Recap: Generalised Linear Mixed Models

Bert van der Veen

Department of Mathematical Sciences, NTNU

Questions so far?

Outline

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

Background

- We can formulate the same models
- But now, parameters come from a distribution

Likelihood formulation: independence

$$\mathcal{L}(\mathbf{y}; \Theta) = \prod_{i}^{n} f(y_i; \Theta) \tag{1}$$

We just multiply! (assumes independence)

Our new likelihood

$$\mathcal{L}(\mathbf{y}; \mathbf{\Sigma}) = \int \prod_{i}^{n} f(y_i|\mathbf{u}) f(\mathbf{u}; \mathbf{\Sigma}) d\mathbf{u} \tag{2}$$

- Fixed effects: what we had so far
- Random effects: new, come from a distribution
- Mixed effects model: contains both

The mixed-effects model

$$g\{\mathsf{E}(\mathbf{y}|\mathbf{u})\} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} \tag{3}$$

1. Link-function
2. Conditional mean
3. Fixed effects design matrix
4. Random effects design matrix

The mixed-effects model

$$g\{\mathsf{E}(\mathbf{y}|\mathbf{u})\} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} \tag{3}$$

1. Link-function
2. Conditional mean
3. Fixed effects parameter vector
4. Random effects parameter vector

The random effects design matrix

- it's the kind of thing as the fixed effects design matrix!

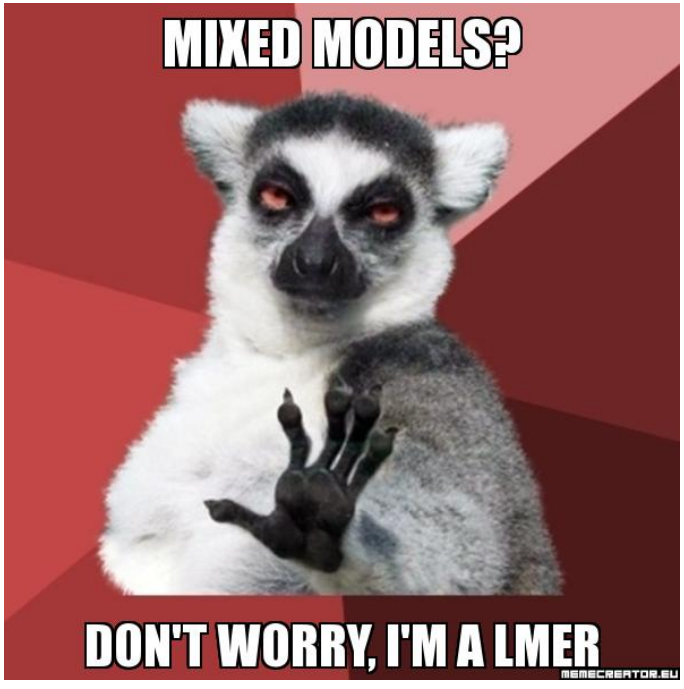
When to include a random effect

- Unobserved effect vs. observed effect
- To account for pseudo replication
- Nuisance vs. of interest
- If parameter comes from a population

To incorporate (species) correlation

A linear mixed-effects model

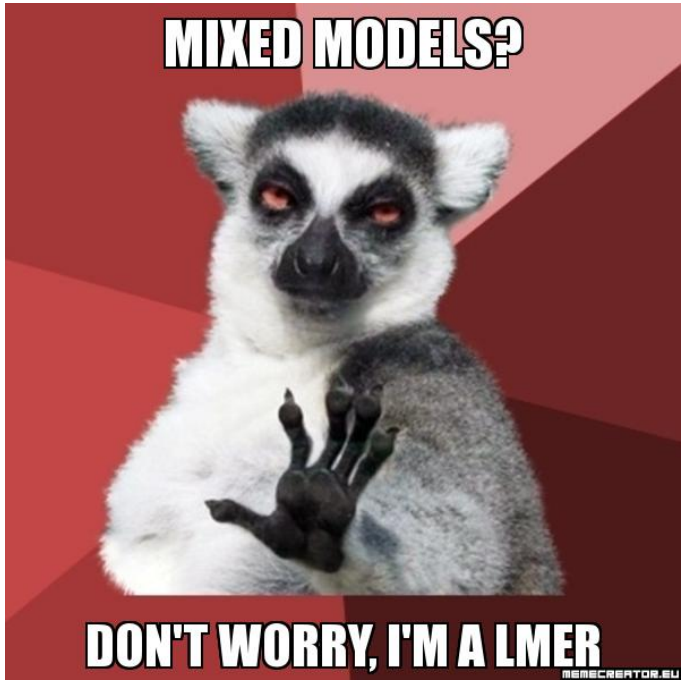
$$\mathsf{E}(\mathbf{y}|\mathbf{u}) = \boldsymbol{\mu} \qquad (4)$$



A linear mixed-effects model

$$\mathsf{E}(\mathbf{y}|\mathbf{u}) = \boldsymbol{\mu} \tag{4}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{5}$$

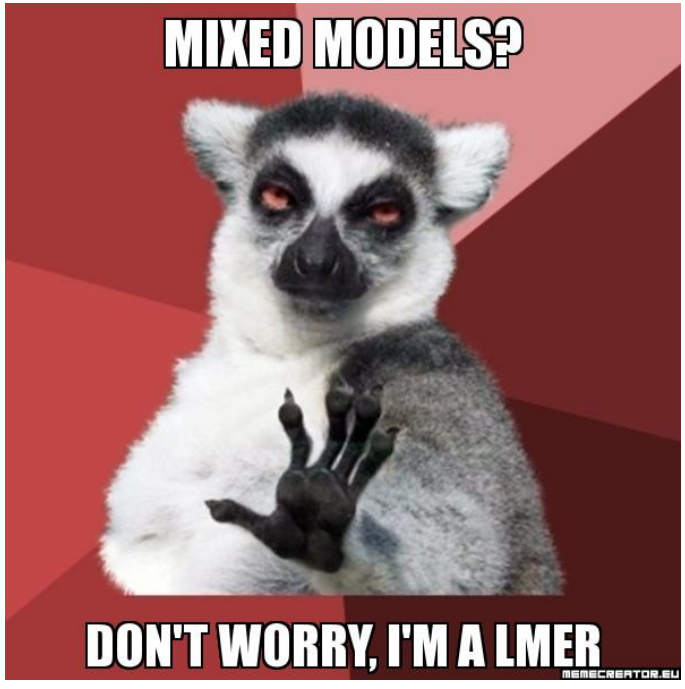


A linear mixed-effects model

$$\mathsf{E}(\mathbf{y}|\mathbf{u}) = \boldsymbol{\mu} \tag{4}$$

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{5}$$

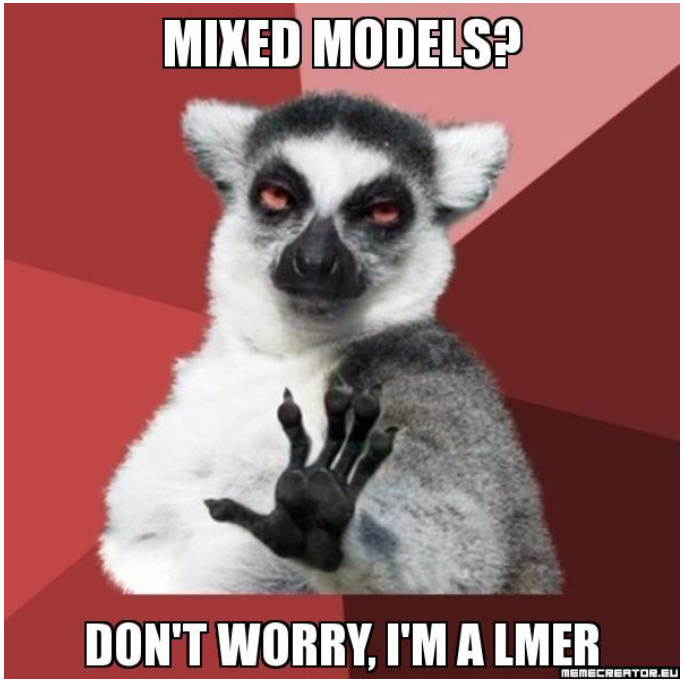
with $\mathbf{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$
with $\mathbf{e} \sim \mathcal{N}(0, \mathbf{I}\sigma^2)$



A linear mixed-effects model

We can rewrite the model in terms of the complete error term.

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon} \qquad (6)$$

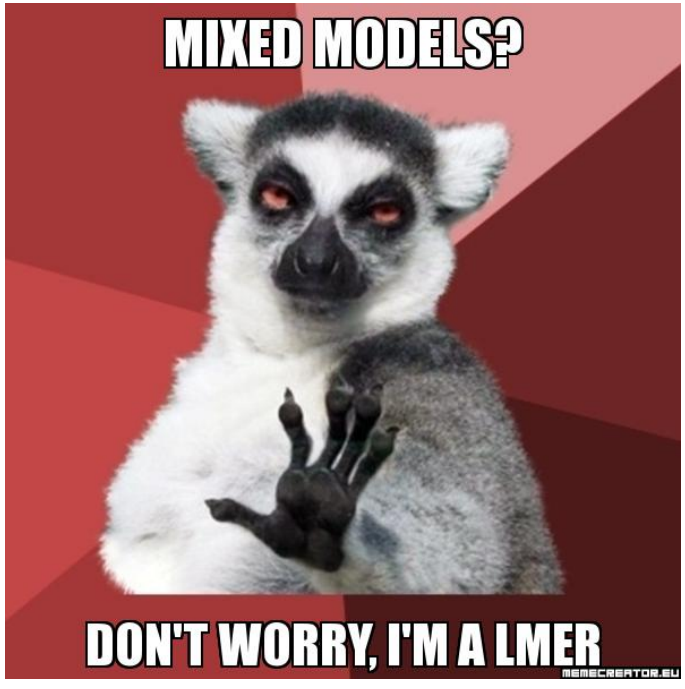


A linear mixed-effects model

We can rewrite the model in terms of the complete error term.

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon} \qquad (6)$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^{\top} + \mathbf{I}\sigma^2)$$



So, we are including covariance between our errors in the model.

The objective function

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e} \tag{7}$$

with $\boldsymbol{\epsilon} = \mathbf{Zu} + \mathbf{e}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{Z\Sigma Z}^\top + \mathbf{I}\sigma^2)$

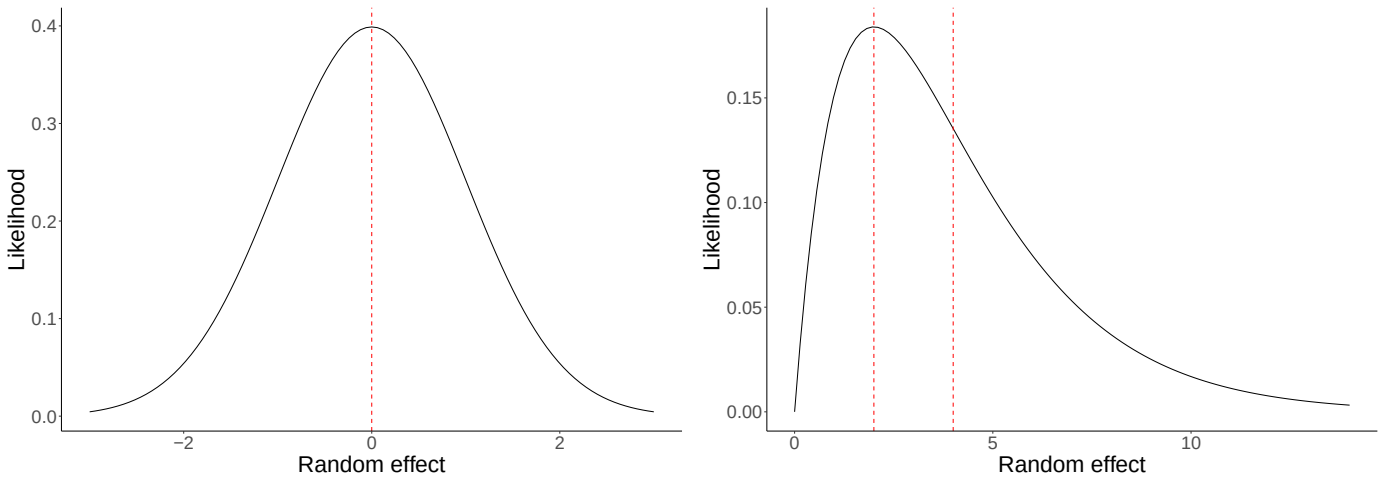
we have the marginal distribution $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \mathbf{Z\Sigma Z}^\top + \mathbf{I}\sigma^2)$

This is not how things are done in practice (because the covariance matrix can get quite big!)

Estimation

- Penalized quasi-likelihood methods
- Adaptive GH quadrature
- Laplace approximation
- Variational approximations
- Et cetera (see e.g., Bolker et al. 2009)

Measure of central tendency: Mean or Mode



Maximum Likelihood Estimation

At the maximum of the likelihood:

- The gradient is zero (tangent is straight)
- The hessian (of -LL) should
  - have positive diagonals
  - positive eigenvalues
  - be symmetric
  - and is thus invertible (we go up in both directions)
- Asymptotic covariance matrix is given by the inverse of the negative Hessian

These are important concepts to understand error messages and convergence in mixed-models.

There are many R-packages

- `nlme`
- `lme4`
- `glmmTMB` (or `glmmADMB`)
- `sdmTMB`
- `MASS`
- `glmmML`
- `repeated`
- `glmm`
- `hglm`
- `spaMM`
- `gllvm`
- `mcmcGLMM`
- `INLA`
- `inlabru`
- MCMC frameworks (JAGS, STAN, NIMBLE, greta)

`lme4` and `glmmTMB` are most commonly used.

`lme4` (Bates et al. 2015)

- Correlation between random effects
- Sparse matrices
- Modern matrix algebra libraries
- Likelihood profiling

Can be fussy about convergence

glmmTMB (Brooks et al. 2017)

- Correlation between and within random effects (e.g., spatial)
- Uses state-of-the art AD software (TMB, Kristensen et al. 2015)
- More supported distributions
  - Tweedie
  - Conway-Maxwell-Poisson
  - Zero-inflation
- Double hierarchical GLMs
- Can fit GLLVMs

Specification with formula syntax in R

- We can think of our model in the same way
  - Intercepts for categorical covariates
  - Slopes for continuous covariates
  - Interactions
- Now the "parameters" can be correlated
- With the R syntax we formulate:
  - The design matrix $\mathbf{Z}$
  - The covariance matrix $\mathbf{\Sigma}$
- Just as before: intercepts are categorical, slopes for continuous covariates

Random effects R formula

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

`y ~ (continuous and/or categorical | categorical)`

"Nested":

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

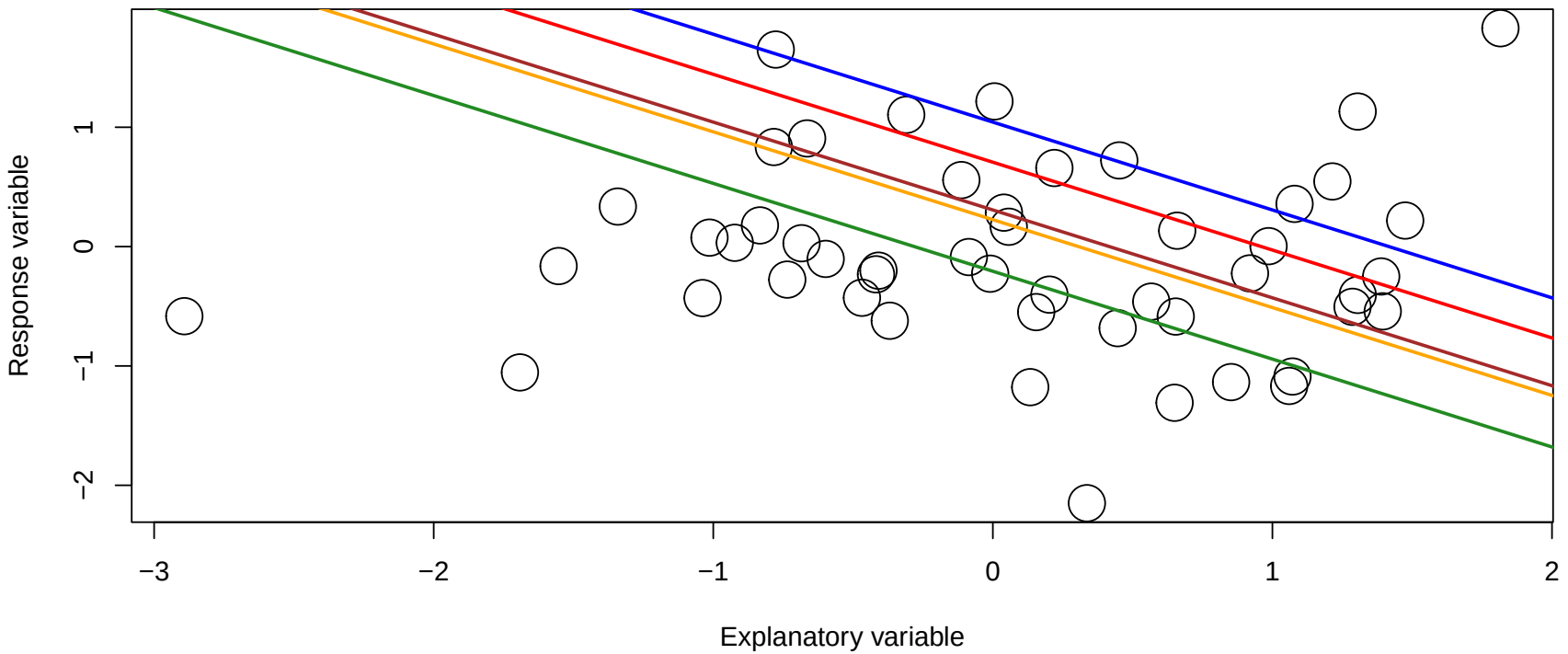
"Crossed":

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

Variation in mean abundance

$$y_{ij} = \mathbf{x}_i\boldsymbol{\beta} + \alpha_j, \qquad \text{with } \alpha_j \sim \mathcal{N}(0, \sigma^2)$$

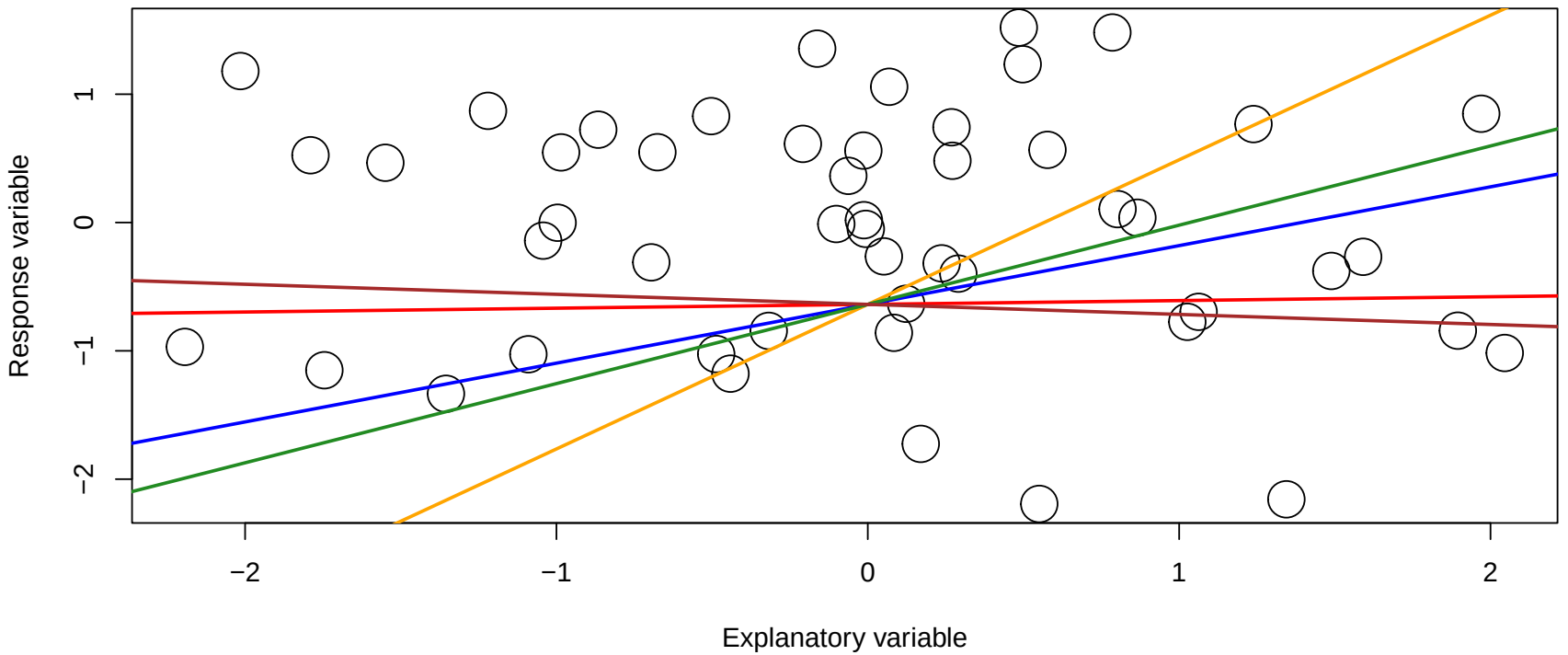
```
y ~ fixed effects + (1|species)
```



Variation in environmental responses

$$y_{ij} = \mathbf{x}_i\boldsymbol{\beta} + z_i u_j, \qquad \text{with } u_i \sim \mathcal{N}(0, \sigma^2)$$

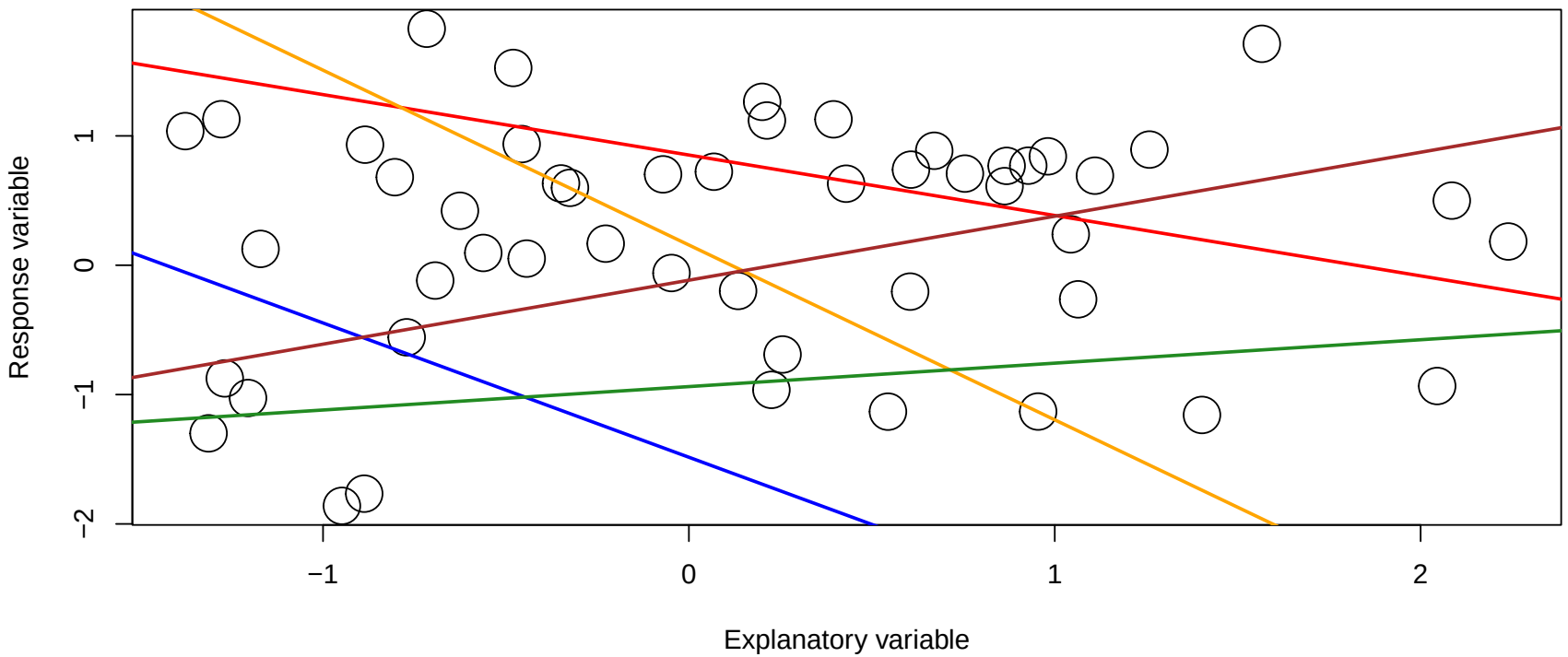
```
y ~ fixed effects + (0+covariate|species)
```



Variation of mean abundance and environmental responses

$$y_{ij} = \mathbf{x}_i\boldsymbol{\beta} + \alpha_j + z_i u_j, \text{ with } \begin{pmatrix} \alpha_j \\ u_j \end{pmatrix} \sim \mathcal{N}\left\{\mathbf{0}, \begin{pmatrix} \sigma_1^2 & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \end{pmatrix}\right\}$$

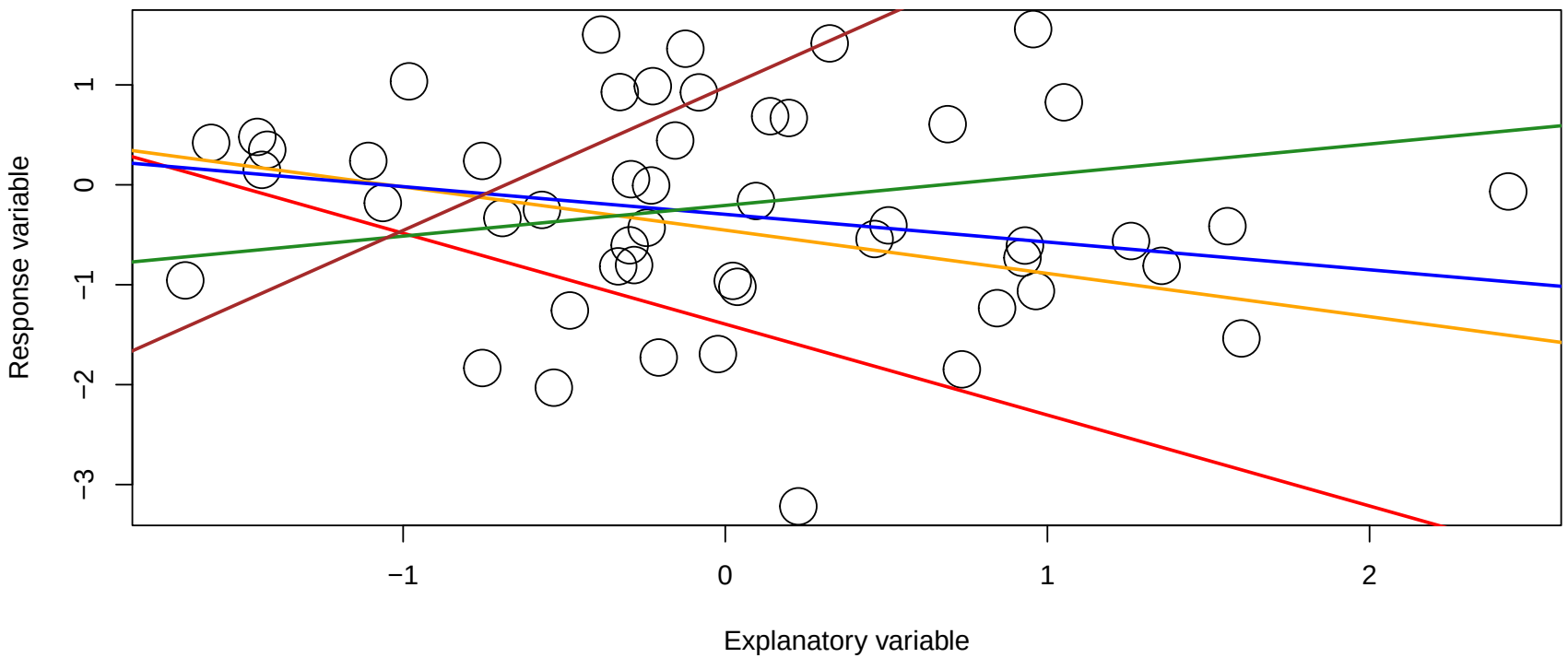
```
y ~ fixed effects + (1|species)+(0+covariate|species)
```



Correlation of mean abundance and enviromental responses

$$y_{ij} = \mathbf{x}_i\boldsymbol{\beta} + \alpha_j + z_i u_j, \text{ with } \begin{pmatrix} \alpha_j \\ u_j \end{pmatrix} \sim \mathcal{N}\left\{\mathbf{0}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\right\}$$

```
y ~ fixed effects + (random slope|random intercept)
```



Species correlation

If we fit a GLM to data of multiple species, we assume **independence**

But, observations of the same species form groups. Co-occurring species have more similar observations than for other species

In GLMM language: **observations of species exhibit correlation**

1) Part of this can be explained by shared environmental responses
2) The other part remains

Joint Species Distribution Model (JSDM)

- For community data, we want to incorporate correlation of species
- We have Multivariate data (in contrast to multivariable)
  - Each species is a response variable

$$g\{\mathbb{E}(\mathbf{y}_i|\boldsymbol{\epsilon}_i)\} = \boldsymbol{\beta}_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{8}$$

- $\boldsymbol{\epsilon}_i$ is our error
- The error takes care of the left-over variation between species
- so we assume $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$
- $\boldsymbol{\Sigma}$ is the matrix of species associations

Spiders: JSADM

```
model <- lme4::glmer(Count ~ Species + Species*soil.dry + (0+Species|Site)
                     family = "poisson", data = datalong)
```

```
## Warning in commonArgs(par, fn, control, environment()): maxfun < 10
## length(par)^2 is not recommended.

## Warning in optwrap(optimizer, devfun, start, rho$lower, control = co
## convergence code 1 from bobyqa: bobyqa -- maximum number of function
## evaluations exceeded

## Warning in (function (fn, par, lower = rep.int(-Inf, n), upper = rep
## failure to converge in 10000 evaluations

## Warning in optwrap(optimizer, devfun, start, rho$lower, control = co
## convergence code 4 from Nelder_Mead: failure to converge in 10000 ev

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$ch
## unable to evaluate scaled gradient
```

Species associations: strategies

There few different methods for dealing with this:

1) "Reduced rank" or "Factor analytic" approach (GLLVMs, tomorrow)
2) Penalize the matrix Pichler and Hartig 2023
3) ?

So, we do not usually (explicitly) fit JSDMs unconstrainedly

Convergence?

see Ben Bolker's GLMM FAQ, `lme4` page on performance, and the `glmmTMB` troubleshooting vignette

Assessing arrival at the MLE

1. Stopping criteria
   - Maximum iterations
   - Gradient close to zero
   - Relative criterion: objective function value improvement
   - Absolute criterion: objective function becomes zero (say)
2. Gradient
3. Hessian

lme4 warnings: hessian

- `Warning: Problem with Hessian check (infinite or missing values?)`
- `Warning: Hessian is numerically singular: parameters are not uniquely determined`
- `Warning: Model failed to converge: degenerate Hessian with 2 negative eigenvalues`
- `Warning: Model is nearly unidentifiable: very large eigenvalue - Rescale variables?`
- `Warning: Model is nearly unidentifiable: very large eigenvalue ratio - Rescale variables?`

Singular matrix

- determinant is zero
  - has zero eigenvalue(s)
- does not have inverse

$$\mathbf{HA} = \mathbf{I} \tag{9}$$

Numerical optimisation: best practices

1. Standardise (center and scale) explanatory variables
2. Try different optimisation routines
3. Different starting values
4. Rethink your model

Mixed-effects model troubleshooting

see Ben Bolker's GLMM FAQ

- check data for mistakes
- check model formulation
  - correct distribution and link-function
  - few random effects levels
  - few (non-zero) observations in a category
  - overly complex: drop terms with zero variances
- double-check hessian calculation (finite differences)
- use random effect as fixed effect
- '?convergence' (and see the last line "convergence issues" for large datasets)

glmmTMB diagnose()

diagnose {glmmTMB} R Documentation

diagnose model problems

Description

EXPERIMENTAL. For a given model, this function attempts to isolate potential causes of convergence problems. It checks (1) whether there are any unusually large coefficients; (2) whether there are any unusually scaled predictor variables; (3) if the Hessian (curvature of the negative log-likelihood surface at the MLE) is positive definite (i.e., whether the MLE really represents an optimum). For each case it tries to isolate the particular parameters that are problematic.

Inference

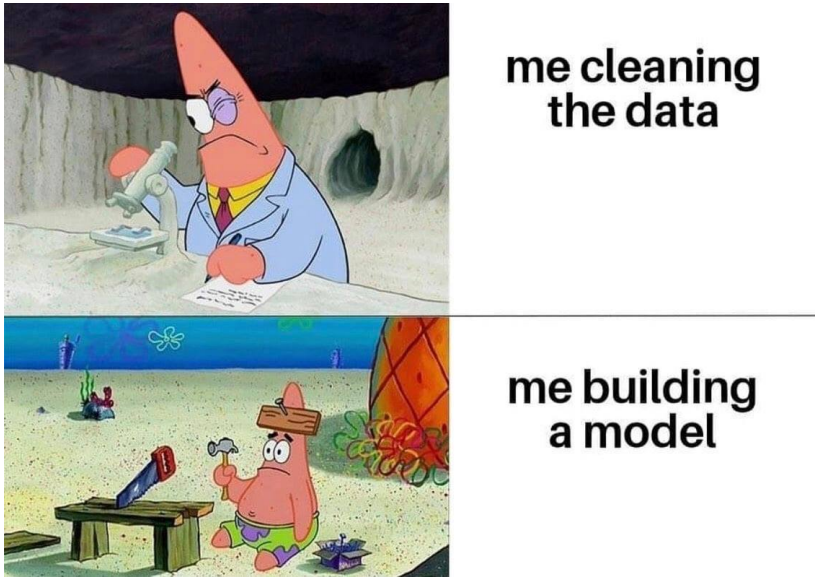
We have a good model!

Inference

We have a good model!

Now we want to do inference

- Hypothesis tests (t-test, LRT)/ P-values
- Model-selection (e.g., with AIC; Akaike 1973)
- Et cetera (R^2).



Next crisis

- Inference in glmms can be difficult
- Many tools are not well-defined (tests, information criteria, residuals)

Take away tips

No free lunch in statistics

- Scale your predictors
- Carefully consider the model structure
- Keep your model as simple as possible, but not simpler
- Different packages have different benefits
  - `glmmTMB` vs. `lme4`
- Try not to -blindly- assume approximations perform well
- Always check (residual) assumptions

There are many uses for random effects for community ecology

End