Generalised Linear (Mixed) Models for data of multiple species

Bert van der Veen

Department of Mathematical Sciences, NTNU

Who-is-who

Logstics

- slides, slack

Outline of this session

- GLMs assume independence
- Mixed-effects can relax that assumption
- Also allow to incorporate correlation (between species)
  - I.e., JSDM

Outline

- Data collection and common data types
- Generalised Linear Models background
  - Assumption checking
- "Vector" models
- GLMMs
  - Estimating variation in species responses
  - Estimating correlation in species responses
- Model diagnostics and selection
- Building on some material from the GLM workshop

Philosophy

- Adjust the model, not the data
- If you're not sure how to adjust the model, you just need to figure out
- Unlike "classically" where the data is adjusted to make the method fit

Classical approach

Classically:

1) Decide on a distance
2) Apply a multivariate analysis
3) Make a plot
4) Do a hypothesis test

Plant Ecol (2015) 216:669–682
DOI 10.1007/s11258-014-0366-3

Model-based thinking for community ecology

David I. Warton · Scott D. Foster · Glenn De'ath · Jakub Stoklosa · Piers K. Dunstan

Statistical modeling

Instead of a focus on data, we consider the *data generating process*

- We collect data
- Decide on a research question for *the population*
- Learn about the variation in the data
  - Which requires formulating a model
- Work out distribution of the estimates
  - And find the "best" estimate
- Conclude if our answer is robust for the population

Collecting data

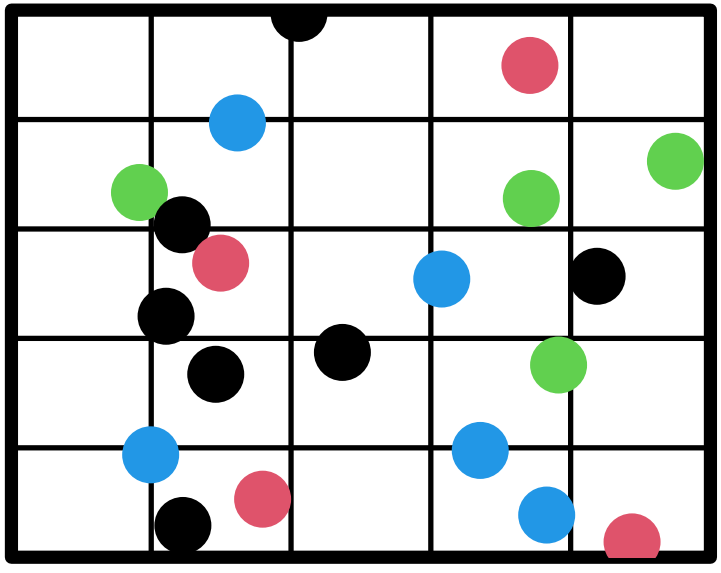
-maybe make this about data properties - we can design the ideal dataset and it would always give us the correct answer - sometimes, things don't work out so well and there are properties that "hide" the true ecological process

Data collection

Figure 1: dw.com

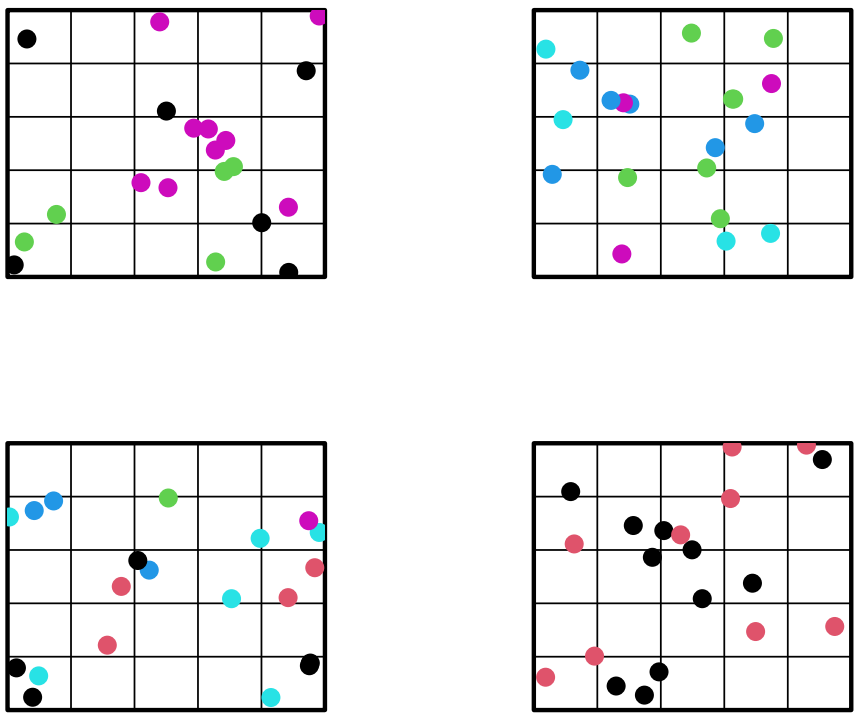
A meadow in the Dutch dunes.

How many plant species are there in this meadow?



We go into the meadow and count plants in quadrats. We find 4 species.

Resampling the dunes



We resample the dune meadow, and find different numbers of species: 3, 3, 6, 2. And different compositions. On average we have found 3.5 species per quadrat.

Sampling variation

1) Each time we sample, we have a slightly different sample
2) Each time we estimate a parameter, it might be slightly different due to this sampling variation
3) The more data we sample, the better we can represent this variability of our estimate
4) And the better we can infer the ecological process that generates patterns of richness/cover in the meadow

The ecological process

What do we know of the processes that generate these data?

- Meta-community theory
- Assembly processes (filtering)
- Ecological gradient theory

The ecological process

What do we know of the processes that generate these data?

- Meta-community theory
- Assembly processes (filtering)
- Ecological gradient theory

Multispecies models provide a statistical connection to these ecological frameworks. We do not just use a fancy tool, we use a fancy tool because we believe it aligns well with our understanding of the ecological process.

The ecological process (2)

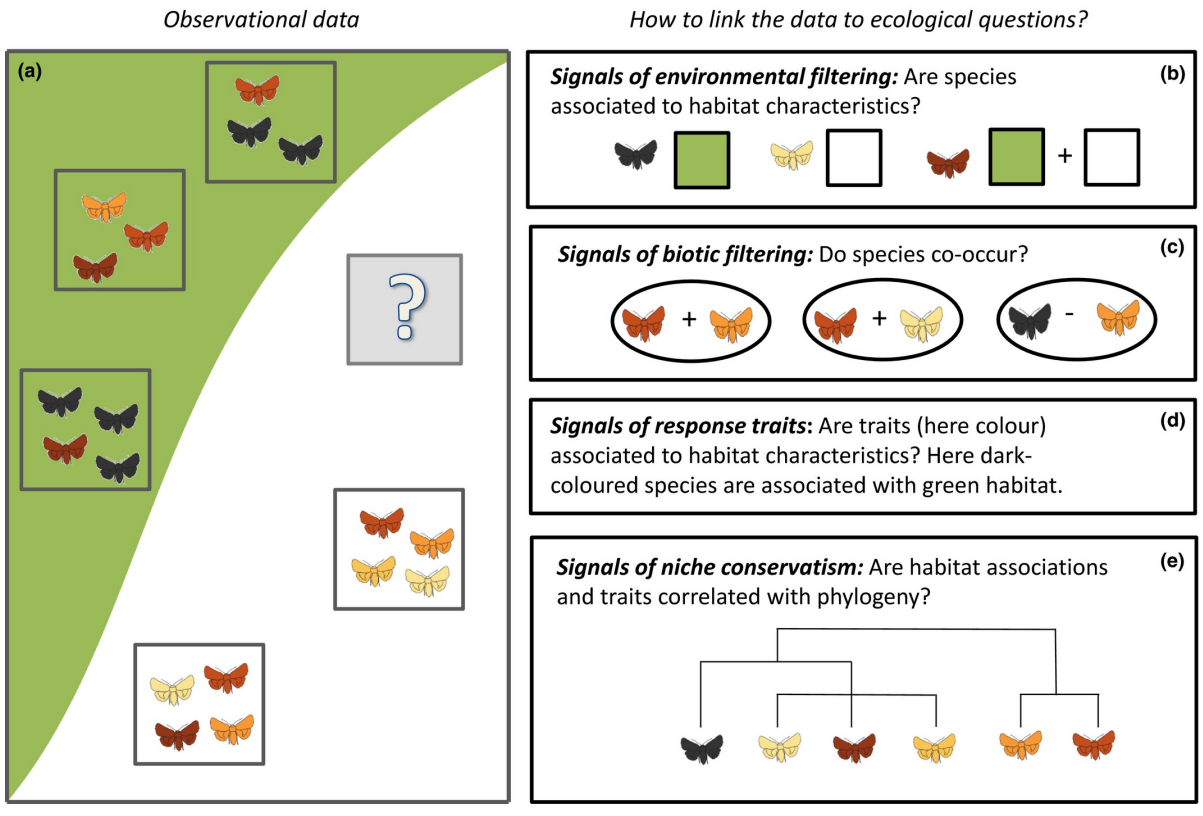


Figure 2: Figure 2 from Ovaskainen et al. (2017)

On ecological communities

The concept of an ecological community is of limited use. By definition:

An ecological community is a group or association of two or more species occupying the same geographical area at the same time

- We often think of ecological communities as groups
- We can also think of a community as a continuum that changes along a gradient (Austun 1985)
- We can also think of them as the species in our data

On ecological communities

The concept of an ecological community is of limited use. By definition:

An ecological community is a group or association of two or more species occupying the same geographical area at the same time

- We often think of ecological communities as groups
- We can also think of a community as a continuum that changes along a gradient (Austun 1985)
- We can also think of them as the species in our data

Connecting model outputs to ecological concepts requires some deep thoughts

Multispecies models

There are multiple statistical frameworks for studying the processes:

- Generalised Linear Models
- Generalised Linear Mixed-effects Models
- Generalised Additive Models (not covered here)
- Generalised Linear Latent Variable Models

Multispecies models

There are multiple statistical frameworks for studying the processes:

- Generalised Linear Models
- Generalised Linear Mixed-effects Models
- Generalised Additive Models (not covered here)
- Generalised Linear Latent Variable Models

Or ecologically:

- Species distribution models
- Joint Species Distribution Models
- Ordination

Multispecies models

There are multiple statistical frameworks for studying the processes:

- Generalised Linear Models
- Generalised Linear Mixed-effects Models
- Generalised Additive Models (not covered here)
- Generalised Linear Latent Variable Models

Or ecologically:

- Species distribution models
- Joint Species Distribution Models
- Ordination

and more. Each method has its limitations (assumptions). It is up to us to assess which are appropriate.

Generalised linear models (GLMs)

GLMs as a framework were introduced by Nelder and Wedderburn (1972) uniting many different models. With a special focus on teaching statistics.

- Linear regression
- Logistic regression
- Probit regression
- Complementary log-log regression
- Log-linear regression
- Gamma regression

Generalised linear models

GLMs extend the linear model framework to address:

- Variance changes with the mean
- Range of **y** is bounded

The basis of many statistical models in Ecology

Components of a GLM

- Systematic component: η
- Random component: data/distribution)
- The link function: connects these components
  - This is not a data transformation
- The variance function

GLM Likelihood

- We use MLE for estimation
- With a distribution in the "exponential family" (for fixed ϕ)

All GLMs have the likelihood:

$$\mathcal{L}(y_i; \Theta) = \exp\{\frac{y_i \eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\} \quad (1)$$

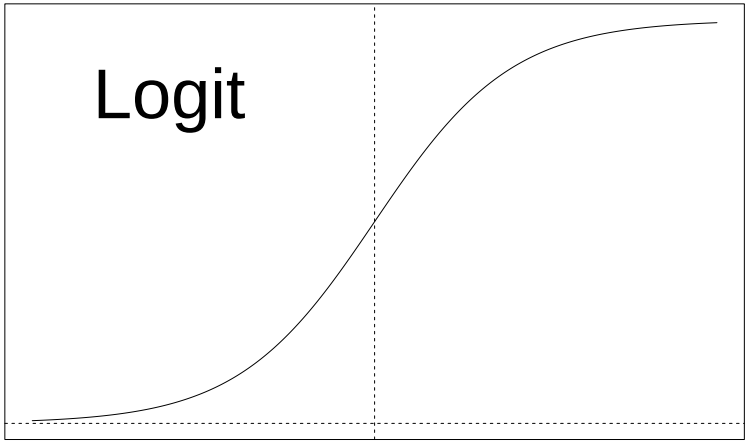
Generalised linear model

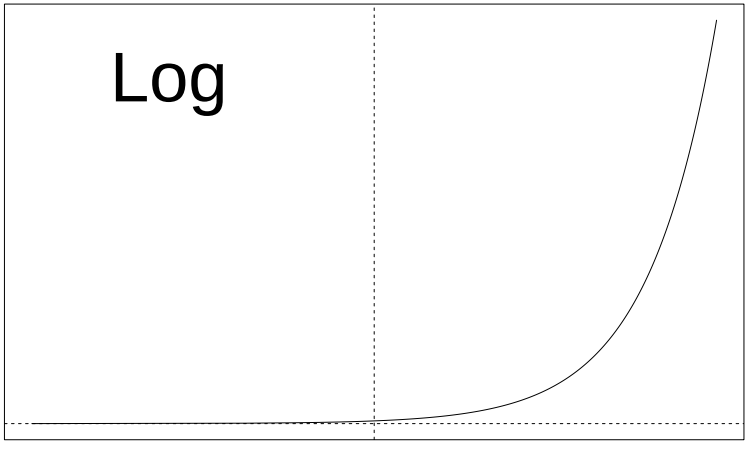
$$
\begin{aligned}
g\{\mathbb{E}(y_i|x_i)\} &= \eta_i = \alpha + x_i\beta \\
\mathbb{E}(y_i|x_i) &= g^{-1}(\eta_i) = g^{-1}(\alpha + x_i\beta)
\end{aligned}
\tag{2}
$$

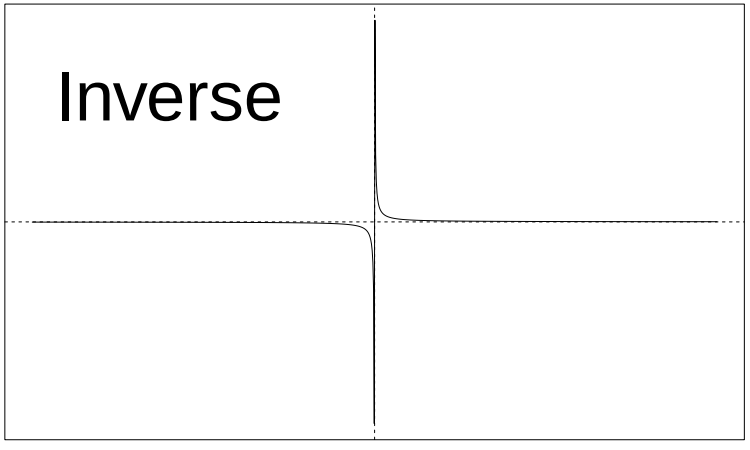
g$(\cdot)$ is the **link function**

The link function

- Is a smooth/monotone function
- Has an inverse $g^{-1}(\cdot)$
- Restricts the scale
- g$(\cdot)$ can be e.g.







Assumptions

- No outliers
- Independence
- Correct distribution (in the EF)
- Correct link function
- Correct variance function (implied by previous two)

We can check the validility of these assumptions with residual plots.

Often used distributions in ecology

- Binomial: occurrence/counts. Presence of species, number of germinated seeds out of a total
- Poisson: counts. Abundance
- Negative binomial (fixed dispersion): counts. Number of species or abundance
- Gamma: (positive) continuous. Body size or biomass
- Ordinal (cumulative link). Cover classes
- Beta (logit link). Cover (note: not a GLM)
  - can be challenging to deal with
- And so on

Using GLMs

We use GLMs to predict or infer species' whereabouts with respect to environmental drivers

1) Do species respond to an environmental gradient?
2) Do species increase or decrease along an environmental gradient?
3) Which environmental gradients best represent the pool of species that we have observed?

Example: macroinvertebrate counts in USA desert

Observation process: count of macroinvertebrates in three "dips"
Alternatively: The proportion of a species in a dip
Alternatively: Was this species found in the dip

There are often many ways to observe the same ecological process. We need to **disentangle** this from the ecological process.

Example: macroinvertebrate counts in USA desert

- Data by Pina and Lougheed 2022
- Counts of 14 species, in 2018 and 2019, in 14 wetlands
- Main goal: assess impacts of water quality on macroinvertebrates

The abundance data

Ephemeroptera	Hemiptera	Amphipoda	Gastropoda	Anispotera	Zygoptera	Decapoda	Diptera	Chiron
0	1	100	8	0	1	12	0	
21	0	0	5	3	10	0	1	
11	3	5	0	4	32	0	0	
0	1	11	0	0	0	0	0	
80	1	60	25	0	7	0	2	
9	15	0	6	0	10	0	1	
30	0	0	10	7	25	0	1	
10	1	60	190	4	60	0	0	
6	0	0	5	0	3	2	2	
5	41	0	5	0	2	0	0	
32	2	0	0	1	35	0	0	
6	25	0	31	2	13	0	0	
28	15	400	0	21	14	0	0	
28	15	400	0	21	14	0	0	
6	150	200	70	2	48	3	4	
6	150	200	70	2	48	3	4	
9	1	0	35	14	50	0	7	
26	1	0	1	10	21	0	12	
0	1	0	2	2	7	0	0	
1	0	60	6	1	9	0	0	
13	0	15	26	0	8	0	1	
87	4	0	0	0	3	0	0	

The environment data

Year	Hydro	Water_Type	Conductivity	DOC	TDN	Turbidity	Alkalinity	Total_CHL	Correc
2018	permanent	non-waste	4.060	2.846	0.306	4.40	63.050	2.231	
2018	permanent	waste	2.582	23.160	3.544	60.73	412.400	95.211	
2018	permanent	non-waste	8.563	28.120	2.450	17.40	363.708	16.915	
2018	permanent	non-waste	15.710	75.040	7.160	38.50	457.625	26.160	
2018	ephemeral	non-waste	1.029	4.012	0.386	3.78	198.042	12.657	
2018	ephemeral	non-waste	1.204	5.356	0.491	24.70	168.042	30.353	

- 11 environmental variables
  - Water chemistry
  - Water type
  - Presence of hydro power
  - Permanent or temporary wetland

Data format

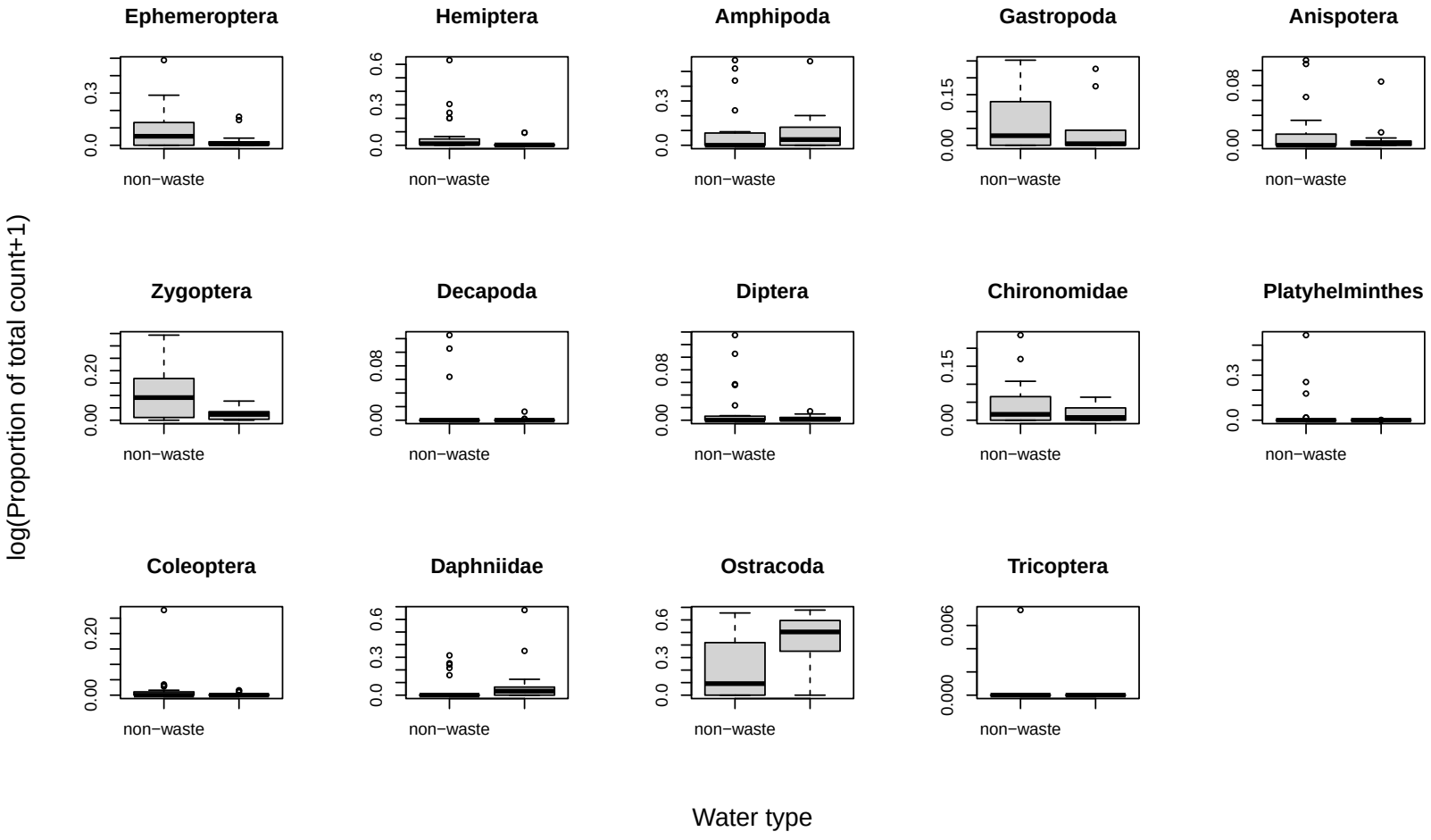
There are two ways to format these data:
Wide format: Species as columns (as presented)
Long format: Species is one column, and "Site" is another column

The format of the data does not affect the model. Some functions accept long format, other wide format, but the formulation of the model is up to us.

To long format

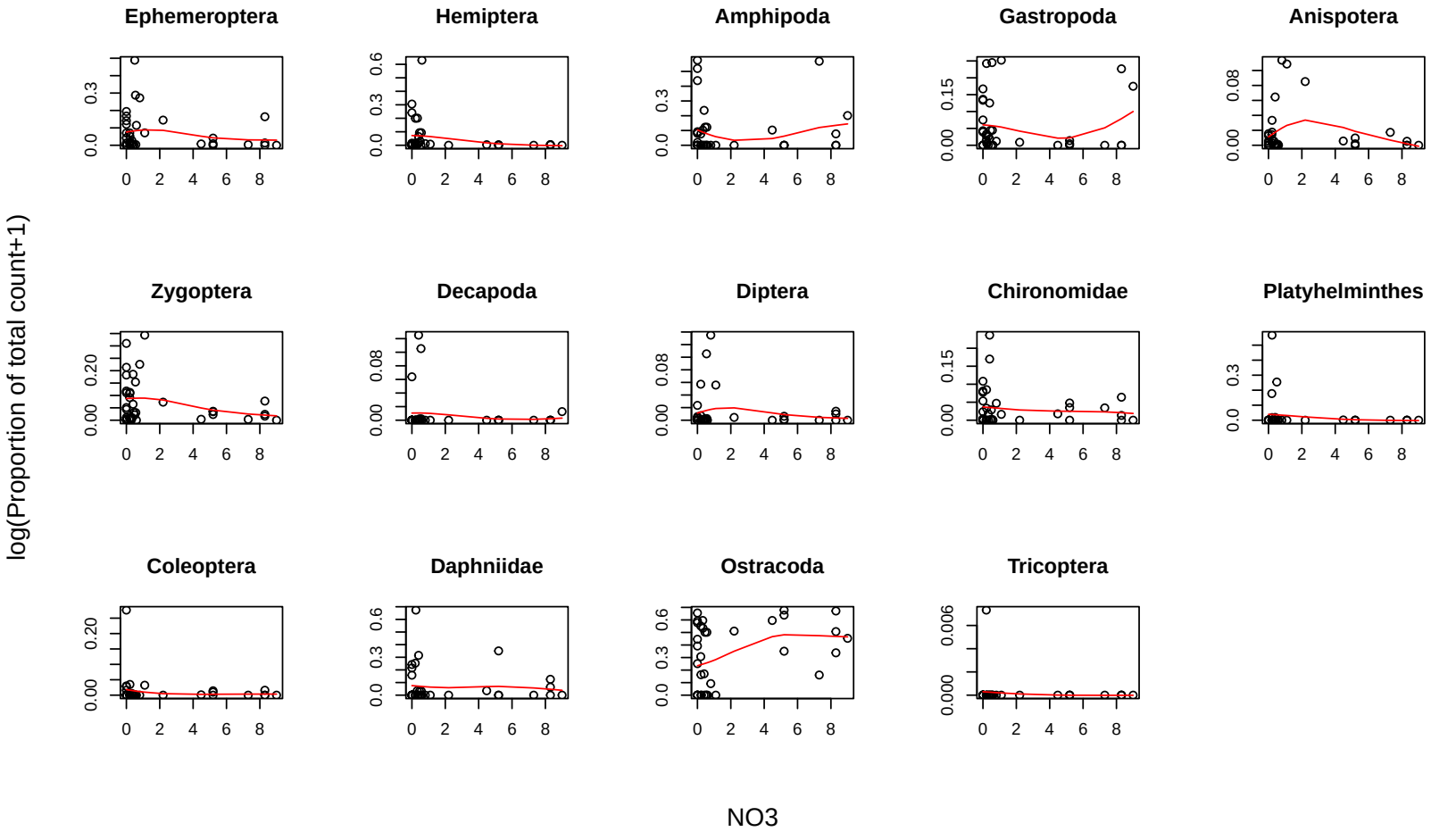
```
dat <- data.frame(y, X, total = rowSums(y))
long <- reshape(dat,
                    varying = colnames(y),
                    v.names = "Count",
                    idvar = "Site",
                    timevar = "Species",
                    direction = "long")

long$Species <- factor(long$Species,
                           labels = colnames(y))
```

Visually inspect the data: categorical covariate



Is there a common effect?

Visually inspect the data: continuous covariate



Is there a common effect?

Wetlands: fit a model

```
model1 <- glm(Count~NO3,
              data = long, family = "poisson")
coef(model1)
```

```
## (Intercept)         NO3
##  3.46872706  0.05872844
```

Intercept and NO3 effect are the same for all species

Multispecies modeling

1) Is the same effect for all species realistic?
2) Is the same (average) count for all species realistic?

Multispecies modeling

1) Is the same effect for all species realistic?
2) Is the same (average) count for all species realistic?
3) We usually assume that species have their own preferred environmental conditions
4) Some species might still like similar conditions; there is a common component
5) We can separate this out with GLMMs or with a "sum-to-zero" contrast

Wetlands: species-specific effects

```
model2 <- glm(Count~NO3:Species, data = long, family="poisson")
coef(model2)
```

```
##                (Intercept)   NO3:SpeciesEphemeroptera
##                 3.69018927                -0.17223172
##       NO3:SpeciesHemiptera      NO3:SpeciesAmphipoda
##                -0.54664249                 0.07189136
##      NO3:SpeciesGastropoda     NO3:SpeciesAnispotera
##                -0.05126546                -0.59016596
##       NO3:SpeciesZygoptera       NO3:SpeciesDecapoda
##                -0.14102521                -6.04052710
##         NO3:SpeciesDiptera   NO3:SpeciesChironomidae
##                -1.75086879                -0.17355409
## NO3:SpeciesPlatyhelminthes     NO3:SpeciesColeoptera
##                -3.13020126                -0.69435033
##      NO3:SpeciesDaphniidae      NO3:SpeciesOstracoda
##                 0.03260345                 0.33691759
##      NO3:SpeciesTricoptera
##               -25.56717960
```

```
mean(coef(model2)[-1]) # average slope
```

Wetlands: species-specific responses

```
summary(model2)
```

```
##
## Call:
## glm(formula = Count ~ NO3:Species, family = "poisson", data = long)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  3.690189   0.009294 397.040  < 2e-16 ***
## NO3:SpeciesEphemeroptera    -0.172232   0.013926 -12.368  < 2e-16 ***
## NO3:SpeciesHemiptera        -0.546642   0.041808 -13.075  < 2e-16 ***
## NO3:SpeciesAmphipoda         0.071891   0.005835  12.320  < 2e-16 ***
## NO3:SpeciesGastropoda       -0.051265   0.009121  -5.621 1.90e-08 ***
## NO3:SpeciesAnispotera       -0.590166   0.046182 -12.779  < 2e-16 ***
## NO3:SpeciesZygoptera        -0.141025   0.012512 -11.271  < 2e-16 ***
## NO3:SpeciesDecapoda         -6.040527   0.369560 -16.345  < 2e-16 ***
## NO3:SpeciesDiptera          -1.750869   0.132832 -13.181  < 2e-16 ***
## NO3:SpeciesChironomidae     -0.173554   0.013989 -12.407  < 2e-16 ***
## NO3:SpeciesPlatyhelminthes  -3.130201   0.198059 -15.804  < 2e-16 ***
## NO3:SpeciesColeoptera       -0.694350   0.057031 -12.175  < 2e-16 ***
## NO3:SpeciesDaphniidae        0.032603   0.006735   4.841 1.29e-06 ***
## NO3:SpeciesOstracoda         0.336918   0.002326 144.875  < 2e-16 ***
## NO3:SpeciesTricoptera      -25.567180   4.779080  -5.350 8.80e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Contrasts

There are other "contrast" treatments in R than "dummy" - We can instead use "sum-to-zero" contrasts - If the sum is zero, the mean must be too - The coefficient of the last species is set to the negative sum

```
(contr <- contr.sum(levels(long$Species)))
```

```
##                [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
## Ephemeroptera     1    0    0    0    0    0    0    0    0     0     0     0
## Hemiptera         0    1    0    0    0    0    0    0    0     0     0     0
## Amphipoda         0    0    1    0    0    0    0    0    0     0     0     0
## Gastropoda        0    0    0    1    0    0    0    0    0     0     0     0
## Anispotera        0    0    0    0    1    0    0    0    0     0     0     0
## Zygoptera         0    0    0    0    0    1    0    0    0     0     0     0
## Decapoda          0    0    0    0    0    0    1    0    0     0     0     0
## Diptera           0    0    0    0    0    0    0    1    0     0     0     0
## Chironomidae      0    0    0    0    0    0    0    0    1     0     0     0
## Platyhelminthes   0    0    0    0    0    0    0    0    0     1     0     0
## Coleoptera        0    0    0    0    0    0    0    0    0     0     1     0
## Daphniidae        0    0    0    0    0    0    0    0    0     0     0     1
## Ostracoda         0    0    0    0    0    0    0    0    0     0     0     0
## Tricoptera       -1   -1   -1   -1   -1   -1   -1   -1   -1    -1    -1    -1
##                [,13]
```

Wetlands: species-specific responses with common effect

```
model3 <- glm(Count~NO3+NO3:Species, data = long, family = "poisson", contrasts = list(Species = contr))
```

```
## Warning: glm.fit: fitted rates numerically 0 occurred
```

```
coef(model3)
```

```
##   (Intercept)           NO3  NO3:Species1  NO3:Species2  NO3:Species3
##     3.6901893    -2.7440428     2.5718111     2.1974003     2.8159342
##  NO3:Species4  NO3:Species5  NO3:Species6  NO3:Species7  NO3:Species8
##     2.6927774     2.1538769     2.6030176    -3.2964843     0.9931740
##  NO3:Species9 NO3:Species10 NO3:Species11 NO3:Species12 NO3:Species13
##     2.5704887    -0.3861584     2.0496925     2.7766463     3.0809604
```

- The intercept is the same for all species
- One NO3 effect that is the same for all species (the mean of effects)
- One NO3 effect per species, relative to the common effect

Wetlands: species-specific responses with common effect

```
##
## Call:
## glm(formula = Count ~ NO3 + NO3:Species, family = "poisson",
##     data = long, contrasts = list(Species = contr))
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.690189   0.009294 397.040  < 2e-16 ***
## NO3           -2.744043   0.342989  -8.000 1.24e-15 ***
## NO3:Species1   2.571811   0.343176   7.494 6.67e-14 ***
## NO3:Species2   2.197400   0.345025   6.369 1.90e-10 ***
## NO3:Species3   2.815934   0.342991   8.210  < 2e-16 ***
## NO3:Species4   2.692777   0.343047   7.850 4.17e-15 ***
## NO3:Species5   2.153877   0.345485   6.234 4.54e-10 ***
## NO3:Species6   2.603018   0.343132   7.586 3.30e-14 ***
## NO3:Species7  -3.296484   0.483976  -6.811 9.67e-12 ***
## NO3:Species8   0.993174   0.363907   2.729  0.00635 **
## NO3:Species9   2.570489   0.343178   7.490 6.87e-14 ***
## NO3:Species10 -0.386158   0.388415  -0.994  0.32013
## NO3:Species11  2.049693   0.346826   5.910 3.42e-09 ***
## NO3:Species12  2.776646   0.343004   8.095 5.72e-16 ***
## NO3:Species13  3.080960   0.342960   8.983  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 98883  on 517  degrees of freedom
```

Wetlands: conclusions

We can conclude that NO3 has, **on average**, a negative effect on our species pool
Some species are more negatively affected than the average, some more positive
Some species are positively affected by NO3, but most positively

Wetlands

The two models have the same number of parameters, but are just differently parameterized. So, their log-likelihoods and AIC are the same:

```
AIC(model2, model3)
```

```
##         df      AIC
## model2 15 77231.31
## model3 15 77231.31
```

Interpreting the coefficients

```
##                (Intercept)   NO3:SpeciesEphemeroptera
##                 3.69018927                -0.17223172
##       NO3:SpeciesHemiptera       NO3:SpeciesAmphipoda
##                 -0.54664249                 0.07189136
##      NO3:SpeciesGastropoda      NO3:SpeciesAnispotera
##                 -0.05126546                -0.59016596
##       NO3:SpeciesZygoptera        NO3:SpeciesDecapoda
##                 -0.14102521                -6.04052710
##         NO3:SpeciesDiptera    NO3:SpeciesChironomidae
##                 -1.75086879                -0.17355409
## NO3:SpeciesPlatyhelminthes      NO3:SpeciesColeoptera
##                 -3.13020126                -0.69435033
##      NO3:SpeciesDaphniidae       NO3:SpeciesOstracoda
##                  0.03260345                 0.33691759
##      NO3:SpeciesTricoptera
##                -25.56717960
```

- Negative means a decrease in the response and positive increase
- More specifically here: the coefficient is multiplicative decrease in `exp(intercept)` for a unit change in `NO3`
- E.g., for "Ephemeroptera": `exp(3.69)*exp(-0.17) = exp(3.69)/exp(0.17) = 40*0.84`

Interpreting the coefficients

Or with `predict`:

```
predict(model2, newdata =
          data.frame(Species = factor("Ephemeroptera", levels = colnames(y)), NO3 = 1),
        type = "response")
```

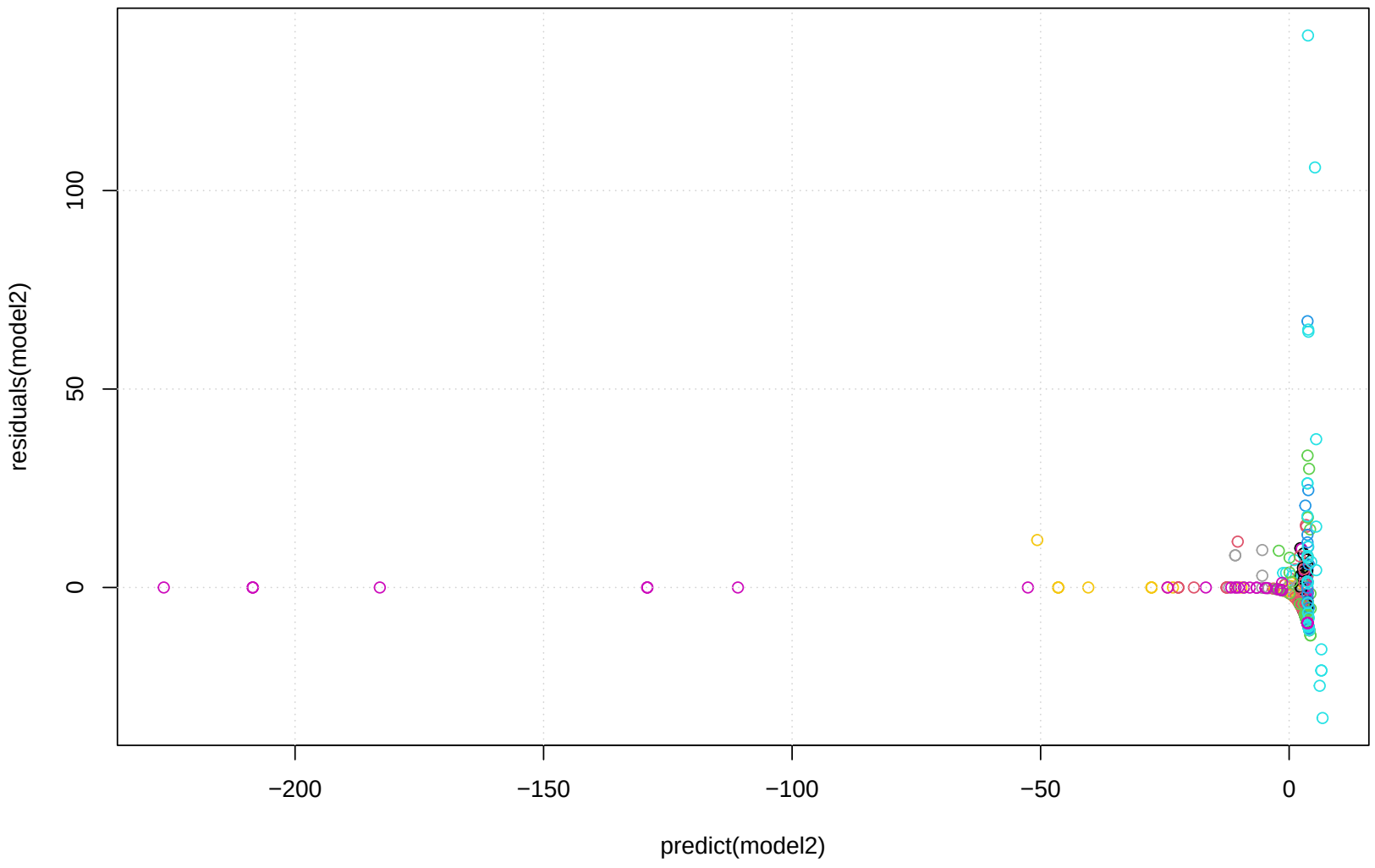
```
##       1
## 33.7155
```

Assumptions

We have made various assumptions that we should (technically) be checking:

1) Poisson distributed responses (we could change to negative-binomial)
2) Same intercept and/or slopes for species (adjust the model structure)
3) Same dispersion for species (move to vector GLM)
4) Species responses are independent (move to JSDM)

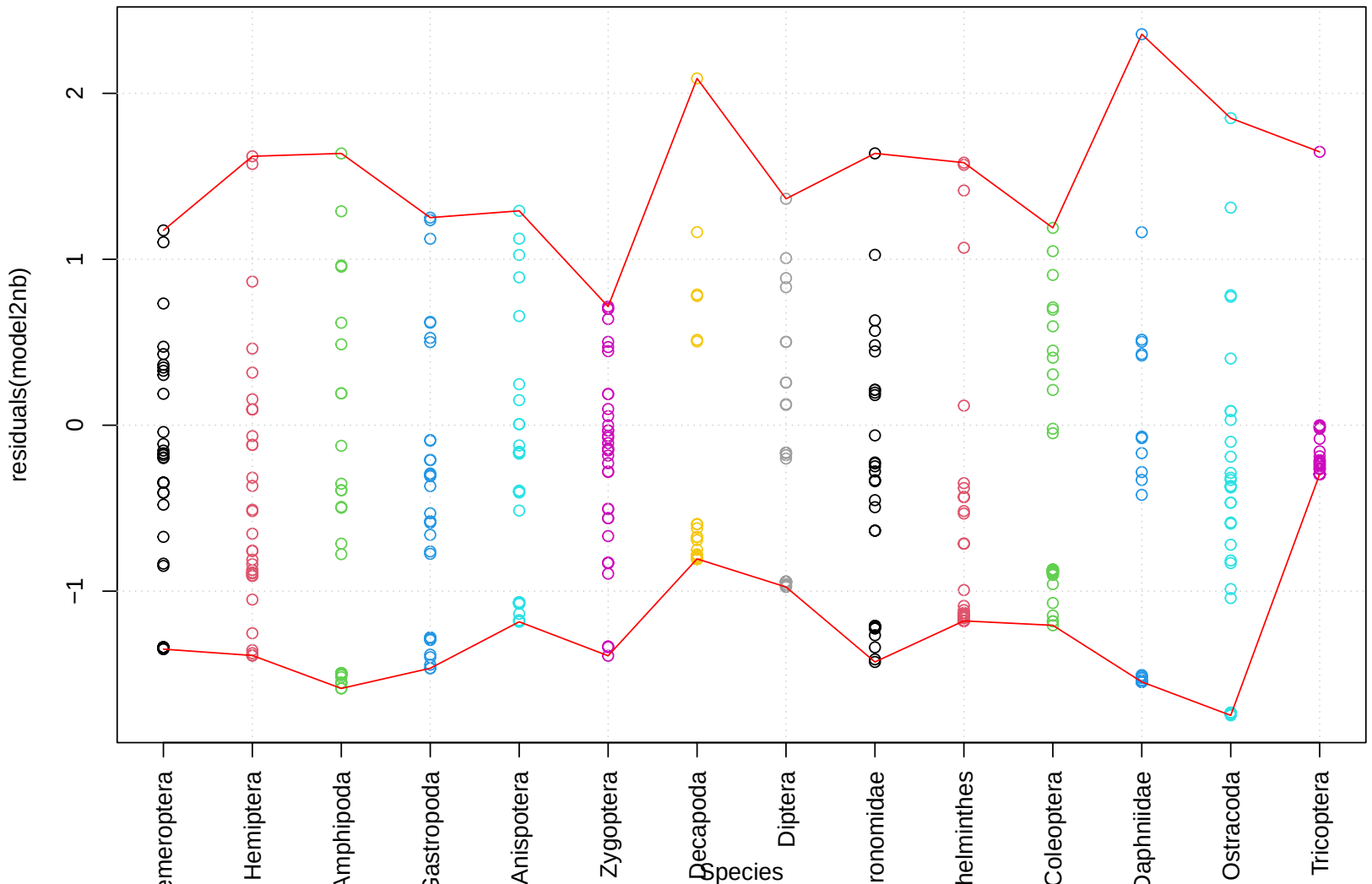
Residual diagnostics



Residual diagnostics

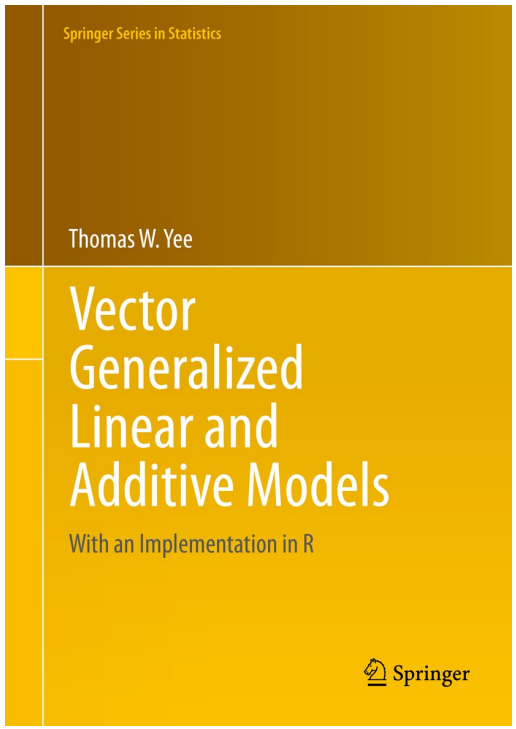
```
model2nb <- MASS::glm.nb(Count~Species + NO3:Species,
                         data = long)
```

Residual diagnostics



Vector GLMs

- One GLM per species
- Each gets their own dispersion parameter
- Slightly more flexible than what we have done so far



Methods in Ecology and Evolution
British Ecological Society
Methods in Ecology and Evolution 2012, **3**, 471–474 doi: 10.1111/j.2041-210X.2012.00190.x

mvabund – an R package for model-based analysis of multivariate abundance data

Yi Wang[1,2], Ulrike Naumann[1], Stephen T. Wright[1], and David I. Warton[1,3*]

Fitting vector GLMs

A few software implementations exist:

- The `VGAM` `R`-package
- The `glmmTMB` `R`-package
- The `gllvm` `R`-package

Clearly, we will use the last one.

VGLM Likelihood

- We use MLE for estimation
- With a distribution in the "exponential family" (for fixed ϕ)

All GLMs have the likelihood:

$$\mathcal{L}(y_{ij};\Theta) = \exp\{\frac{y_{ij}\eta_{ij} - b(\eta_{ij})}{a(\phi_j)} + c(y_{ij},\phi_j)\} \tag{3}$$

So, now we have ϕ_j instead of ϕ

gllvm

Received: 7 May 2019 | Accepted: 5 September 2019

DOI: 10.1111/2041-210X.13303

APPLICATION

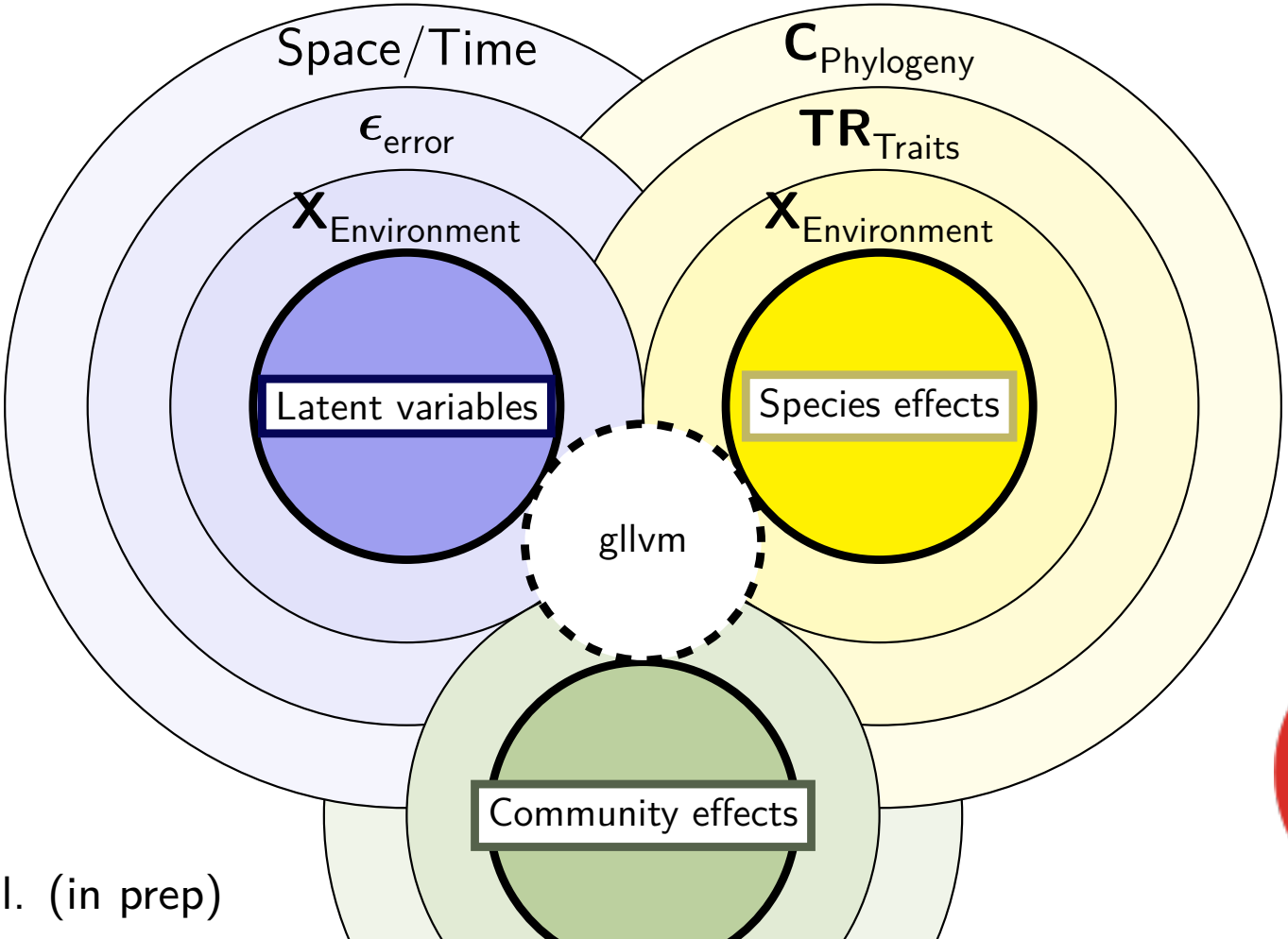
Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R

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

- Originally published in 2019 by Niku et al. I "joined in" shortly after
- For model-based multivariate analysis of community ecological data
- Models are fitted in C++ (Kristensen et al. 2015)
- Can fit many different models: VGLM(M), JSDM, and ordination

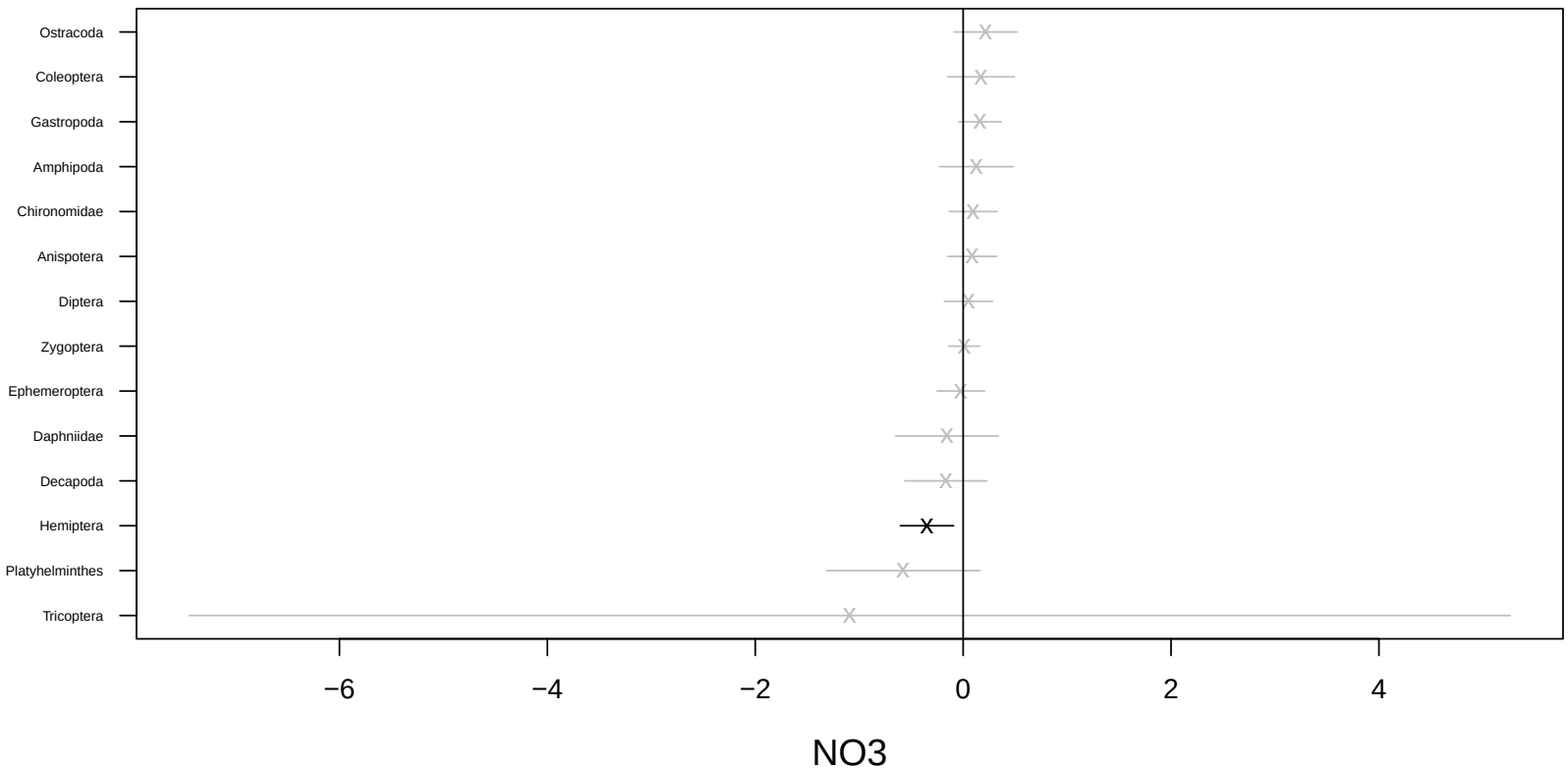
gllvm 2.0



Korhonen et al. (in prep)

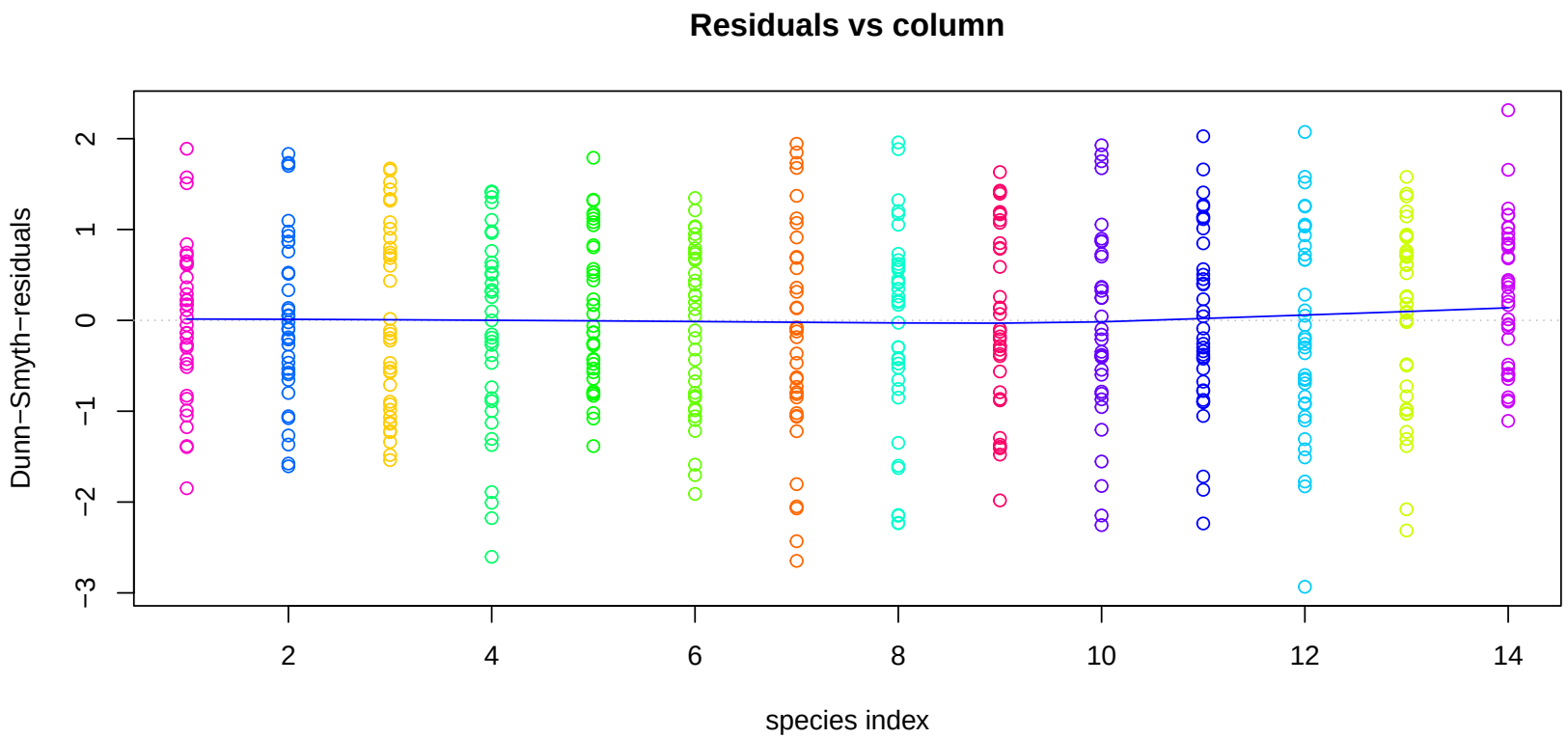
VGLM in gllvm

```
model4 <- gllvm::gllvm(y, X = X, formula = ~NO3, family = "negative.binomial")
gllvm::coefplot(model4)
```



VGLM in gllvm

```
plot(model4, which = 4)
```



Multispecies mixed-effects models

When to include a random effect:

- Unobserved effect
- To account for pseudo replication
- Nuisance
- To induce correlation
- Shrinkage

Random effects in `gllvm`

In the `gllvm` R-package there are three formula interfaces:

- `row.eff`
- `formula`
- `lv.formula`

Random effects in `gllvm`

In the `gllvm` R-package there are three formula interfaces:

- `row.eff` : for species-common fixed/random effects
- `formula`
- `lv.formula`

Random effects in `gllvm`

In the `gllvm` R-package there are three formula interfaces:

- `row.eff` : for species-common fixed/random effects
- `formula` : for species-specific fixed/random effects
- `lv.formula`

Random effects in `gllvm`

In the `gllvm` R-package there are three formula interfaces:

- `row.eff` : for species-common fixed/random effects
- `formula` : for species-specific fixed/random effects
- `lv.formula` : for effects in the ordination

For now, we focus on ‘formula‘

Random effects in `gllvm`

Our model is of the form:

$$\eta_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta}_j \tag{4}$$

but now, $\boldsymbol{\beta}_j$ is a random effect (intercept or slope). Specifically, $\boldsymbol{\beta}_j \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

Random effects in `gllvm`

Our model is of the form:

$$\eta_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta}_j \tag{4}$$

but now, $\boldsymbol{\beta}_j$ is a random effect (intercept or slope). Specifically, $\boldsymbol{\beta}_j \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

- $\boldsymbol{\mu}$: the "common effect" from before
- $\boldsymbol{\Sigma}$: variation in species random effects

The "common effect" from before

These are controlled with `row.eff`

- `row.eff` is a mixed-effects formula
- `row.eff = ~1` omits the common effects
- `row.eff = "random` incorporates row-specific random effects
- `row.eff = (1|group) + NO3` is a random effect and a fixed effect
- Can also incorporate spatial or temporal random effects

Turnover

When we use random effects, we induce correlation. Here, between site i and k:

$$\text{cov}(\mathbf{x}_i^\top \boldsymbol{\beta}_j, \mathbf{x}_k^\top \boldsymbol{\beta}_j) = \mathbf{x}_i^\top \Sigma \mathbf{x}_k, \qquad \forall j = 1 \dots m \tag{5}$$

So, Σ is a also a measure of how much the community changes from one site to another. Two sites with strong negative correlation are predicted to have different communities, and the same community with strong positive correlation. Thus, we can **predict** turnover.

Random effects `R` formula in `gllvm`

Now some examples of how it works in `R`. Generally:

```
formula = ~ (0 + continuous | categorical)
```

(the 0 is to omit an intercept term)

"Nested":

`formula = ~ (1|a/b)` is the same as `formula = ~ (1|a:b + b)`

"Crossed":

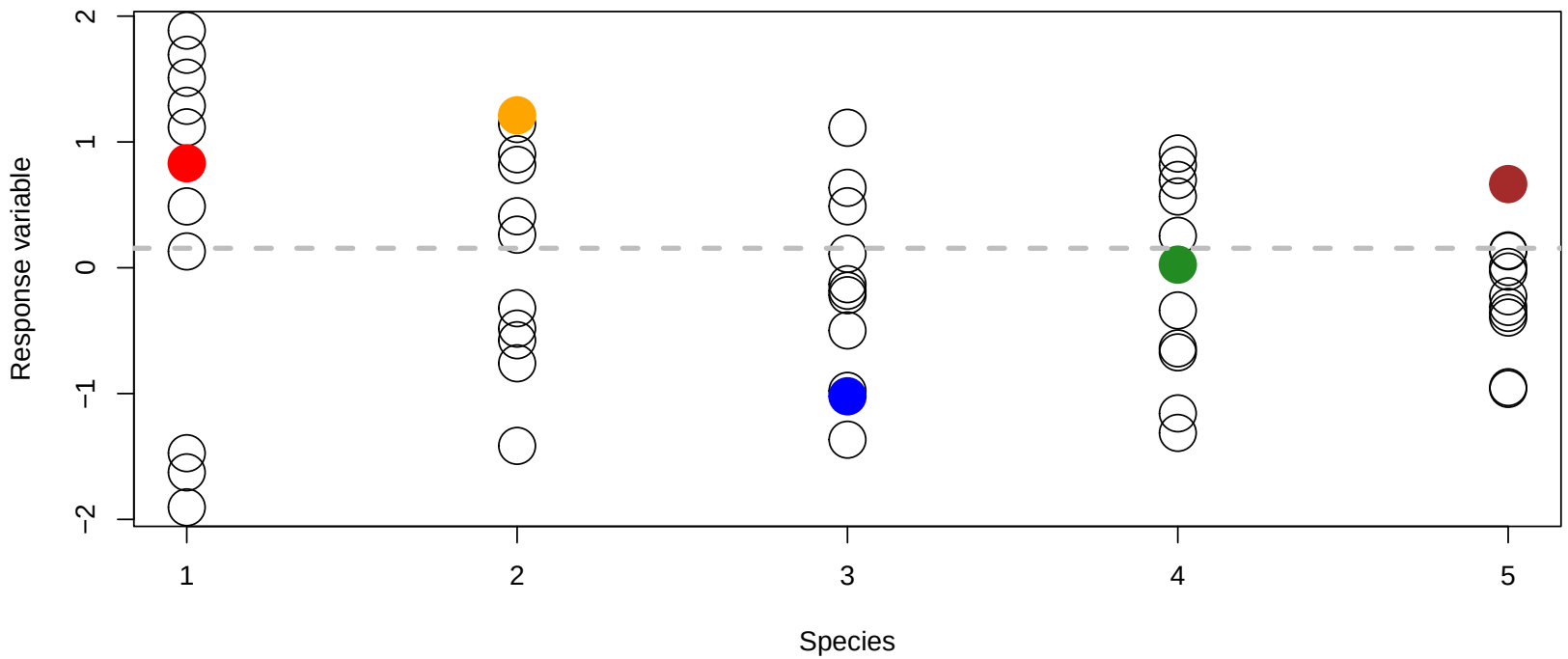
```
y ~ (1|a) + (1|b)
```

Effects within the same brackets are assumed to be correlated

Variation in mean abundance

$y_{ij} = \alpha_j, \qquad$ with $\alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma^2_\alpha)$

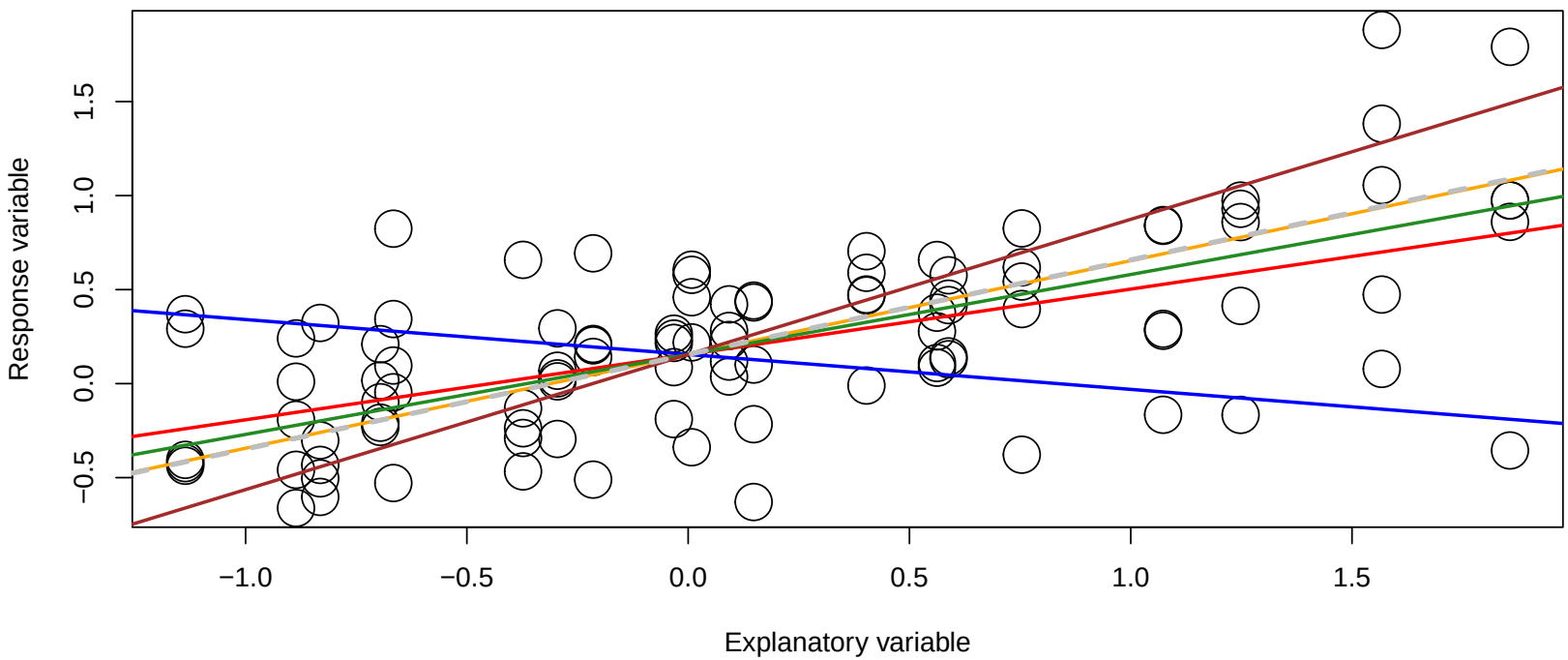
```
formula = ~ (1|1), beta0com = TRUE
```



Variation in environmental responses

$$y_{ij} = \mu_\alpha + x_i \beta_j, \qquad \text{with } \beta_j \sim \mathcal{N}(\mu_\beta, \sigma^2_\beta)$$

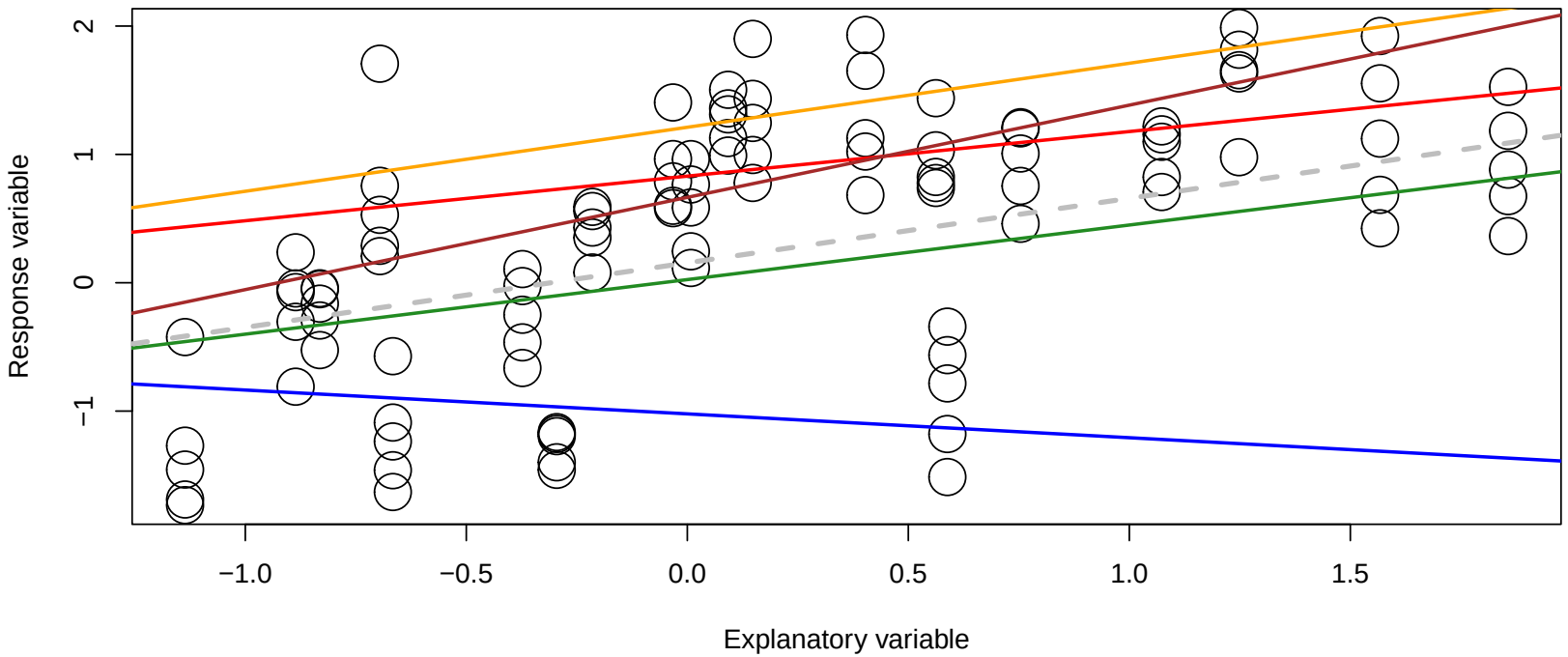
```
formula= ~ (0+covariate|1), beta0com = TRUE
```



Variation of mean abundance and environmental responses

$$y_{ij} = \alpha_j + x_i\beta_j, \text{ with } \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & 0 \\ 0 & \sigma_\beta^2 \end{pmatrix} \right\}$$

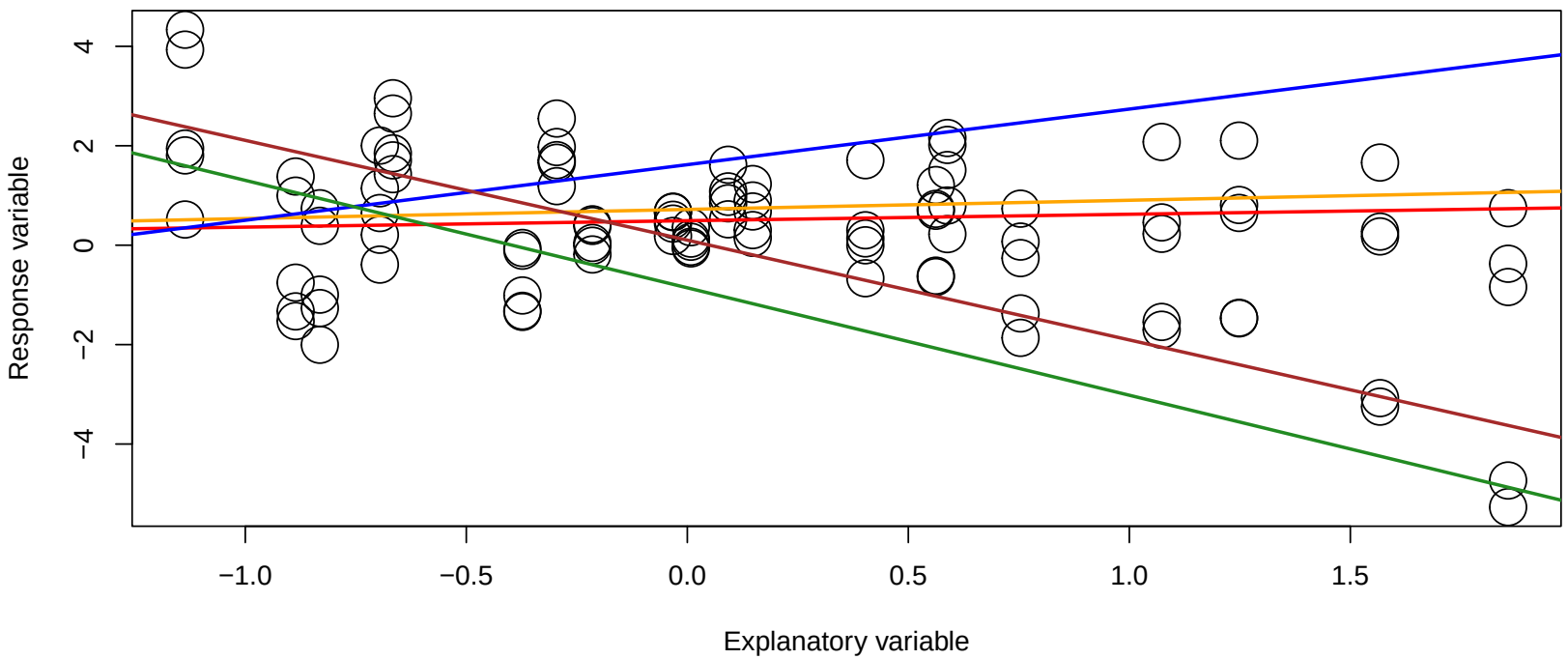
```
formula= ~ (1|1)+(0+covariate|1), beta0com = TRUE
```



Correlation of mean abundance and enviromental responses

$$y_{ij} = \alpha_j + x_i\beta_j, \text{ with } \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma^2_\alpha & \sigma^2_{\alpha,\beta} \\ \sigma^2_{\beta,\alpha} & \sigma^2_\beta \end{pmatrix} \right\}$$

```
formula= ~ (covariate|1), beta0com = TRUE
```



Example 2

```
X <- data.frame(sapply(X,function(x)if(is.numeric(x)){scale(x)}else{x}, simplify = FALSE))
```

```
library(gllvm)
```

```
## Loading required package: TMB

##
## Attaching package: 'gllvm'

## The following object is masked from 'package:stats':
##
##     simulate
```

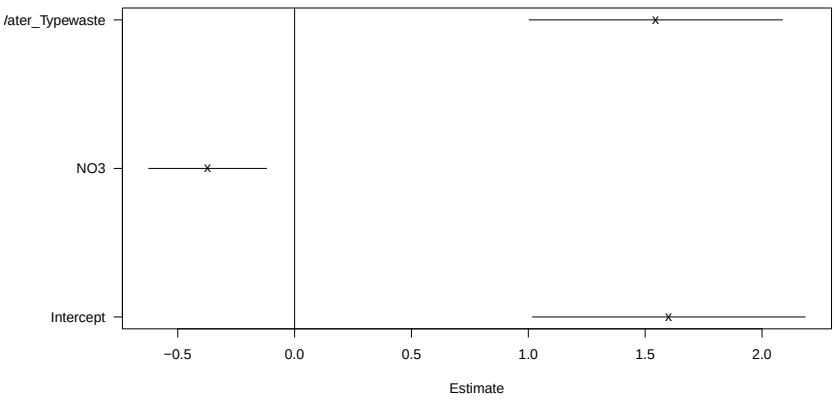
```
model <- gllvm(y, X = X, formula = ~(0+NO3|1) + (Water_Type|1), beta0com = TRUE, family = "negative.binomial",
```

Wetlands: summary

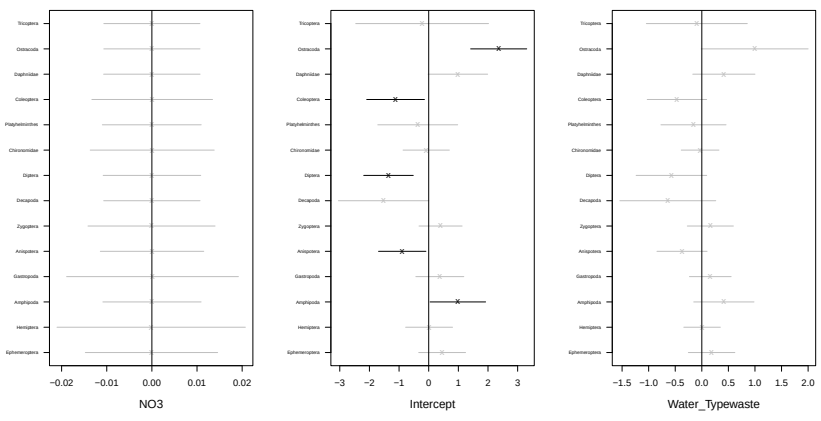
```
##
## Call:
## gllvm(y = y, X = X, formula = ~(0 + NO3 | 1) + (Water_Type |
##     1), family = "negative.binomial", num.lv = 0, beta0com = TRUE)
##
## Family:  negative.binomial
##
## AIC:  2666.544 AICc:  2668.407 BIC:  2755.793 LL:  -1312.3 df:  21
##
## Informed LVs:  0
## Constrained LVs:  0
## Unconstrained LVs:  0
##
## Formula:  ~(0 + NO3 | 1) + (Water_Type | 1)
## LV formula:  ~ 0
## Row effect:  ~ 1
##
## Random effects:
##  Name            Variance Std.Dev Corr
##  NO3             0.0000   0.0054
##  Intercept       1.2367   1.1121  0.0000
##  Water_Typewaste 0.2205   0.4696  0.0000 0.9939
##
## Coefficients predictors:
##                 Estimate Std. Error z value Pr(>|z|)
## Intercept         1.6011     0.3550   4.510 6.47e-06 ***
## NO3              -0.3722     0.1533  -2.428   0.0152 *
## Water_Typewaste   1.5450     0.3295   4.689 2.75e-06 ***
```

Results: plots

```
plot(summary(model))
```

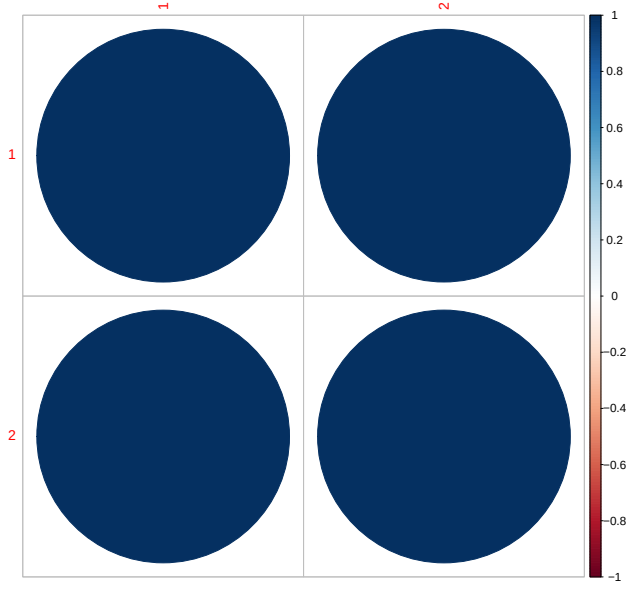



```
randomCoefplot(model)
```



Results: turnover

```
S <- model$params$sigmaB
x <- matrix(c(0, 1, 0,0,1,1), ncol = 2)
corrplot::corrplot(cov2cor(t(x)%*%S%*%x))
```



So, regardless of the water type, we expect to see the same community.

We'll give you something more interesting to look at in the exercise!

Summary

- GLMs are fun, but not usually suitable for multispecies data
- VGLMs; fitting one model per species gives more flexibility
- This facilitates adding components that are shared across species
- Which is especially helpful when working with random effects
- From which we can predict community similarity via turnover

So far we have assumed that species do not influence each other