GLLVMs from A-Å

Using the models in a complete study context

Audun Rugstad, Ph.D. candidate

Department of Mathematical Sciences, NTNU

2025-04-01

Intro

Outline of the session

Five-step approach:

1. Formulate your question in terms of a statistical model
2. Data exploration and preparation
3. Model fitting
4. Model checking
5. Making inferences

Outline of the session

We'll go through the full analysis of an ecological dataset

Example: Dou et al. (2022): *Influence of environmental variables on macroinvertebrate community structure in Lianhuan Lake*

Data: Counts of 74 aquatic invertebrates from 44 lake samples in northern China., with simultaneous measures of 13 physical-chemical properties of the water.

Formulating the question

Formulating the question

Formulate your research question in terms of a statistical analysis

First question: **What is the goal of the analysis?**

- **Prediction:** Find the model that best predict future observations
- **Explanation:** Investigate relationships between explanatory variables and response(s)

Prediction

Prediction

- Find the model that best predict future values of one/more response variables (i.e. species distribution)
- Variable selection based on optimizing for prediction
- I.e. forward/backward selection using AIC, Root mean square error etc.

Example: A model that predicts future distributions of wood beetles under different climate change scenarios

Explanation

Two types:

Confirmatory

- Test a **specific, clear hypothesis** with one/a few models
- Predictor variables selected based on *a priori* knowledge
- Ideally **no variable selection** and pre-registration

Explanation

Two types:

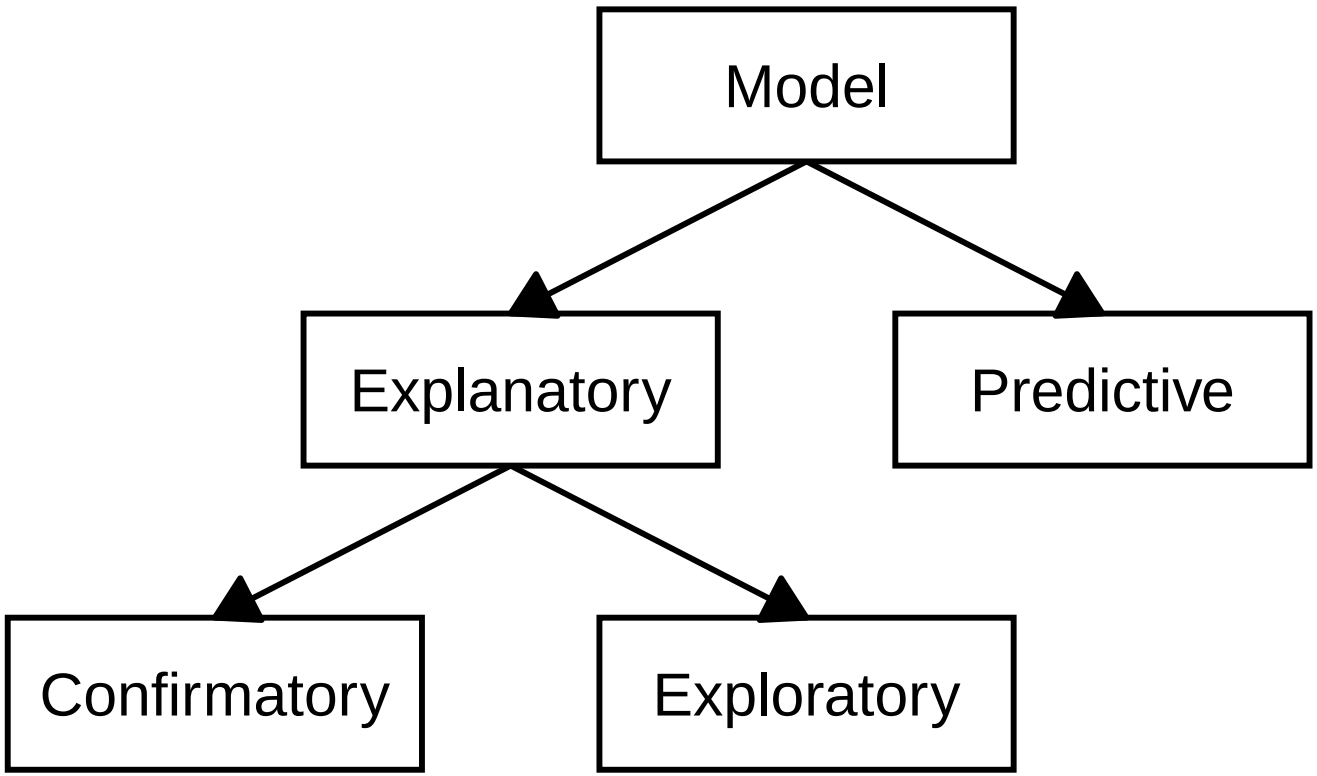
Confirmatory

- Test a **specific, clear hypothesis** with one/a few models
- Predictor variables selected based on *a priori* knowledge
- Ideally **no variable selection** and pre-registration

Exploratory

- Use model(s) to **explore possible** causal relationships
- Should only be used to generate hypotheses
- Avoid automatic model selection

Flowchart



How does this apply to community ecology

Exploratory "throwing everything at the wall" approaches perhaps most common

JSDMs more associated with prediction than other methods?

"True" confirmatory analyses are rare

- The line between the three can often be blurry

How does this apply to community ecology

How can GLLVMs help?

- More streamlined and focused analysis
- Ideally, a single model of the community that is as complete as possible
- Can then be used to make inferences about different causal relationships etc.

How does this apply to community ecology

Q: What type of analysis does a GLLVM with only **unconstrained** latent variables fit into?

How does this apply to community ecology

Q: What type of analysis does a GLLVM with only **unconstrained** latent variables fit into?

A: Primarily exploratory, generating hypothesis about factors that structure the community.

But: if the goal is to test specific hypothesis about species co-occurrences, could potentially be used more confirmatory

How does this apply to community ecology

Q: What type of analysis does a GLLVM with **constrained** latent variables fit into?

How does this apply to community ecology

Q: What type of analysis does a GLLVM with **constrained** latent variables fit into?

A: All three, depending on approach?

Example: Invertebrates in Lihuan

Dou et al. (2022): Macroinvertebrates in 13 lakes in Lihuan, CN:

From the intro: *The community composition of macro-invertebrate assemblages and their relationships with environmental variables were investigated*

What type of modeling approach makes sense here?

Example: Invertebrates in Lihuan

Dou et al. (2022): Macroinvertebrates in 13 lakes in Lihuan, CN:

From the intro: *The community composition of macro-invertebrate assemblages and their relationships with environmental variables were investigated*

What type of modeling approach makes sense here?

A: Seems very clearly **exploratory** (which variables are potential drivers of community change?)

Example: Invertebrates in Lihuan

What should the model look like?

Example: Invertebrates in Lihuan

What should the model look like?

A **concurrent ordination** might be a good fit

- Allows for assessing the relative effect of many predictor variables at once, on a small number of latent variables
- Also allows us to account for potential residual variation in the L.V.s (i.e. variation not explained by the predictors)

Example 2

Haven't yet found a good dataset/study that is not exploratory in the same way:(

Data preparation

Data exploration and preparation

Important to always **visually inspect** key properties of the data before modeling

- Ensure the data meets the assumptions of the model
- Act as a "sanity check" for modeling output

Standardize and scale variables as needed

As with most other models, we recommend "common sense" practices and guidelines such as Zuur, Ieno, and Elphick (2010)

Data exploration and preparation

Some factors that can mess up a GLLVM:

- Predictor variables that are too co-linear
- Several species with very little information on abundances (e.g. zero-inflation)
- Sites with very little information on abundances
- Highly unbalanced sampling

Model philosophy

The underlying philosophy of GLLVMs is:

- Data is ideally gathered with a spesific model or analysis in mind
- If not the case: tailor a model to the data properties
- The model should account for different properties of the data, as far as possible

The model should suit the data, not the other way around.

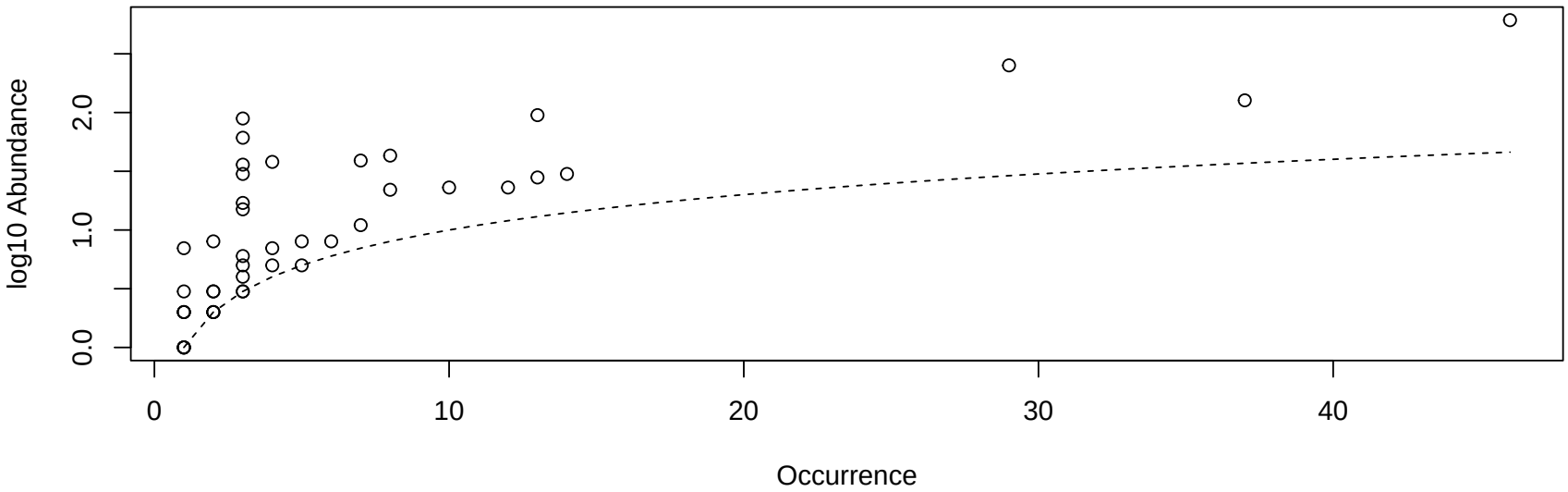
Example: Lake invertebrates

```
# load data
lake_sp <- read.csv("../data/lake_inverts_sp.csv", sep = "\t")
lake_env <- read.csv("../data/lake_inverts_env.csv", sep = "\t")

# clean (according to paper)
lake_sp <- lake_sp |> select(-Site, -Group, -Season, -Lakes)
# reduce to actual counts!
lake_sp <- lake_sp / 16
```

Example: Species

We can use a (Log-)abundance occupancy plot:



X = Number of plots a species occupies, **Y** = average log-abundance

Can help identify outlier species + get a sense of the commonness of species in the dataset

Species data

Table of total species abundances:

1	2	3	4	5	6	7	8	11	15	17	22	23	28	30	36	38	39
4	6	5	1	3	1	2	3	1	1	1	1	2	1	2	1	1	1
89	95	127	252	612													
1	1	1	1	1													

9 species with 1-2 occurrences in the dataset, and three species with very high abundances

Filtering data

Suggestion: remove species occurring in only one plot, and plots containing only one species (remove 8 species and 9 plots).

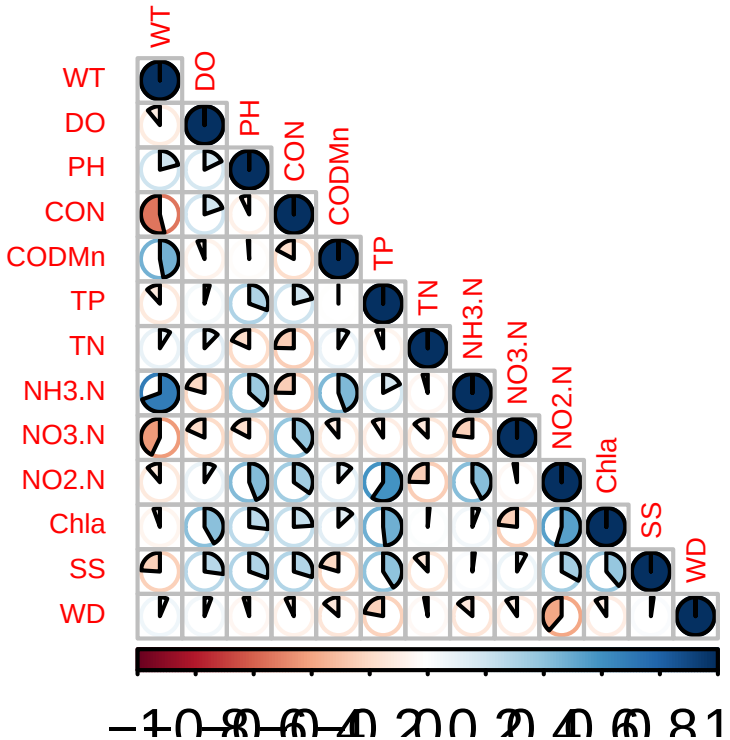
```
# remove rare species
lake_sp_filt <- lake_sp |>
  select_if(colSums(lake_sp !=0)>1)

# remove rare sites
lake_sp_filt <- lake_sp_filt[rowSums(lake_sp_filt != 0)>1,]

# remove same sites in env. dataframe
lake_env_filt <- lake_env[row.names(lake_sp_filt),]
```

Environmental variables

Correlation of all predictors (lake chemical attributes):



Environmental variables

Why is this important?

- High co-linearity makes it hard for the model to converge (find estimates)
- Also makes the parameter uncertainties larger

Environmental variables

What can be done?

- Omit predictors
  - Need to consider properties of the system and analysis
- "Model tricks"
  - Regularization ("reigning in" wild estimates)
  - Shrinkage (finding ways to make unimportant parameters -> 0)
  - Ex: Switching to random effects or putting constraints on the response distribution

How high is too high?

Rule of thumb: no more than 0.8

Our max cor = 0.69

So we "just" standardize and scale to mean = 0 and variance = 1:

```
lake_env[,5:17] <- scale(lake_env[,5:17])
```

- So all coefficient estimates are on the same magnitude

Model setup

Model setup

Our suggestion: Concurrent ordination

What to include in the model?

- Our goal is to explore the impact of **environmental variables**
- As such, we can maybe argue for ignoring the study design here (differences between season/lake assumed to be manifested through the water properties)

Model setup

Initial model setup:

- Two concurrent latent variables informed by the chemical properties of the water
- Random row effects (to account for site variation in overall abundance)
- Poisson distribution
- *Could* include quadratic effects, but might be overloading the model

Model setup

```
mod_lakes <- gllvm(lake_sp_filt, # species dataframe
                   X = lake_env_filt, #envrionemtnal dataframe
                   lv.formula = ~ WT + DO + PH + CON + CODMn +
                         TP + TN + NH3.N + NO3.N + NO2.N + Chla +
                         SS + WD,
                   row.eff = "random", # random row effect
                   family = "poisson", # poisson distribution
                   num.lv.c = 2, #number of concurrent LVs
                   n.init = 20) # run the model a few times
                                # to ensure convergence
```

Model setup

A few (personal) tips for the `gllvm()` function:

- `n.init`: How many starting iterations you run
  - Depending on how long one run takes, maybe 5-50. Tells you when it has converged.
- `trace = TRUE`: tells you when each model run specified in `n.init` is complete (makes it less frustrating to fit)
- `sd.errors = FALSE/TRUE`: when diagnosing/comparing models, you can turn off estimation of standard errors to make each model fitting faster (and fit them retroactively using the `se.gllvm()` function)

Model checking

Model checking with GLLVMs

What to look at to determine whether the model is "good"?

- **Diagnostic plots** (`plot()` function) for model assumptions
  - Most important!
- AIC/BIC to compare model fit (e.g. number of LVs needed)
- Goodness of fit-tests (`gllvm.goodnessOfFit()`)
- Very small/large parameter estimates (particularly sp. coefficients) for convergence

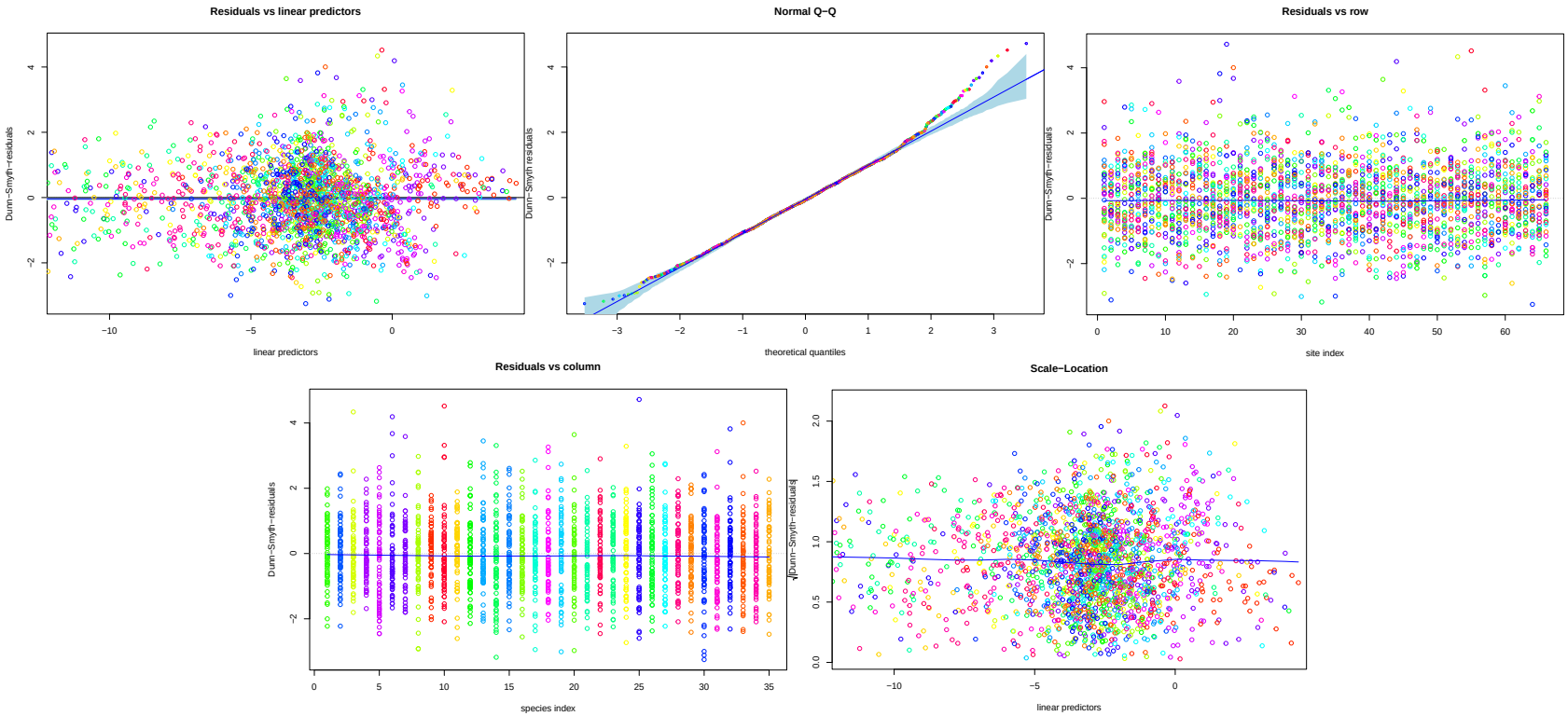
Model checking with GLLVMs

What to do to improve the fit?

- Change response distribution
- Run more iterations
- Change starting values
- Change from fixed to random effects
- Change number of LVs

Model checking: Example

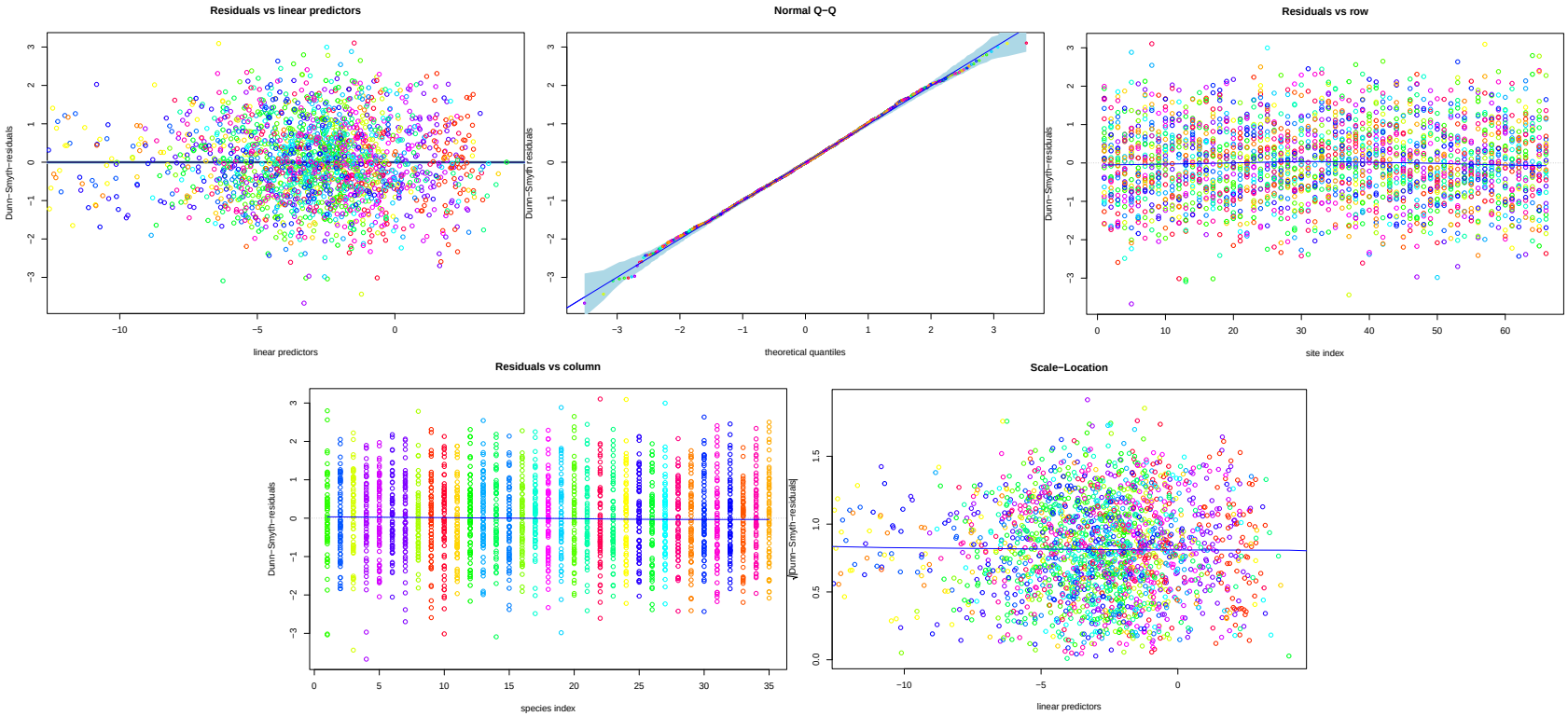
We can look at diagnostic plots:



We see some **fanning** and potential **overdispersion** in the residual and QQ plots.

Model checking: Example

After re-running the model with `family = "negative.binomial"`:



The residual plots indicate that the NB model meets the assumptions of the

Model checking: Example

The difference in **AICc** values between model 1 and 2 also indicates that the negative binomial model is better.

```
AICc(mod_lakes_pois, mod_lakes_nb)
```

```
[1] 2918.967 2476.823
```

Model checking: Example

Looking at some **goodness-of-fit** measures to the original data, however...

	Measure	Poisson	NB
1	cor	0.98382312	0.02060082
2	RMSE	0.78925359	65.12833567
3	MAE	0.28928256	2.24318316
4	MARNE	0.05016865	0.15774418

Model checking: Example

Looking at some **goodness-of-fit** measures to the original data, however...

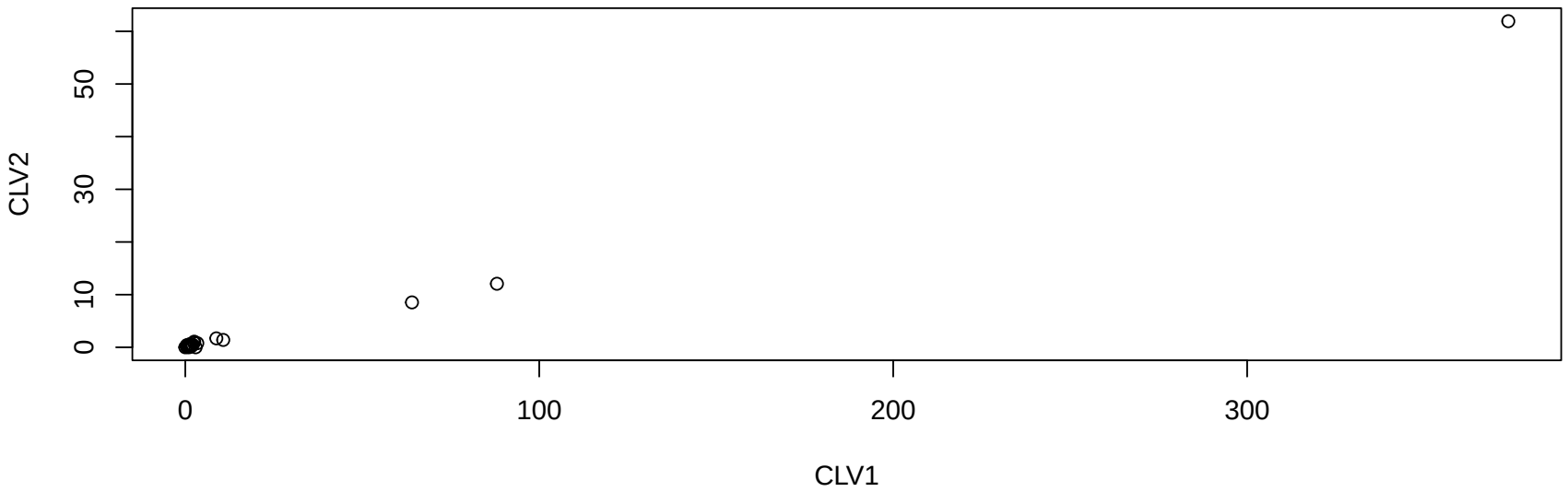
	Measure	Poisson	NB
1	cor	0.98382312	0.02060082
2	RMSE	0.78925359	65.12833567
3	MAE	0.28928256	2.24318316
4	MARNE	0.05016865	0.15774418

Conclusion: Our N.B. model, while meeting the model assumptions (a valid model), is a pretty bad fit to our data

What has happened?

If we look at the plots of the standard errors for the species loadings, we also see that some are very large, indicating a poor convergence of the model:

```
plot(mod_lakes_nb$sd$theta) # theta = species coefficients
```



Improving the NB model

We try two things:

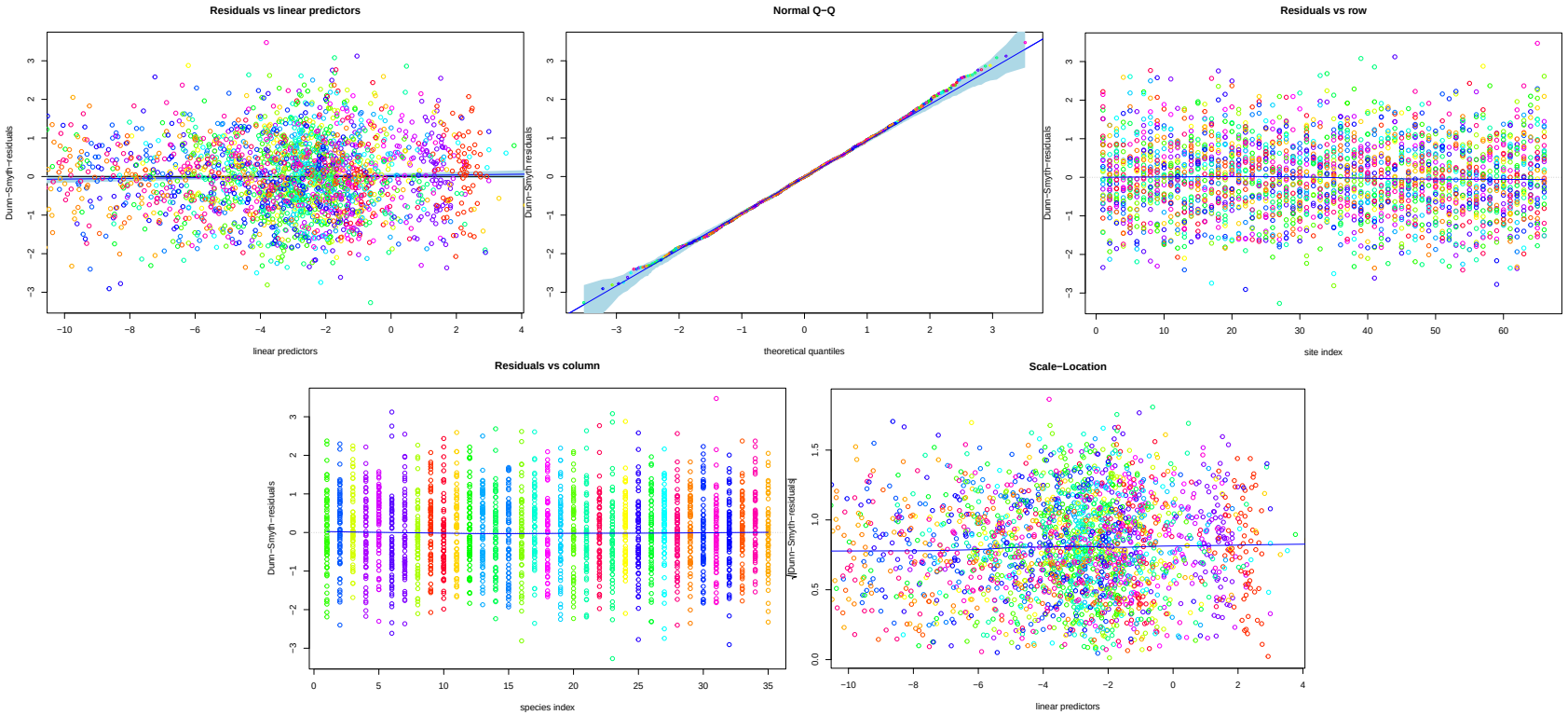
- Specify the LV-coefficients as **random effects**, drawn from a distribution unique for each coefficient
- Tell the model to estimate a common *dispersion* parameter for the negative binomial distribution for each species

Both could help in steering the model away from "extreme" estimates and lessen overfitting (=regularizing)

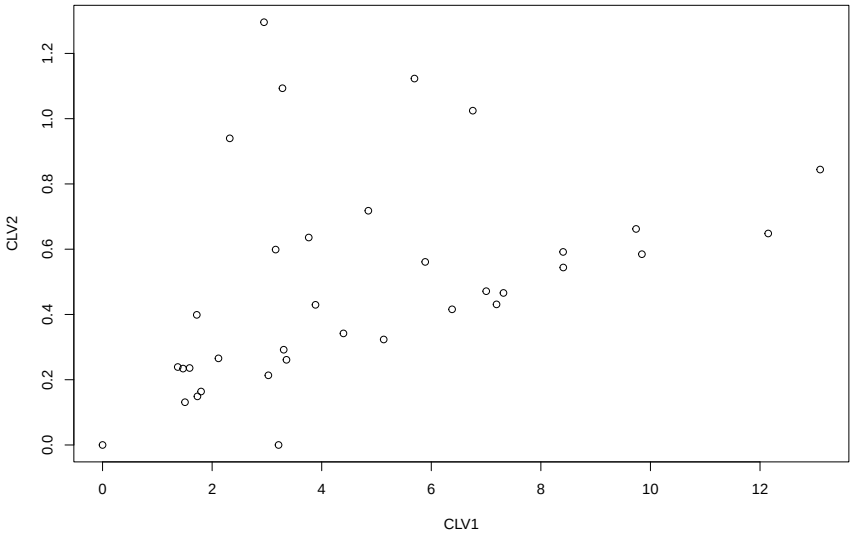
Improving the NB model

```
mod_lakes_nb_2 <- gllvm(lake_sp_filt, X = lake_env_filt,
                        lv.formula = ~ WT + DO + PH + CON + CODMn +
                              TP + TN + NH3.N + NO3.N + NO2.N + Chla +
                              SS + WD,
                        row.eff = "random",
                        family = "negative.binomial",
                        ## LV coefficients are random effects:
                        randomB = "P",
                        ## Only one dispersion parameter estimated:
                        disp.formula = rep(1, ncol(lake_sp_filt)),
                        num.lv.c = 2,
                        n.init = 20)
```

Checking the new model



Checking the new model



	Measure	NB1	NB2
1	cor	0.02060082	0.4658600
2	RMSE	65.12833567	3.8379823
3	MAE	2.24318316	0.8962227
4	MARNE	0.15774418	0.0670415

Making inferences

Making inferences with GLLVMs

What you look at depends on your question.

Typical visualizations:

- Model summary
- Biplots and triplots
- Variance explained
- Coefficient plots (caterpillar plots)

Model summary

Two most important parts for us:

`Residual standard deviation of LVs`:

```
[1] 4.198412e-05 1.926389e-06
```

Very low, implies our predictors explain all the variance in the latent variables

We *could* go back and fit the model again without residual LV variation (as a "classic" constrained ordination)

Model summary

Coefficients LV predictors:

	Name	Std.Dev.CLV1	Variance.CLV1	Std.Dev.CLV2	Variance.CLV2
1	WT	0.2022	0.0409	0.1724	0.0297
2	DO	0.0003	0.0000	0.0003	0.0000
3	PH	0.2332	0.0544	0.1989	0.0396
4	CON	0.0001	0.0000	0.0001	0.0000
5	CODMn	0.0000	0.0000	0.0000	0.0000
6	TP	0.1105	0.0122	0.0942	0.0089
7	TN	0.0004	0.0000	0.0003	0.0000
8	NH3.N	1.0835	1.1739	0.9239	0.8535
9	NO3.N	0.0000	0.0000	0.0000	0.0000
10	NO2.N	0.0005	0.0000	0.0004	0.0000
11	Chla	1.7938	3.2179	1.5296	2.3397
12	SS	0.0003	0.0000	0.0002	0.0000
13	WD	0.0002	0.0000	0.0002	0.0000

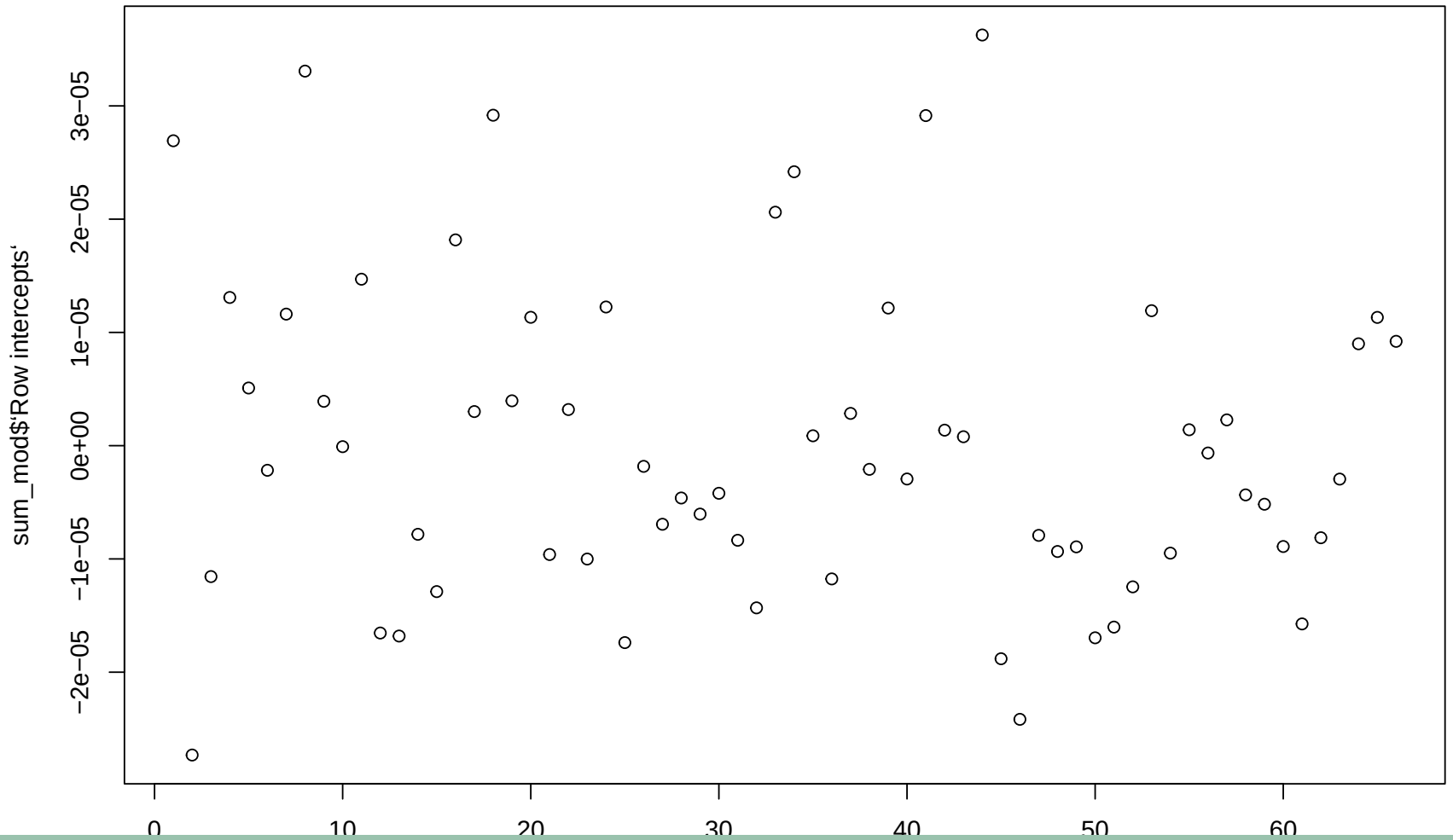
Model summary

The most importation predictors seem to be `WT` (water temperature), `PH` (water PH), and `Chla` (chlorophyll A content)

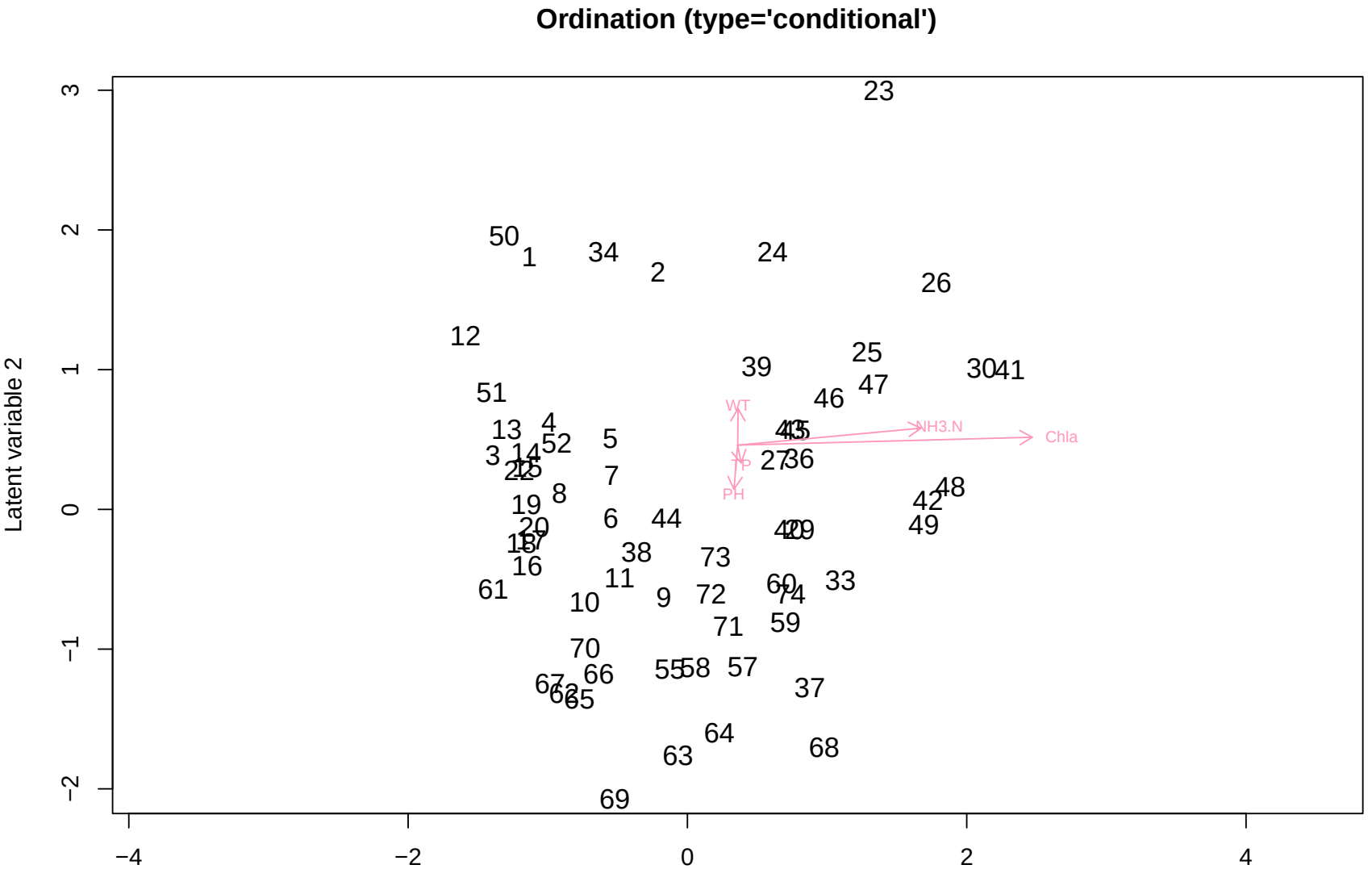
Relatively similar to the conclusion in Dou et al. (2022).

Model summary

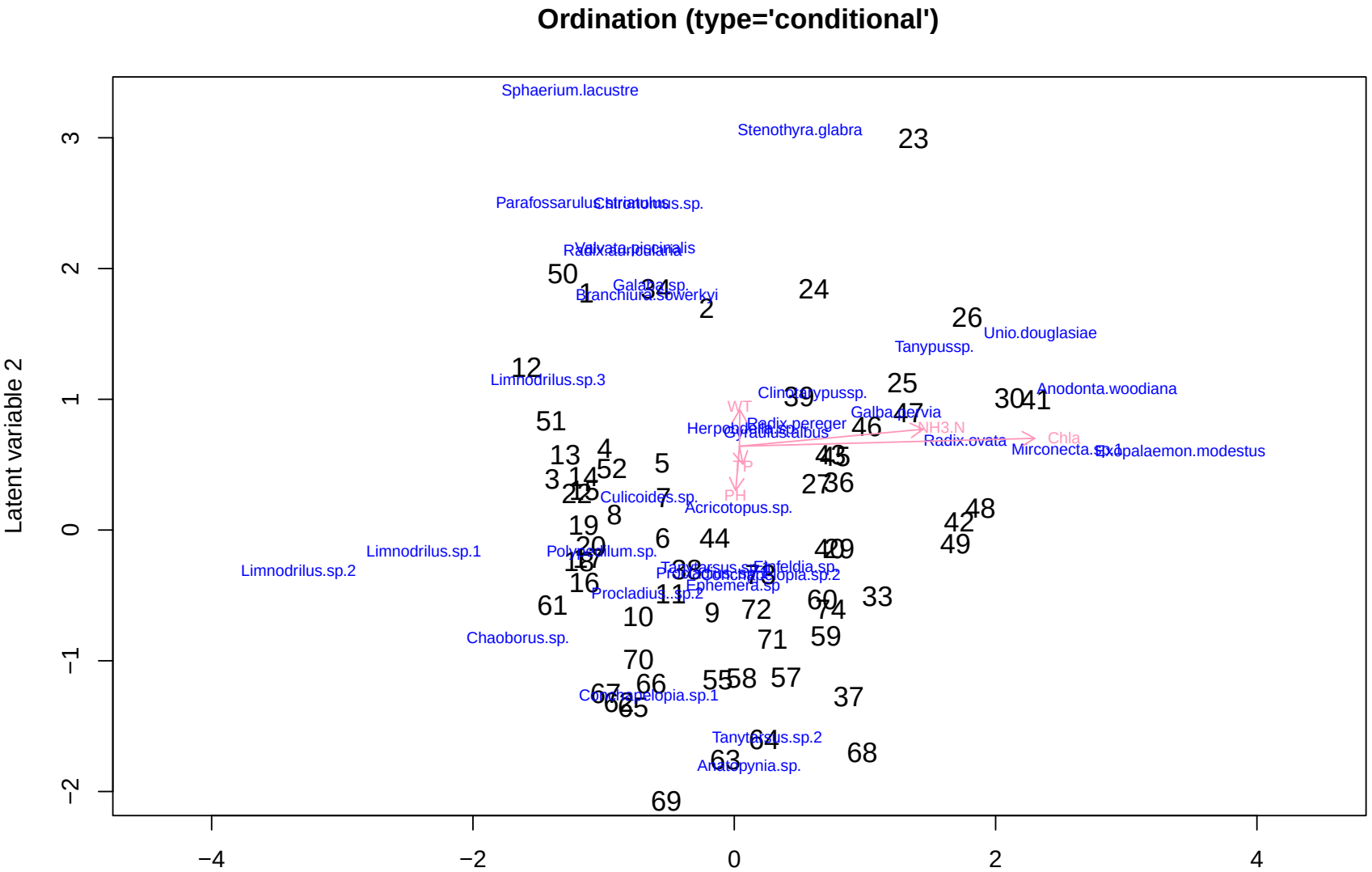
We can also look at the estimates for the random row effects:



Biplot and triplot



Biplot and triplot

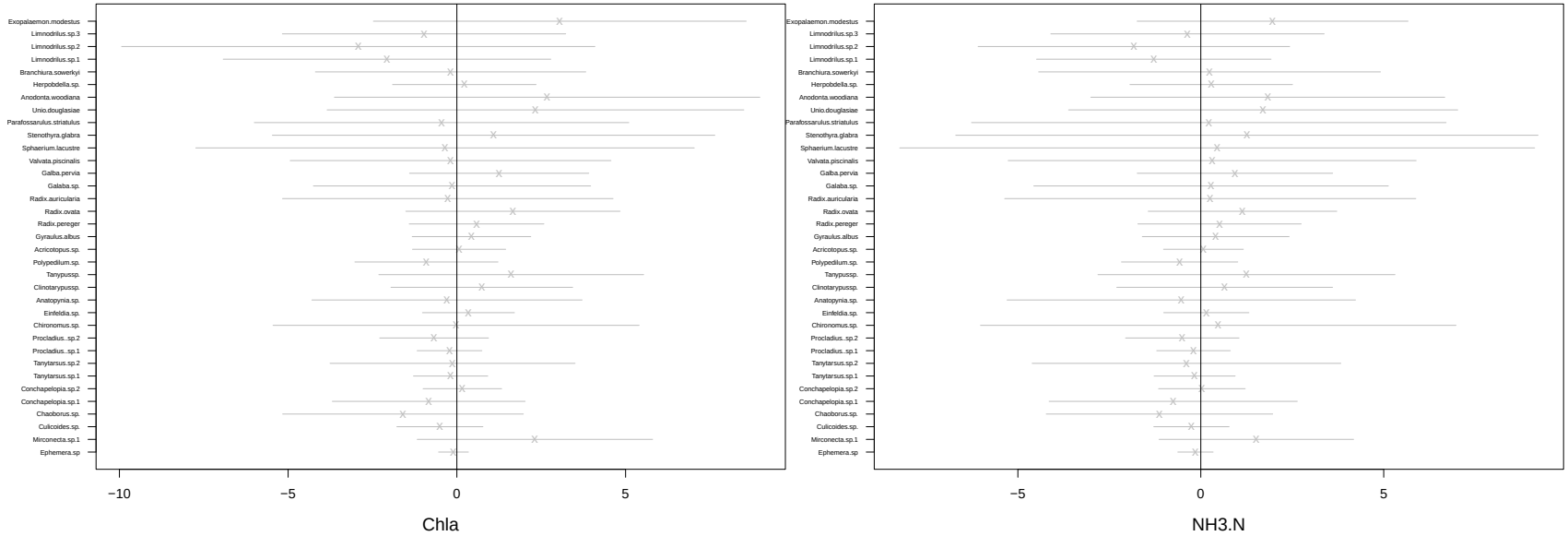


Variance explained

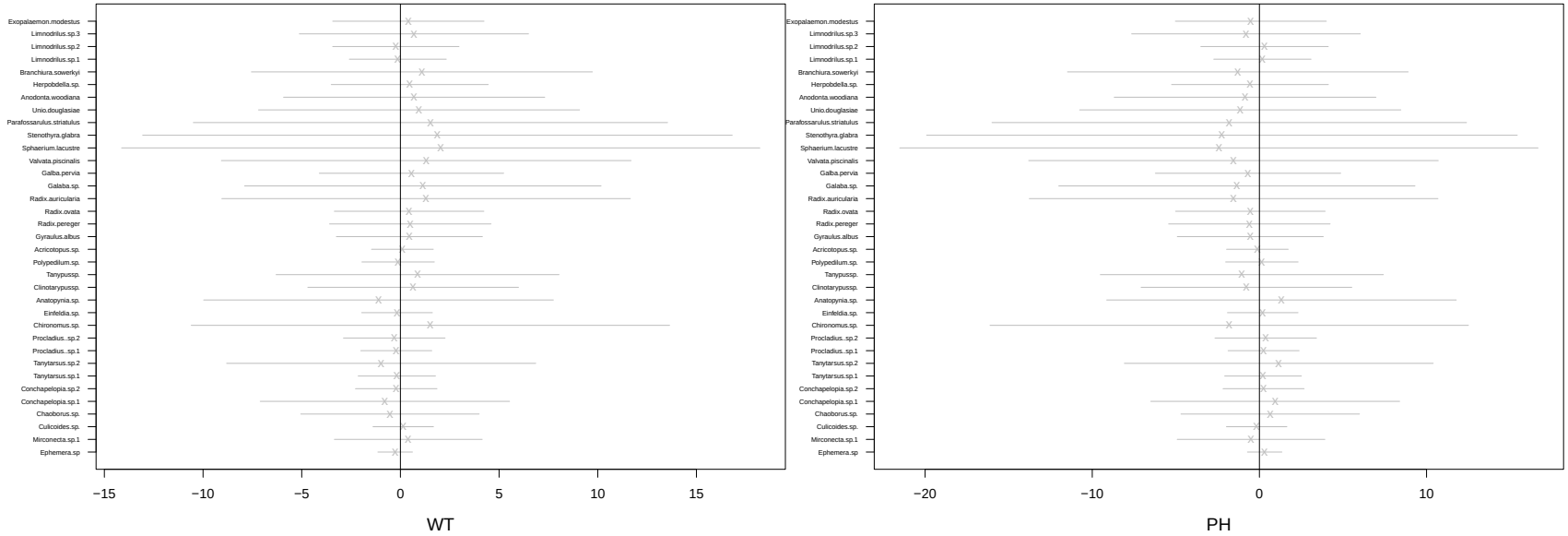
```
varPartitioning(mod_lakes_nb_2)
```

```
Effect                       Mean.explained.variance
CLV:WT                       20.1%
CLV:DO                        0.0%
CLV:PH                       27.9%
CLV:CON                       0.0%
CLV:CODMn                     0.0%
CLV:TP                        5.2%
CLV:TN                        0.0%
CLV:NH3.N                    15.2%
CLV:NO3.N                     0.0%
CLV:NO2.N                     0.0%
CLV:Chla                     31.5%
CLV:SS                        0.0%
CLV:WD                        0.0%
CLV1                          0.0%
CLV2                          0.0%
Row random effect: sample     0.0%
```

Coefficient plots



Coefficient plots



Takeaway so far?

Takeaway so far?

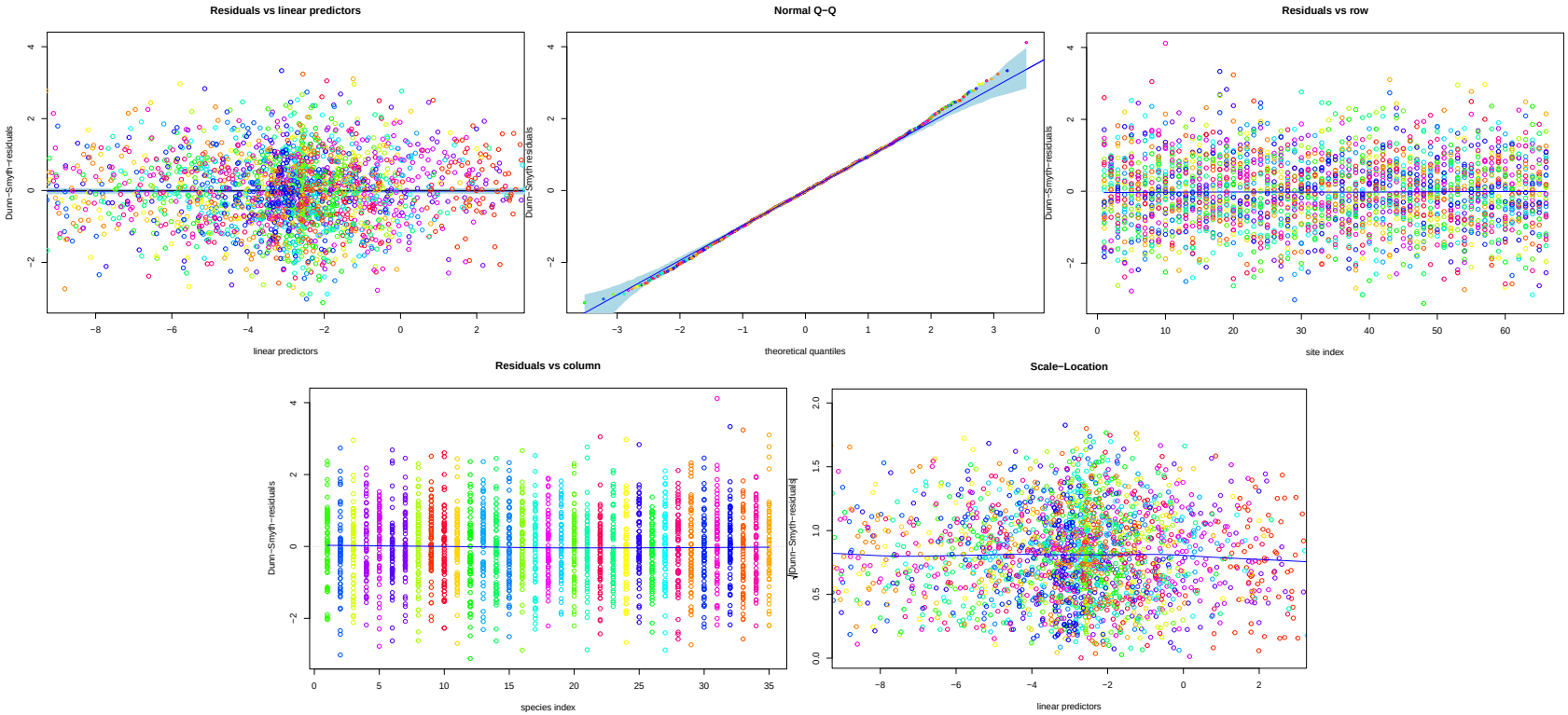
- While the model seems well behaved and reasonably fit to the data, estimates for the effect of predictors on the species are **very** high
- Suggests it it not very useful for explanatory purpose (only very weak evidence)

Comparison to fully constrained model

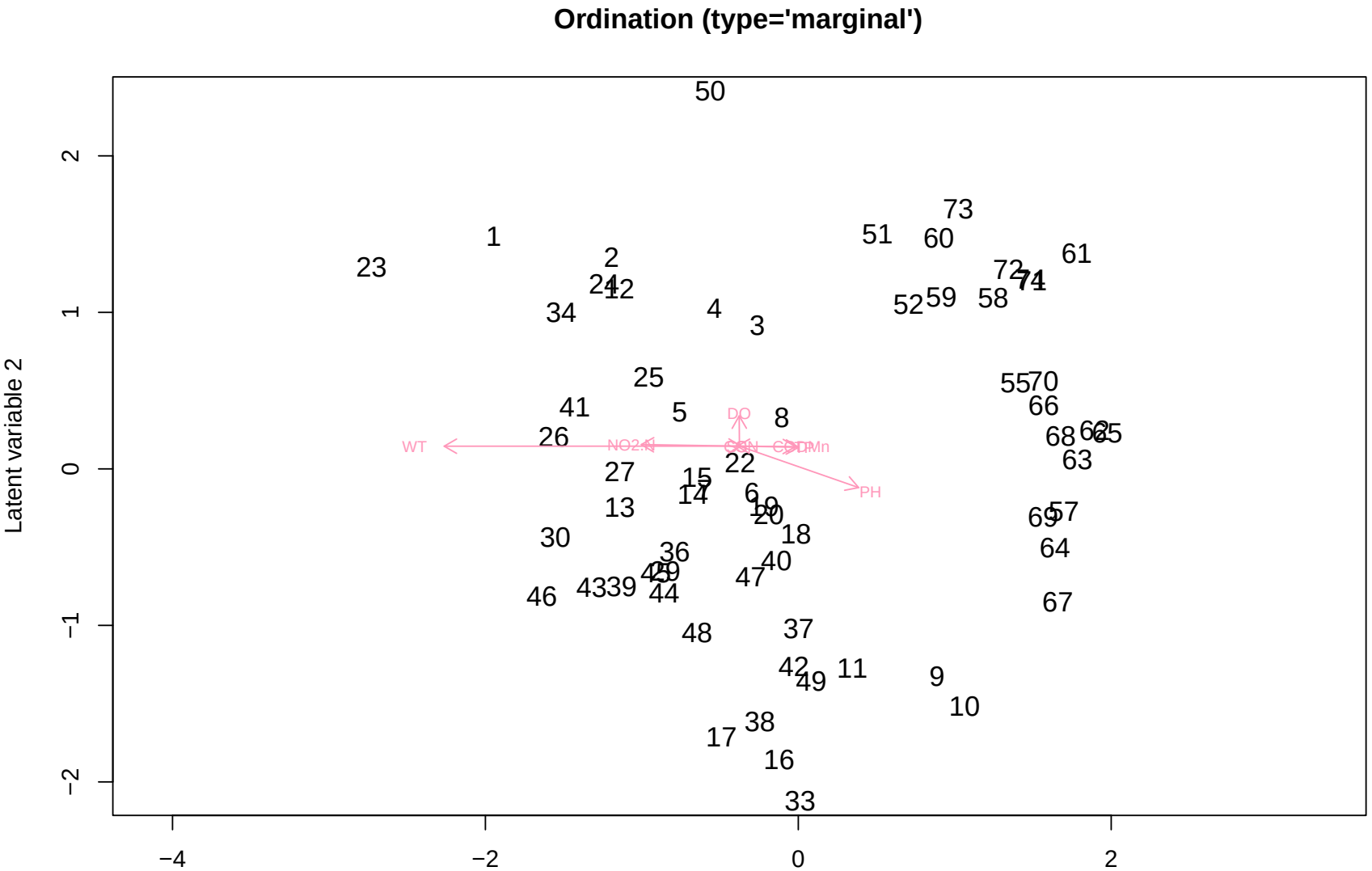
To get a sense of the problem, we could compare it to the slightly more "shaved" model that the parameter estimates seem to suggest:

```
mod_lakes_nb_3 <- gllvm(lake_sp_filt, X = lake_env_filt,
                    lv.formula = ~ WT + DO + PH + CON + CODMn +
                          TP + TN + NH3.N + NO3.N + NO2.N + Chla +
                          SS + WD,
                    ## Remove random row effects:
                    #row.eff = "random",
                    family = "negative.binomial",
                    ## LV coefficients are random effects:
                    randomB = "P",
                    ## Only one dispersion parameter estimated:
                    disp.formula = rep(1, ncol(lake_sp_filt)),
                    ## Fully constrained (reduced rank) coefs:
                    num.RR = 2,
                    n.init = 20)
```

Model checking



Biplot (model 2)



Variance explained (model 2)

```
varPartitioning(mod_lakes_nb_3)
```

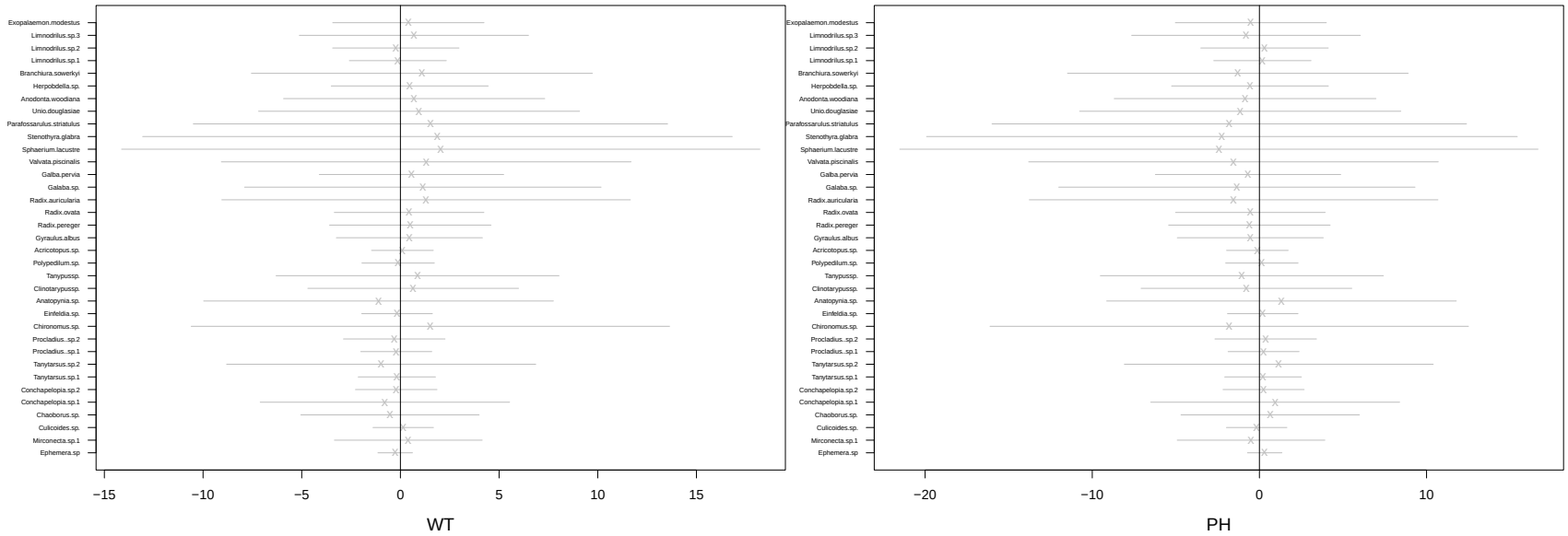
```
Effect    Mean.explained.variance
CLV:WT    50.7%
CLV:DO     9.9%
CLV:PH    29.7%
CLV:CON    0.0%
CLV:CODMn  2.1%
CLV:TP     2.2%
CLV:TN     0.0%
CLV:NH3.N  0.0%
CLV:NO3.N  0.0%
CLV:NO2.N  5.4%
CLV:Chla   0.0%
CLV:SS     0.0%
CLV:WD     0.0%
```

What has changed?

What has changed?

The effect of Chlorophyll and NH3 content has **disappeared**!

Coefficient plots (model 2)



Still extremely high uncertainty

Conclusion

Conclusion

Both models can be argued to be a reasonable, but suggest different drivers of species diversity.

- Still, some indication from both that PH and water temperature are important factors to look into
- Large uncertainties like this often a feature of negative binomial models

If the goal was prediction, we could have done automatic model selection or similar to (probably) get a simpler model with clearer estimates

Conclusion

Highlights the need to not trust model outputs blindly

The tools showed here can be useful in figuring out what is going on.

Next time: Try to collect more data in a way that could fit a Poisson model?

- Some unclearness about the actual sampling effort and pooling of samples in this study
- Unbalanced study design can have had an influence(more sampling in alkaline lakes than

Conclusion

In other words:

- We can say both **that** more data is needed, and something about **why** more data is needed:)

References

References

Dou, Qianming, Xue Du, Yanfeng Cong, Le Wang, Chen Zhao, Dan Song, Hui Liu, and Tangbin Huo. 2022. "Influence of Environmental Variables on Macroinvertebrate Community Structure in Lianhuan Lake." *Ecology and Evolution* 12 (2): e8553. https://doi.org/10.1002/ece3.8553.

Zuur, Alain F., Elena N. Ieno, and Chris S. Elphick. 2010. "A Protocol for Data Exploration to Avoid Common Statistical Problems." *Methods in Ecology and Evolution* 1 (1): 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x.