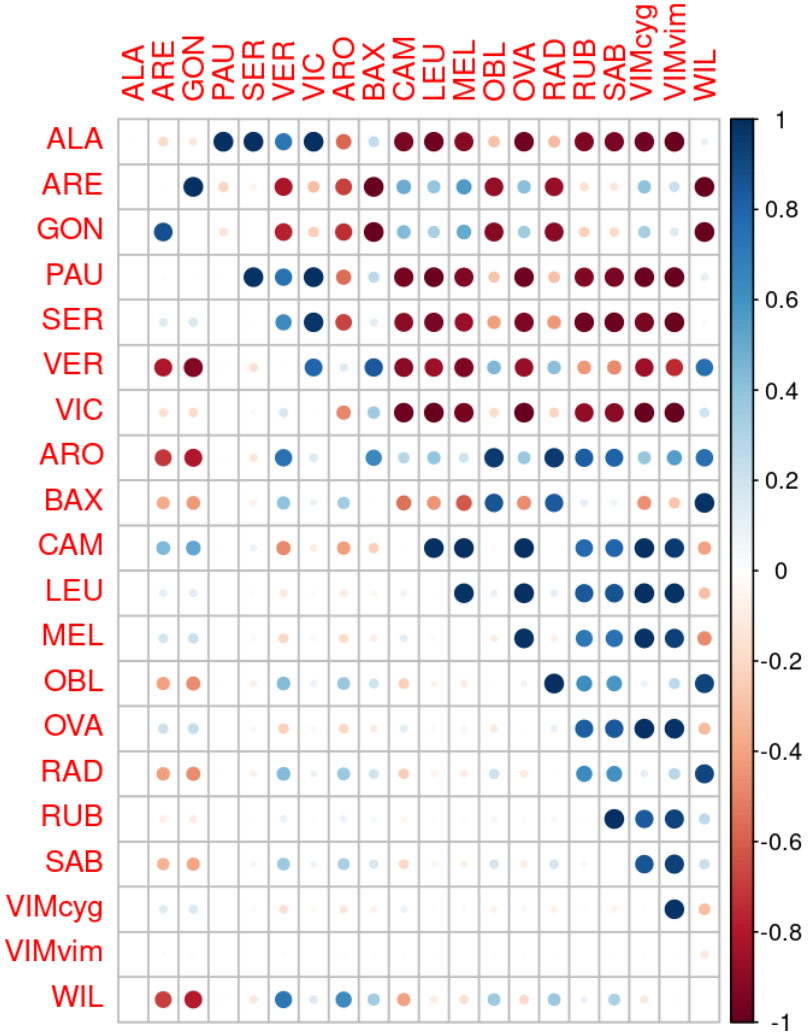
```
cnord <- gllvm::gllvm(Y, X = X, lv.formula =
  ~ Rockiness + Loaminess + Sandiness + cvTemp + PPTann + VallyBotFlat,
  family = "binomial", num.lv.c = 2, randomB="LV", reltol=1e-15,maxit=6e3,
  starting.val="res",seed=1, optim.method="L-BFGS-B",
  link = "logit",n.init=100, n.init.max=10)
```

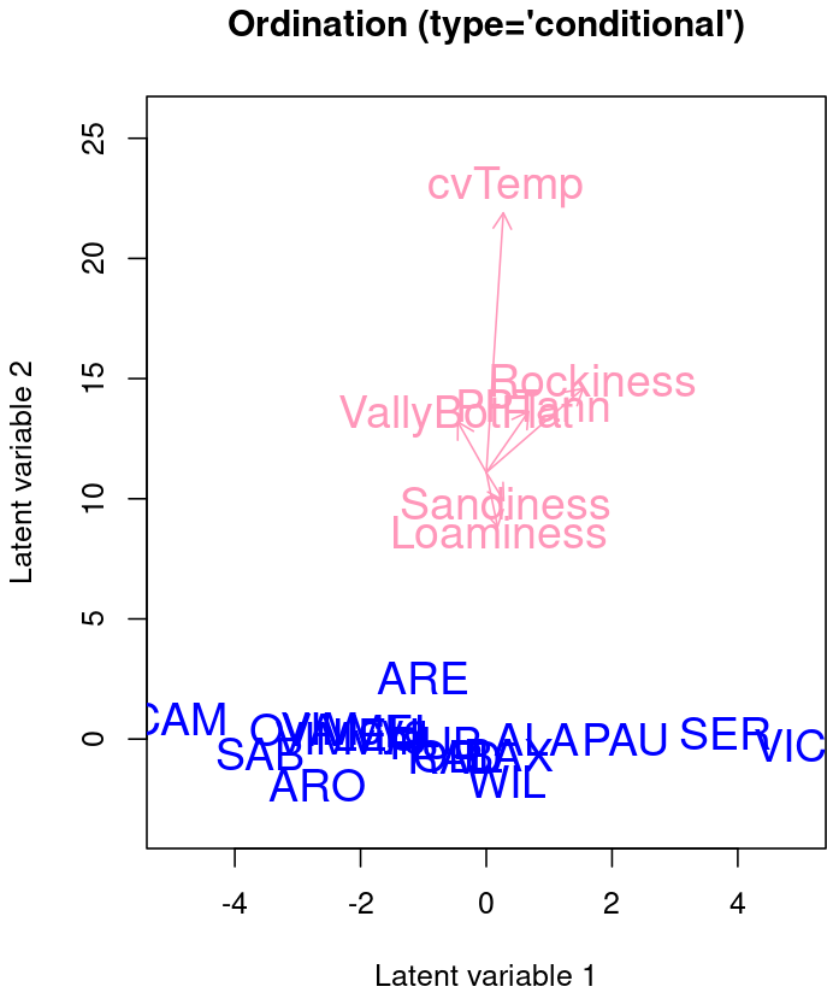
- First LV has little unmeasured variation
- Second LV has a lot of unmeasured variation

Example with Pollock et al (2015) data

```
Covar.cor <- gllvm::getEnvironCor(cnord)
Resid.cor <- gllvm::getResidualCor(cnord)
ReOrderNames <- c("ALA", "ARE", "GON", "PAU", "SER", "VER", "VIC",
  "ARO", "BAX", "CAM", "LEU", "MEL", "OBL", "OVA", "RAD", "RUB", "SAB", "V
Covar.cor <- Covar.cor[ReOrderNames, ReOrderNames]
Resid.cor <- Resid.cor[ReOrderNames, ReOrderNames]
# LV above diagonal, covariates below
Covar.cor.tmp <- Covar.cor
Covar.cor.tmp[lower.tri(Covar.cor.tmp, diag = TRUE)] <- 0
Resid.cor.tmp <- Resid.cor
Resid.cor.tmp[upper.tri(Resid.cor.tmp, diag = TRUE)] <- 0
Correlations <- Covar.cor.tmp + Resid.cor.tmp
# Covariates on lower triangle, residual on upper triangle
corrplot::corrplot(Correlations, tl.cex = 1, win.asp = 1.5)
gllvm::ordiplot(cnord, biplot = TRUE, s.colors = "transparent",
xlim = c(-5,5), cex.spp = 1.5, cex.env = 1.5, arrow.scale = 2, rotate = FA
```

Example with Pollock et al (2015) data

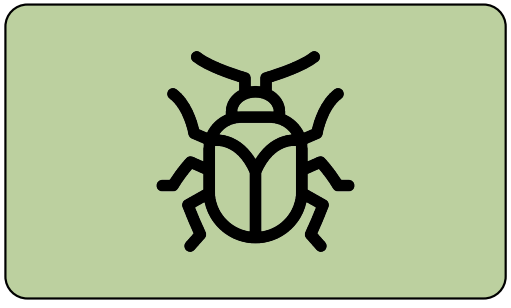


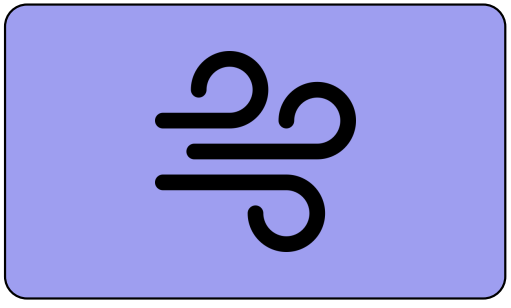


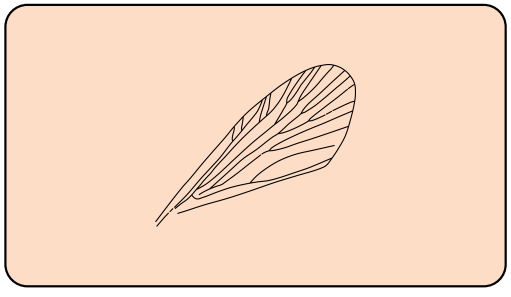
The 4th corner

At present, model-based ordination techniques are lacking that incorporate traits

Fourth corner analysis

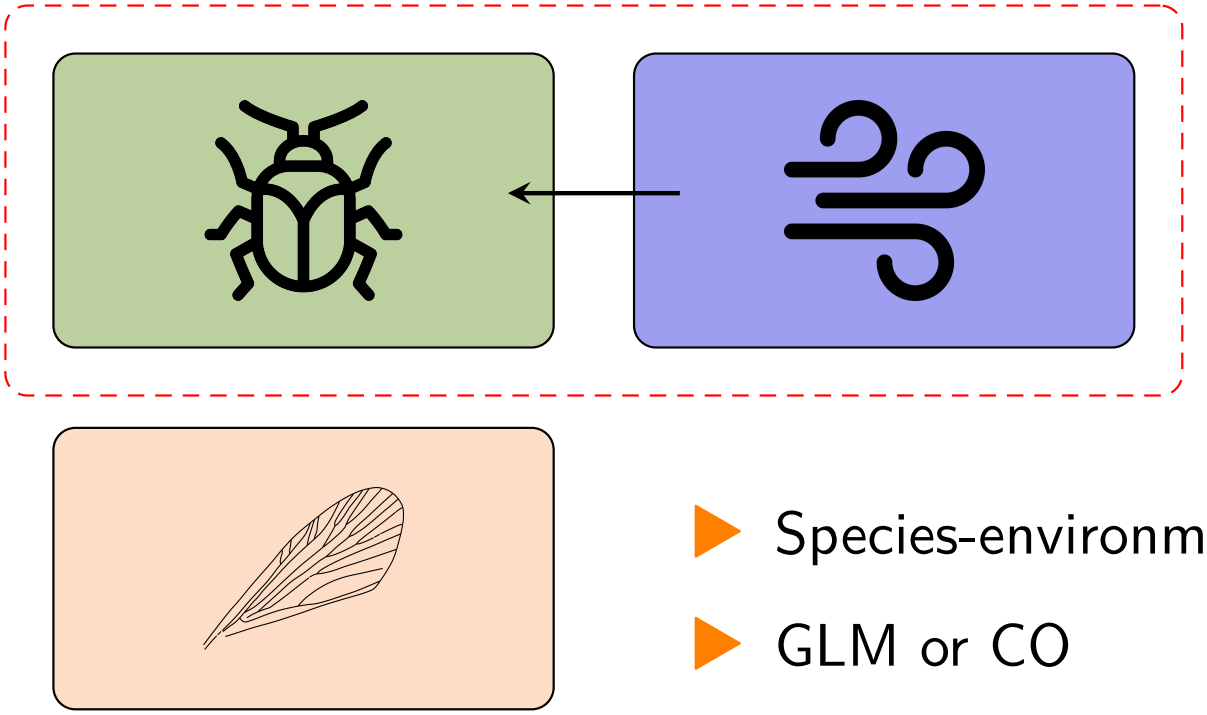






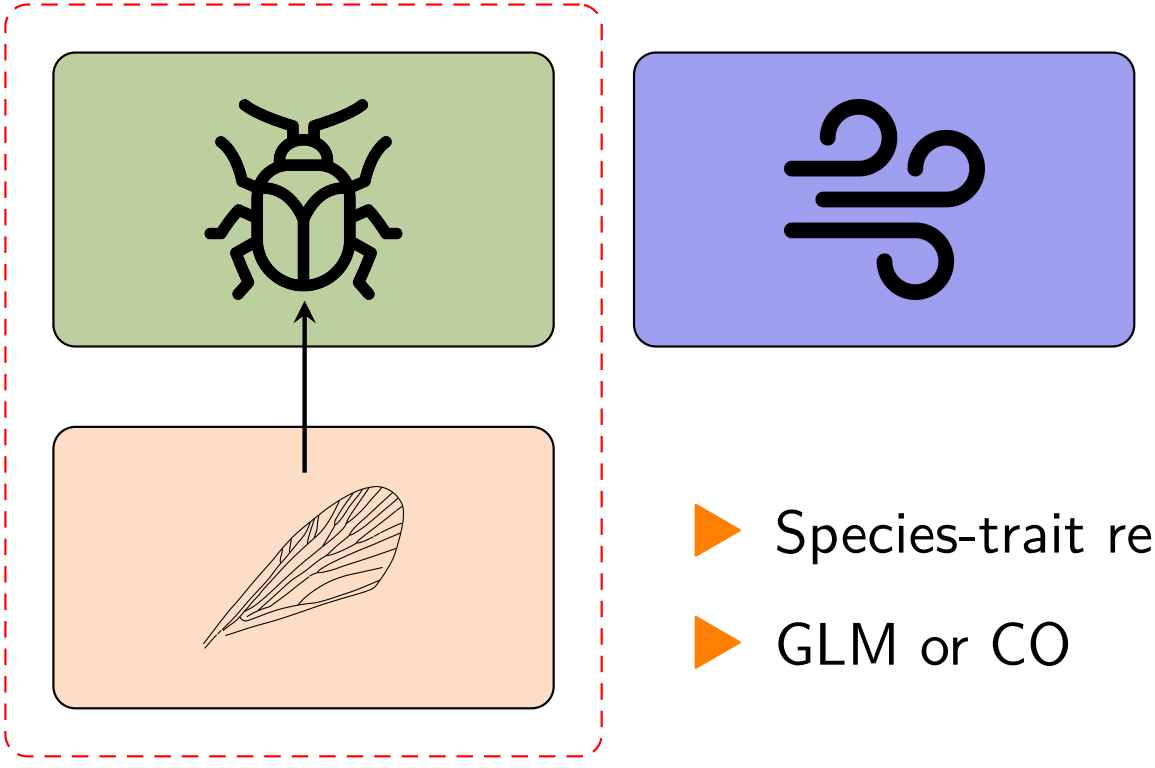
- **Y**: community data
- **X**: environmental variables
- **TR**: species traits

Fourth corner analysis



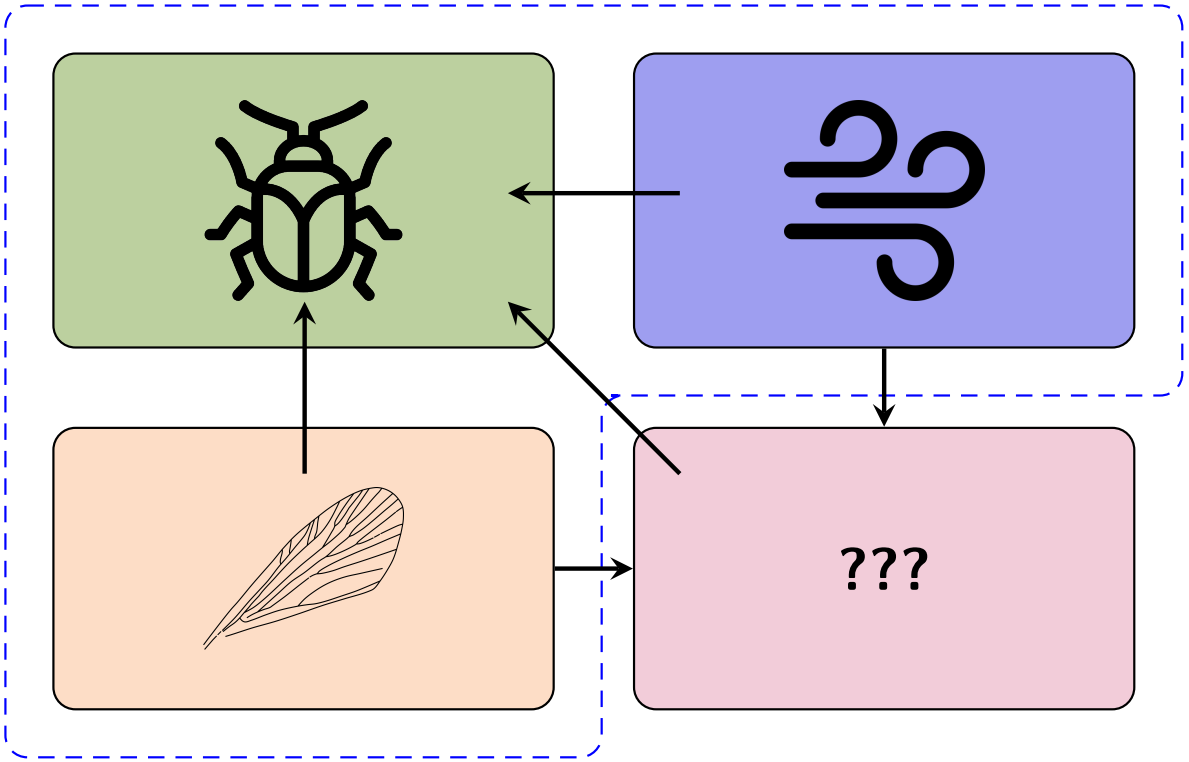
- Species-environment relationship
- GLM or CO

Fourth corner analysis



- Species-trait relationship
- GLM or CO

Fourth corner analysis



Trait-based analysis

1. CWM + RDA *Doledec et al. (1996)*
2. Double constrained ordination *Lebreton et al. (1988), ter Braak et al. (2018)*
3. Fourth corner (LV) Models *Brown et al. (2014), Ovaskainen at al. (2017), Niku et al. (2021)*

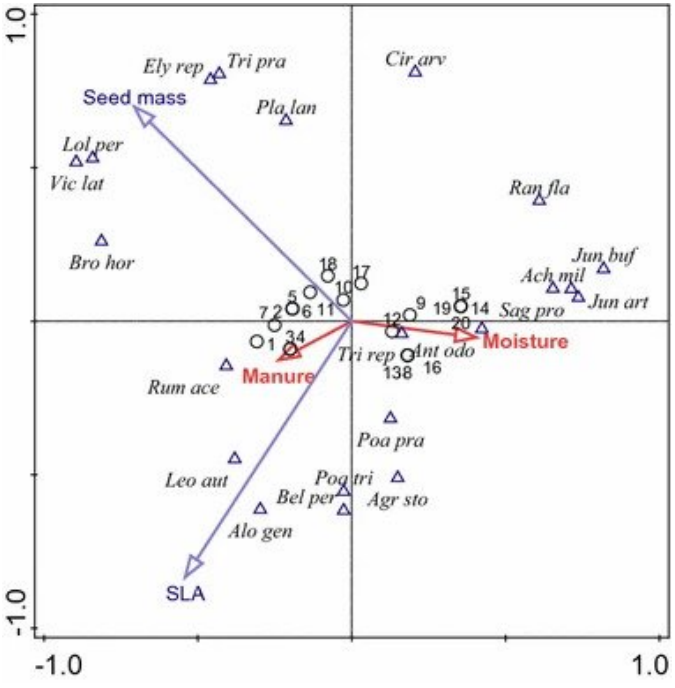


Figure 2: Quadriplot *ter Braak et al. (2018)*

Fourth-corner LVMs

A kind of LVM (JSDM) that also includes traits

Received: 19 February 2020 | Revised: 6 April 2021 | Accepted: 9 April 2021

DOI: 10.1002/env.2683

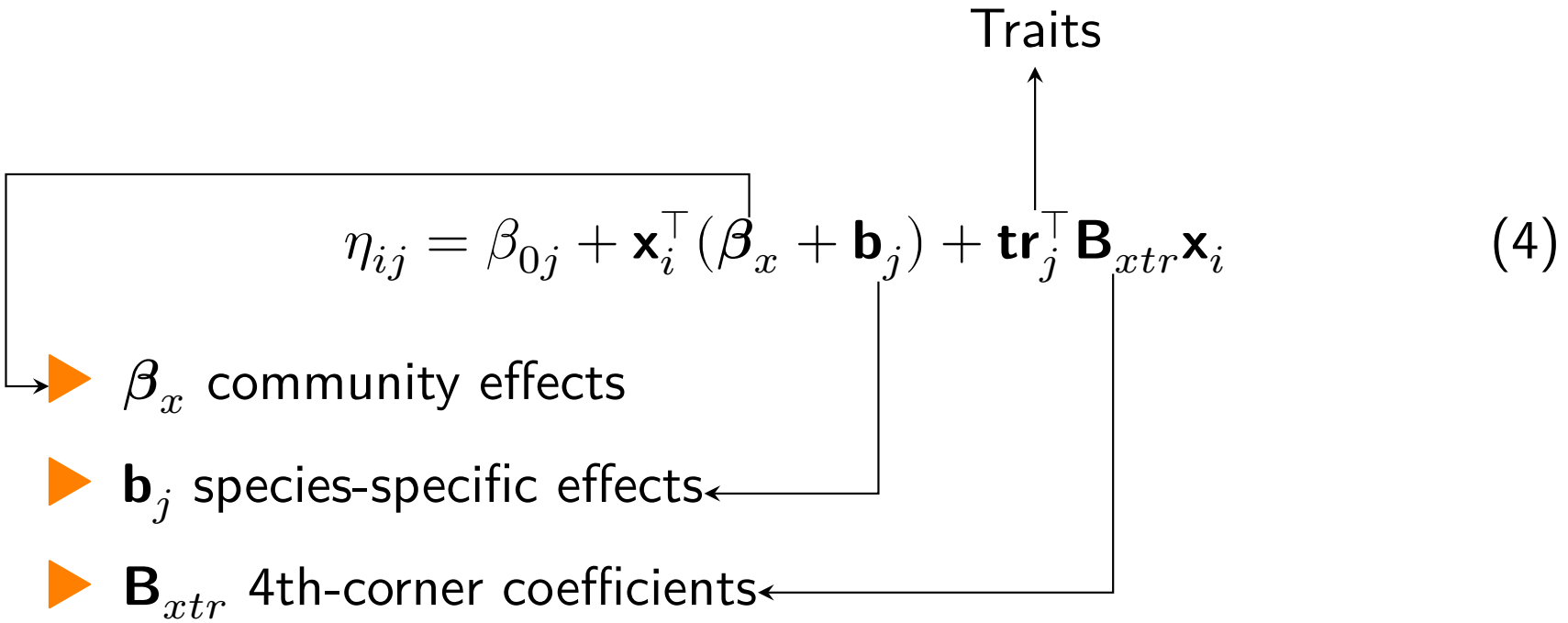
SPECIAL ISSUE PAPER

WILEY

Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models

Jenni Niku[1] | Francis K. C. Hui[2] | Sara Taskinen[1] | David I. Warton[3]

Fourth-corner LVMs



Fourth-corner LVMs

Hierarchically this means:

$$\boldsymbol{\beta}_j = \boldsymbol{\beta}_x + \mathbf{B}_{xtr}\mathbf{tr}_j + \mathbf{b}_j \tag{5}$$

So, we are modeling our species coefficients of covariates

- With an mean effect: $\boldsymbol{\beta}_x$
- Slopes for traits: $\mathbf{B}_{xtr}$, these describe how our species' environment responses depend on traits
  - Can be used nicely for hypothesis testing
- An error term $\mathbf{b}_j$ to capture everything left

Fourth-corner LVMs

- With a 4th corner LVM we can examine trait-environment relationships
- Figure out **why** species prefer particular conditions
- While still incorporating **other reasons** for co-occuring (or not)

Example with Eucalyptus data

```
TR <- read.csv("../data/eucalyptTR.csv");row.names(TR) <- TR$TAXON
```

Example with Eucalyptus data

```
library(gllvm)
TMB::openmp(7)
fit4th <- gllvm(Y, X, TR = TR,
formula = ~ Rockiness + Loaminess + Sandiness + cvTemp + PPTann + VallyBot
(Rockiness + Loaminess + Sandiness + cvTemp + PPTann + VallyBotFlat):
( MedianSLA + MaxHeight.m. + MedianSeedMass.mg.),
randomX = ~ Rockiness + Loaminess + Sandiness + cvTemp + PPTann + VallyBot
                    , family = "binomial", seed = 1, n.init = 3)
```

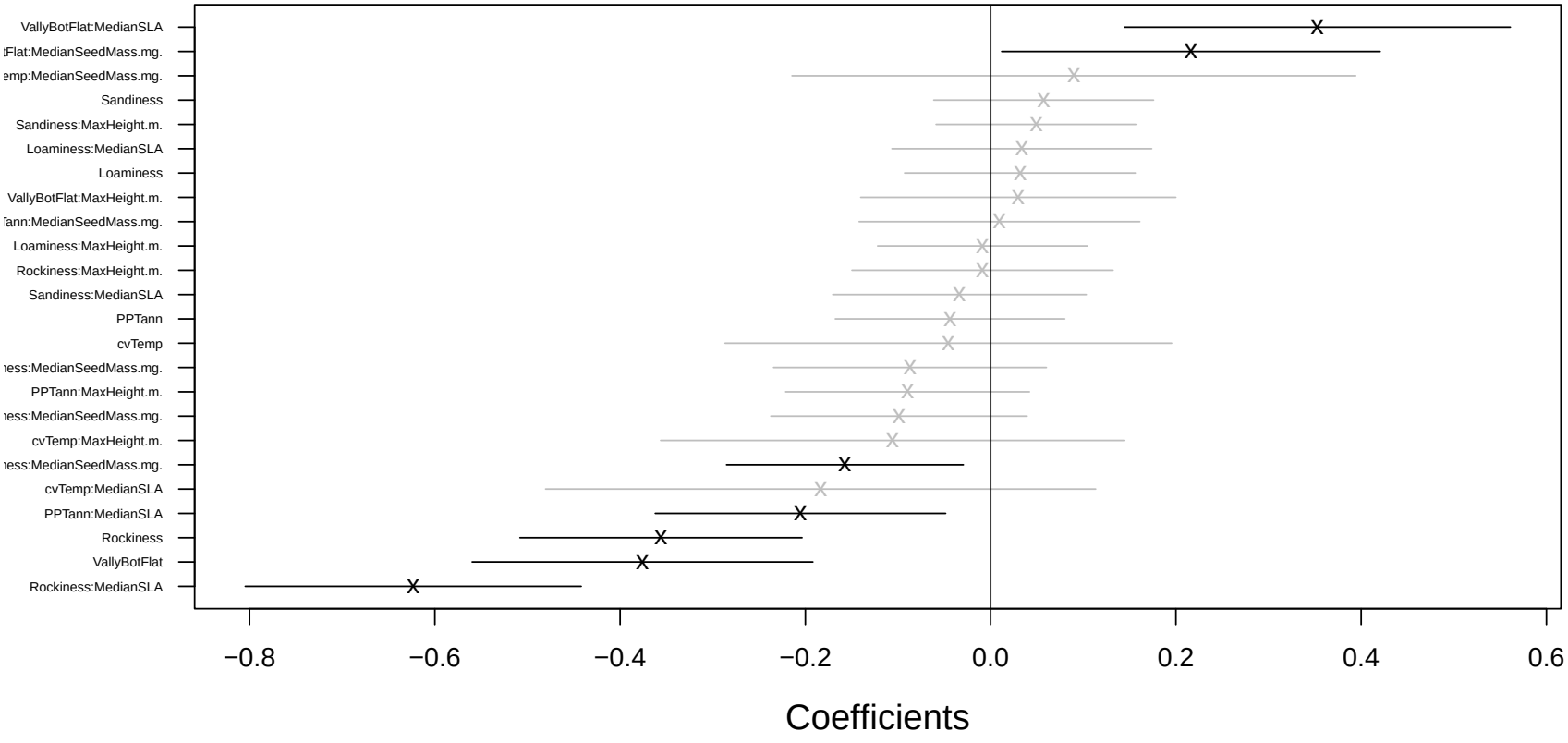
Example with Eucalyptus data

```
summary(fit4th)
```

```
##
## Call:
## gllvm(y = Y, X = X, TR = TR, formula = ~Rockiness + Loaminess +
##     Sandiness + cvTemp + PPTann + VallyBotFlat + (Rockiness +
##     Loaminess + Sandiness + cvTemp + PPTann + VallyBotFlat):(MedianSLA +
##     MaxHeight.m. + MedianSeedMass.mg.), family = "binomial",
##     randomX = ~Rockiness + Loaminess + Sandiness + cvTemp + PPTann +
##         VallyBotFlat, seed = 1, n.init = 3)
##
## Family:  binomial
##
## AIC:  4039.813 AICc:  4042.224 BIC:  4780.563 LL:  -1915.9 df:  104
##
## Informed LVs:  0
## Constrained LVs:  0
## Unconstrained LVs:  2
##
## Formula:  ~Rockiness+Loaminess+Sandiness+cvTemp+PPTann+VallyBotFlat+Rockiness:MedianSLA+Rockiness:MaxHeight.m.+Rockiness:MedianSeedMass.mg.+Loaminess:MedianSLA+Lo
## LV formula:  ~ 0
##
## Random effects:
##  Name         Variance Std.Dev Corr
##  Rockiness    0.0415   0.2038
##  Loaminess    0.0408   0.2019  -0.7920
##  Sandiness    0.0312   0.1765  -0.9638  0.6042
##  cvTemp       0.2678   0.5175   0.1401 -0.4901 -0.0165
##  PPTann       0.0496   0.2227  -0.9335  0.7110  0.8919  0.1719
##  VallyBotFlat 0.0983   0.3135   0.9160 -0.6706 -0.8843 -0.2289 -0.9981
##
## Coefficients predictors:
##                               Estimate Std. Error z value Pr(>|z|)
## Rockiness                    -0.355858   0.077696  -4.580 4.65e-06 ***
## Loaminess                     0.032111   0.063766   0.504 0.614560
## Sandiness                     0.057208   0.060521   0.945 0.344534
## cvTemp                       -0.045654   0.122972  -0.371 0.710450
```

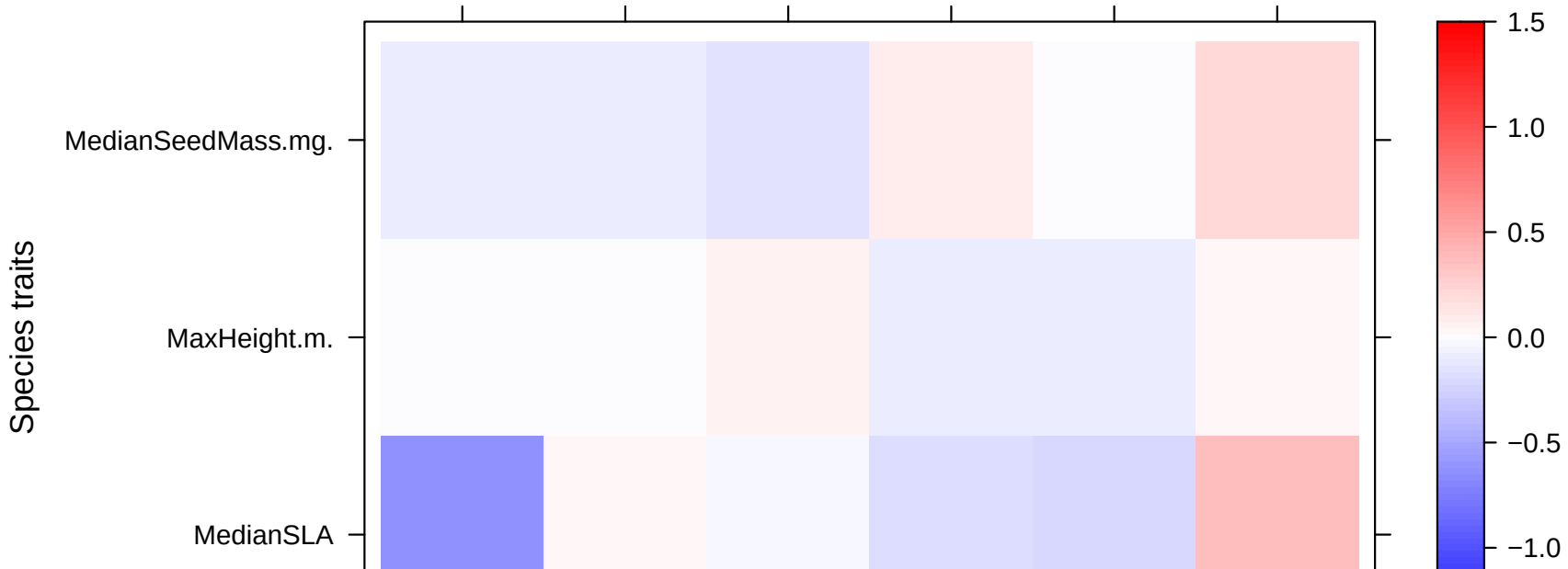
Example with Eucalyptus data

```
gllvm::coefplot(fit4th)
```



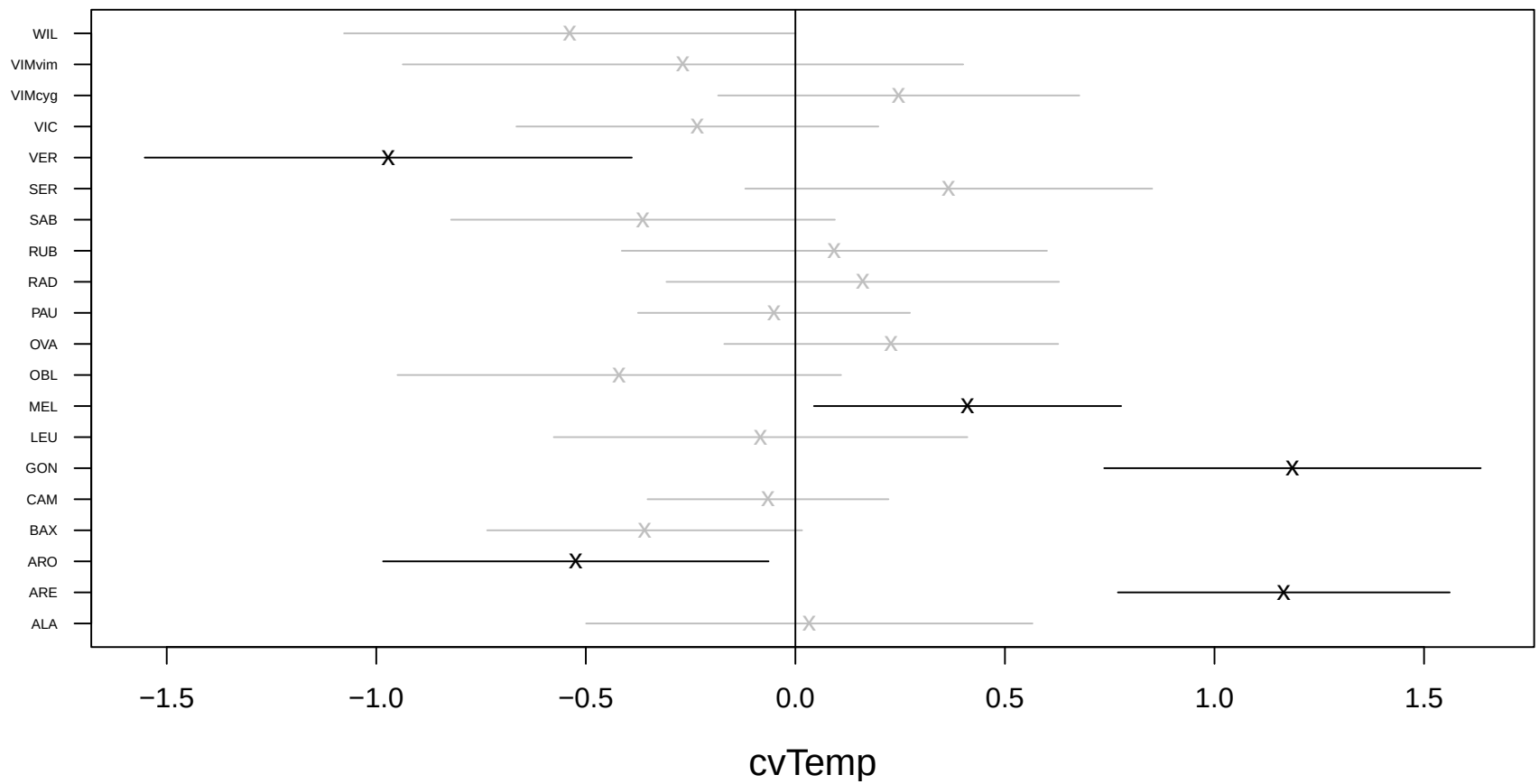
Example with Eucalyptus data

```
a <- 1.5
colort <- colorRampPalette(c("blue", "white", "red"))
plot.4th <- lattice::levelplot((as.matrix(fourth)), xlab = "Environmental
                    ylab = "Species traits", col.regions = colort(100),
                    at = seq(-a, a, length = 100), scales = list(x = lis
plot.4th
```



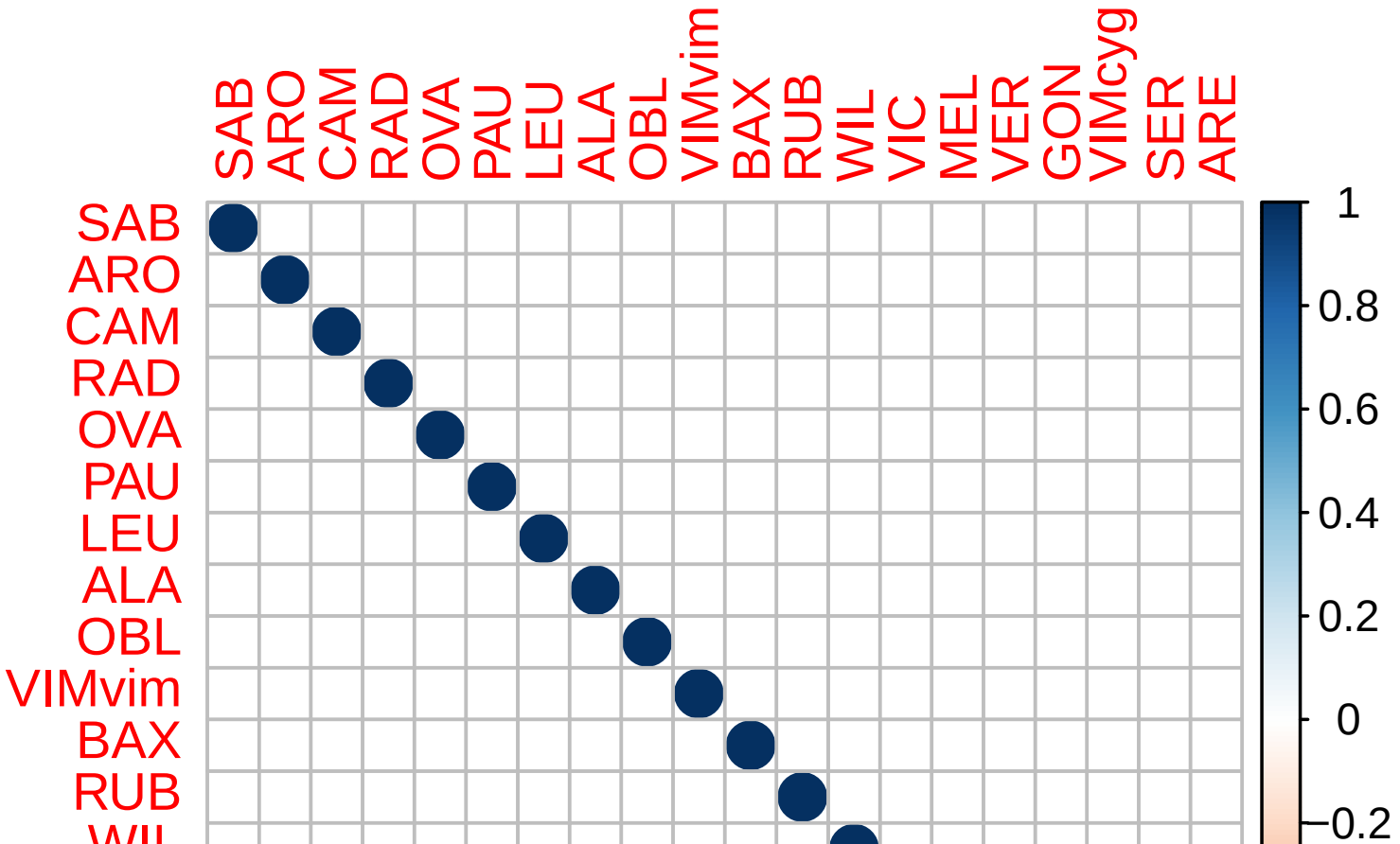
Example with Eucalyptus data

```
gllvm::randomCoefplot(fit4th, which.Xcoef = "cvTemp")
```



Example with Eucalyptus data

```
corrplot::corrplot(gllvm::getResidualCor(fit4th), order = "AO
```



Example with Eucalyptus data

Fit random effect model without traits (in dev version)

```
library(gllvm)
TMB::openmp(parallel::detectCores()-1)
```

```
## gllvm
##     7
## attr(,"autopar")
## [1] FALSE
```

```
fit <- gllvm(Y, X,
formula = ~ (0+Rockiness + Loaminess + Sandiness + cvTemp + PPTann + Vally
                    , family = "binomial", seed = 1, n.init = 3)
```

Note: such random effects can be combined with ordination methods

Example with Eucalyptus data

```
summary(fit)
```

```
##
## Call:
## gllvm(y = Y, X = X, formula = ~(0 + Rockiness + Loaminess + Sandiness +
##     cvTemp + PPTann + VallyBotFlat | 1), family = "binomial",
##     seed = 1, n.init = 3)
##
## Family:  binomial
##
## AIC:  4182.271 AICc:  4183.921 BIC:  4794.815 LL:  -2005.1 df:  86
##
## Informed LVs:  0
## Constrained LVs:  0
## Unconstrained LVs:  2
##
## Formula:  ~(0 + Rockiness + Loaminess + Sandiness + cvTemp + PPTann + VallyBotFlat | 1)
## LV formula:  ~ 0
##
## Random effects:
##  Name         Variance Std.Dev Corr
##  Rockiness    12.9712  3.6016
##  Loaminess     0.0752  0.2742  -0.7204
##  Sandiness     0.0845  0.2906  -0.0985  0.3212
##  cvTemp        0.4916  0.7011  -0.6504  0.1643  0.3765
##  PPTann        0.1714  0.4140  -0.3406  0.4595  0.8996  0.6258
##  VallyBotFlat  1.9478  1.3956   0.7632 -0.7144 -0.6878 -0.7504 -0.8653
##
## Coefficients predictors:
##                Estimate Std. Error    z value Pr(>|z|)
## Rockiness    -7.067e-01  3.790e-20 -1.865e+19   <2e-16 ***
## Loaminess    -2.261e-02  1.825e-19 -1.239e+17   <2e-16 ***
## Sandiness     3.349e-02  6.680e-20  5.014e+17   <2e-16 ***
## cvTemp       -2.376e-02  3.543e-20 -6.705e+17   <2e-16 ***
## PPTann       -6.454e-02  2.720e-19 -2.373e+17   <2e-16 ***
## VallyBotFlat -3.846e-01  1.283e-19 -2.999e+18   <2e-16 ***
```

Example with Eucalyptus data

```
anova(fit4th, fit)
```

```
## Model  1 :  y ~ NULL
## Model  2 :  ~ Rockiness + Loaminess + Sandiness + cvTemp + PPTann + Val
```

```
##   Resid.Df        D Df.diff P.value
## 1     9074   0.0000       0
## 2     9056 178.4587      18       0
```

We accept the alternative hypothesis: species responses to the environment are structured by traits

Community Phylogenetics

Models with species random effects in `gllvm` are new.

This is still in (very) active development, and only available in the development version

Phylogenetic random effects

So far, we have been discussing **unstructured** species associations
But what if we have information to provide?

Phylogenetic random effects

- In the 4th corner model $\mathbf{b}_j$ could be structured by Phylogeny
- I.e., More closely related species have similar responses to the environment
- Can also structure species-random effects

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 and LVs for brevity. So our model is:

$$\boldsymbol{\eta} = \mathbf{1}\beta_{0j}^{\top} + \mathbf{XB} \tag{6}$$

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

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

Phylogenetic random effects

$$\boldsymbol{\Sigma}_m = \mathbf{C}\rho + (1-\rho)\mathbf{I} \tag{7}$$

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

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

So: this model only generates positive species associations.

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]

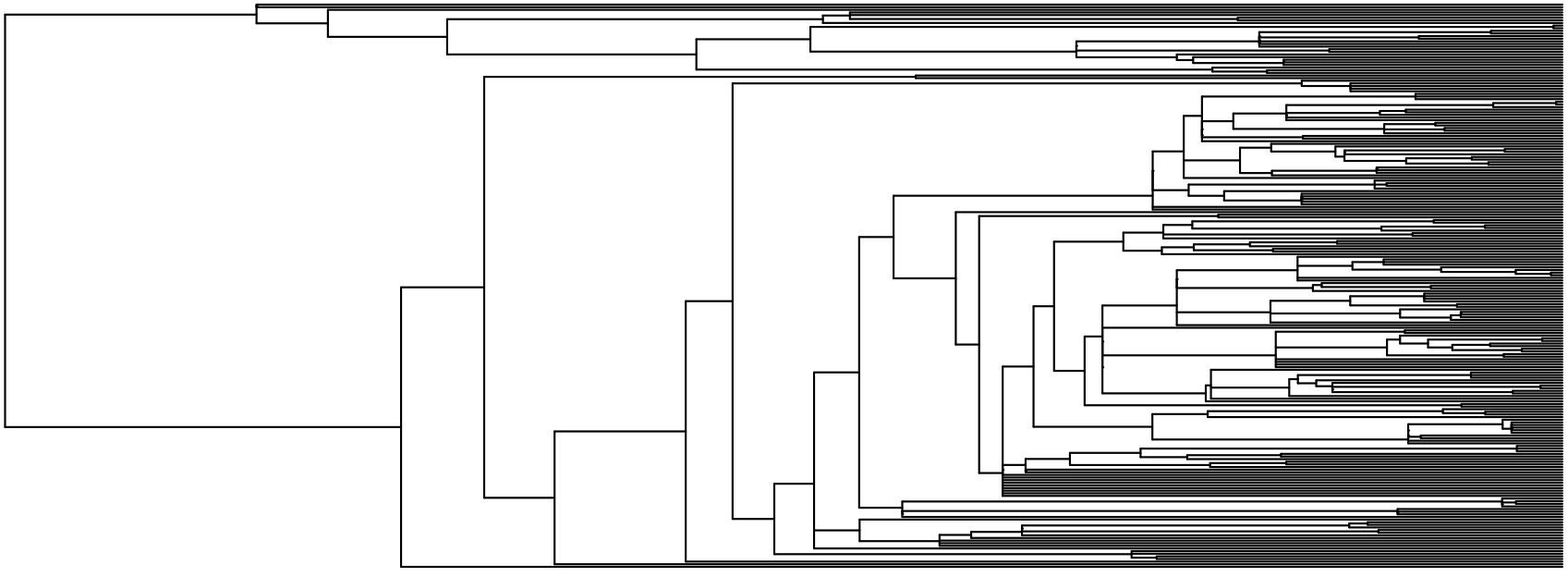
Example with fungi data

```
Y = read.csv("../data/fungiY.csv",)[,-1]
X = read.csv("../data/fungiX.csv")[,-1]
tree = ape::read.tree("../data/fungiTree.txt")
```

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

Example with fungi data

```
plot(tree, show.tip.label = FALSE)
```



Example with fungi data

```
TMB::openmp(parallel::detectCores()-1)
fita <- gllvm(y=Y, X=X, family = "binomial", num.lv = 0, sd.errors = FALSE
      row.eff = ~(1 | REGION/RESERVE), studyDesign = X[,c("REGION","RESERV
      formula = ~(0+DBH.CM+AVERDP+I(AVERDP^2)+CONNECT10+TEMPR+PRECIP+log.A
      colMat = list(ape::vcv(tree)[colnames(Y),colnames(Y)],
              dist = ape::cophenetic.phylo(tree)[colnames(Y),colnames(Y)])
                    nn.colMat = 5, max.iter=10e3, n.init = 3)
```

Example with fungi data

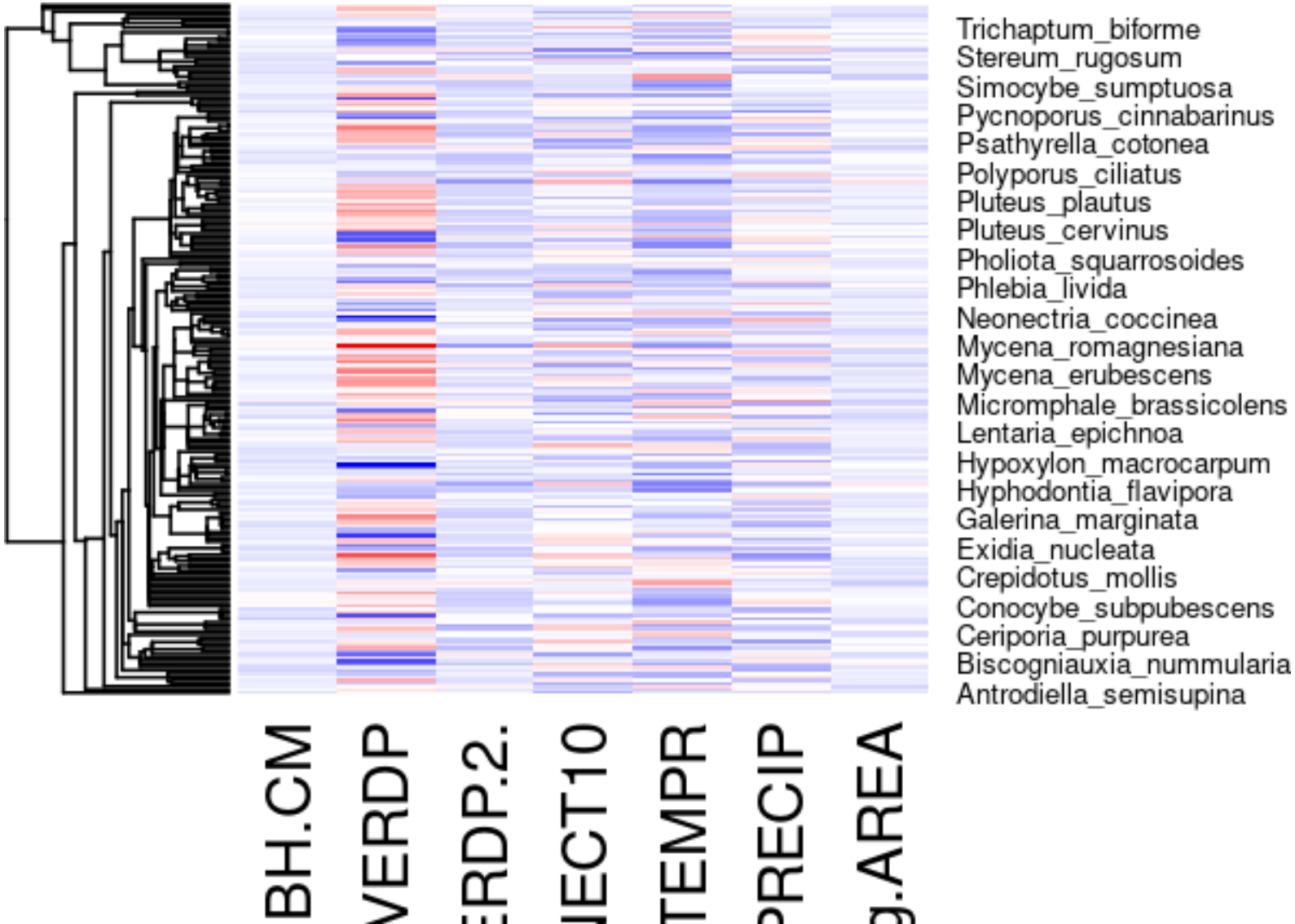
```
summary(fita)
```

```
##
## Call:
## gllvm(y = Y, X = X, formula = ~(0 + DBH.CM + AVERDP + I(AVERDP^2) +
##     CONNECT10 + TEMPR + PRECIP + log.AREA | 1), family = "binomial",
##     num.lv = 0, studyDesign = X[, c("REGION", "RESERVE")], colMat = list(ape::vcv(tree)[colnames(Y),
##         colnames(Y)], dist = ape::cophenetic.phylo(tree)[colnames(Y),
##         colnames(Y)]), row.eff = ~(1 | REGION/RESERVE), sd.errors = FALSE,
##     nn.colMat = 5, max.iter = 10000, n.init = 3)
##
## Family:  binomial
##
## AIC:  102607.7 AICc:  102608 BIC:  105337.3 LL:  -51051 df:  253
##
## Informed LVs:  0
## Constrained LVs:  0
## Unconstrained LVs:  0
##
## Formula:  ~(0 + DBH.CM + AVERDP + I(AVERDP^2) + CONNECT10 + TEMPR + PRECIP + log.AREA | 1)
## LV formula:  ~ 0
##
## Random effects:
##  Name        Signal Variance Std.Dev Corr
##  DBH.CM      0.8490 0.0069   0.0832
##  AVERDP      0.8490 0.3754   0.6127  -0.0081
##  I.AVERDP.2. 0.8490 0.1409   0.3754  -0.2610 -0.2004
##  CONNECT10   0.8490 0.1444   0.3800  -0.0128  0.2124 -0.6836
##  TEMPR       0.8490 0.5304   0.7283  -0.3190 -0.0693  0.9444 -0.4243
##  PRECIP      0.8490 0.1146   0.3386   0.4930 -0.4546  0.5745 -0.7841  0.3557
##  log.AREA    0.8490 0.0523   0.2286   0.2593  0.0339 -0.9854  0.6545 -0.9563
##
##
##
##
##
##
```

Example with fungi data

```
heatmap(t(fita$params$Br),
Rowv = phylogram::as.dendrogram(tree), Colv=NA,
col = colorRampPalette(c("blue","white","red"))(ncol(Y)), scale = "none")
```

Example with fungi data



Summary

- Ordination and JSDMs are two frameworks for analysing species co-occurrence data
- One focuses on inference, the other prediction
- The GLLVM framework is used differently in both
- We formulate models differently
- But can (and really should) learn from each other
- Model-based ordination with traits and Phylogeny to follow in the future