Generalised Linear Mixed Models

for data of multiple species

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

Random-effects

When to include a random effect:

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

The mixed-effects model

$$g\{\mathsf{E}(\mathbf{y}|\mathbf{u})\} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} \tag{1}$$

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

The mixed-effects model

$$g\{\mathsf{E}(\mathbf{y}|\mathbf{u})\} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} \quad (1)$$

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

Likelihood formulation: independence

$$\mathcal{L}(\mathbf{y};\Theta) = \prod_{i}^{n} f(y_i;\Theta) \tag{2}$$

We just multiply! (assumes independence)

Our new likelihood

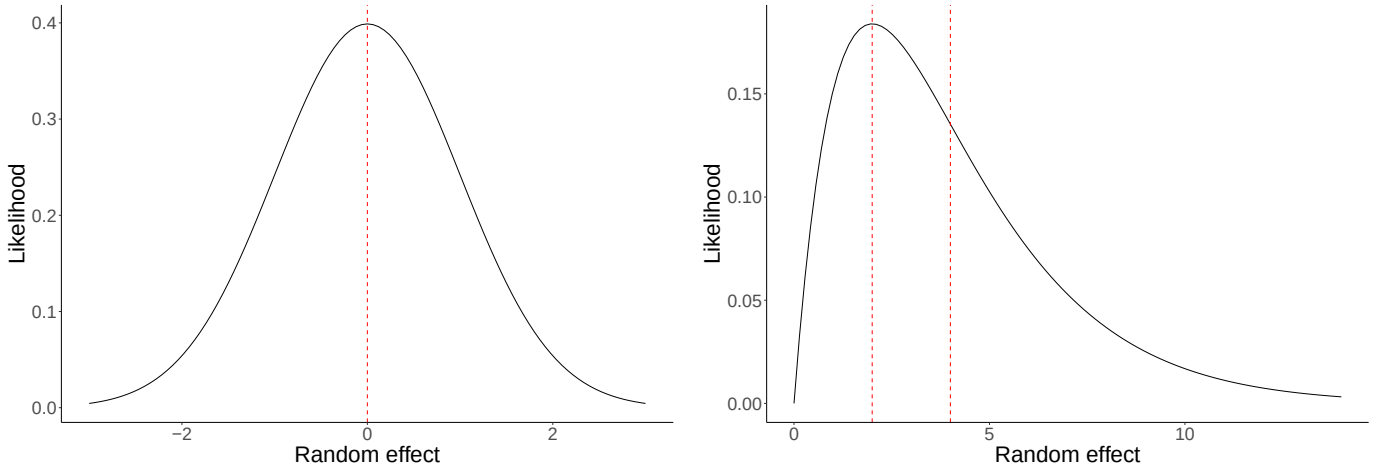
$$\mathcal{L}(\mathbf{y}; \mathbf{\Sigma}) = \int \prod_{i}^{n} f(y_i|\mathbf{u}) f(\mathbf{u}; \mathbf{\Sigma}) d\mathbf{u} \tag{3}$$

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

Estimation

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

Measure of central tendency: Mean or Mode



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
- Uses Laplace approximation

Can be fussy about convergence (and cannot fit VGLMMs)

`glmmTMB` (Brooks et al. 2017)

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

Can fit VGLMMs and GLLVMs

gllvm

- Geared to multispecies data
- Correlation between random effects and (some) within (spatial, temporal)
- Uses state-of-the art AD software (TMB, Kristensen et al. 2015)
- Random-effects matrix if assumed the same for all species
- Many supported distributions (not as many as `glmmTMB` yet)
- Uses Variational (default), Laplace, or a combination (EVA)

Least fussy in convergence

VGLMM

VGLMMs are GLMMs on steroids

- We have extra dispersion parameters
- There are many random effects
- The covariance matrix is the same across responses
- Things can get quite slow
- Optimization tends to be (even) more sensitive

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

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":

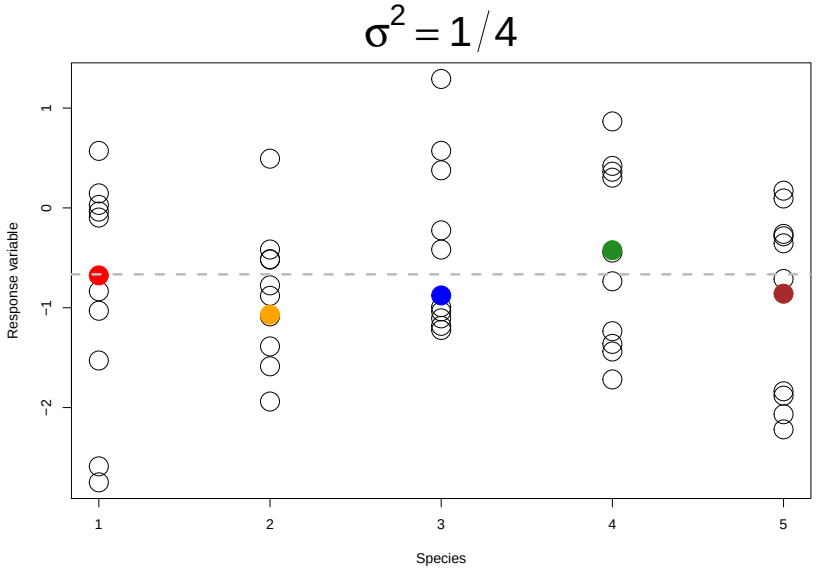
`y ~ (1|a) + (1|b)`

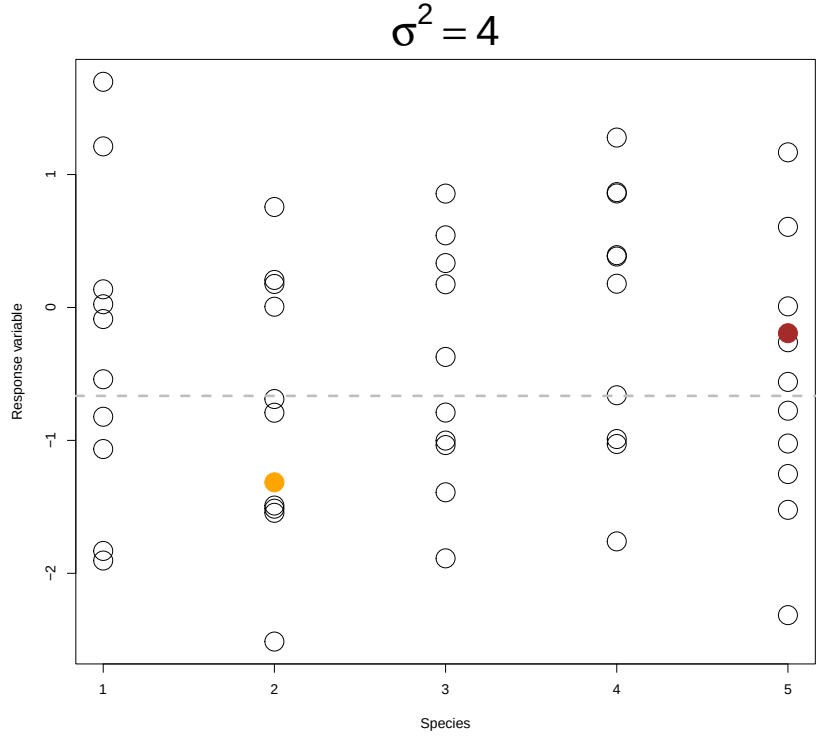
Effects within the same brackets are assumed to be correlated

Variation in mean abundance

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

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



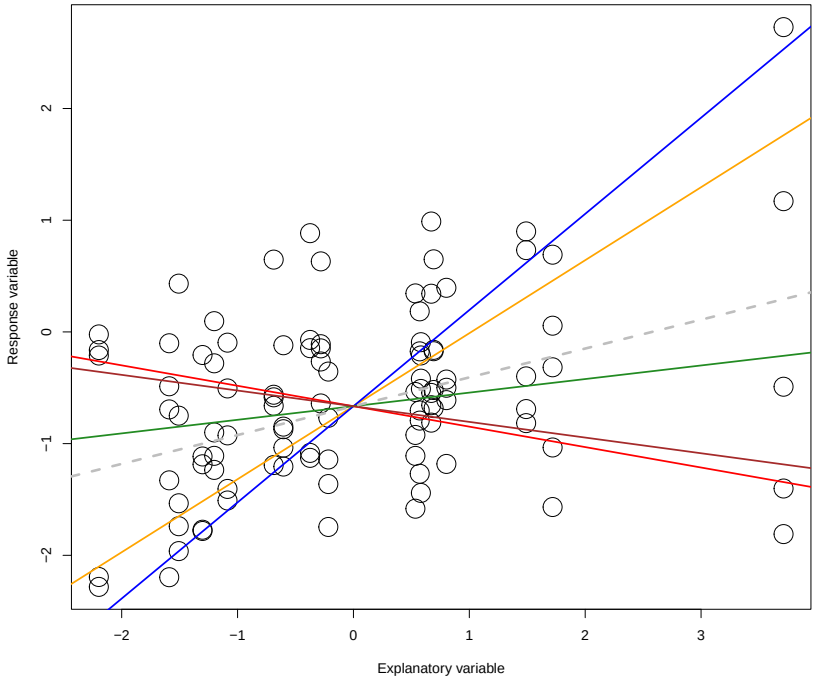


Variation in environmental responses

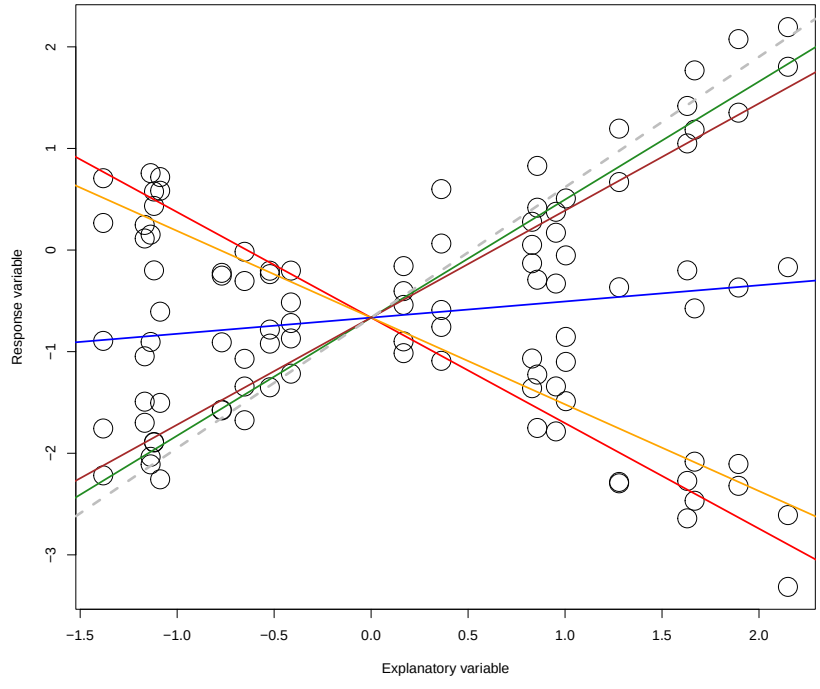
$$y_{ij} = \mu_\alpha + x_i\beta_j, \qquad \text{with } \beta_j \sim \mathcal{N}(\mu_\beta, \sigma^2_\beta)$$

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

$\sigma^2 = 1/4$



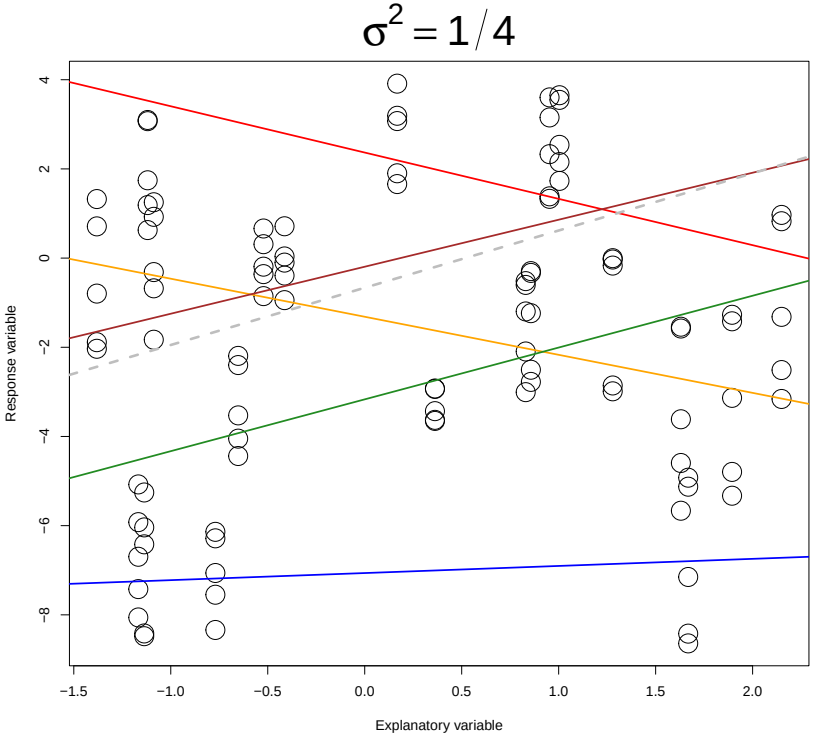
$\sigma^2 = 4$

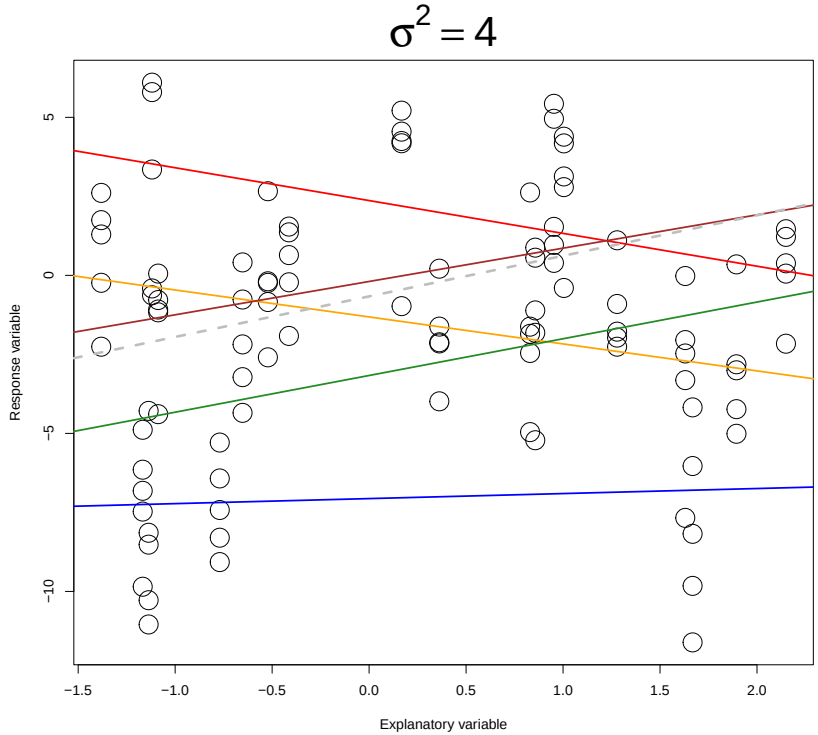


Variation of mean abundance and environmental responses

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

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



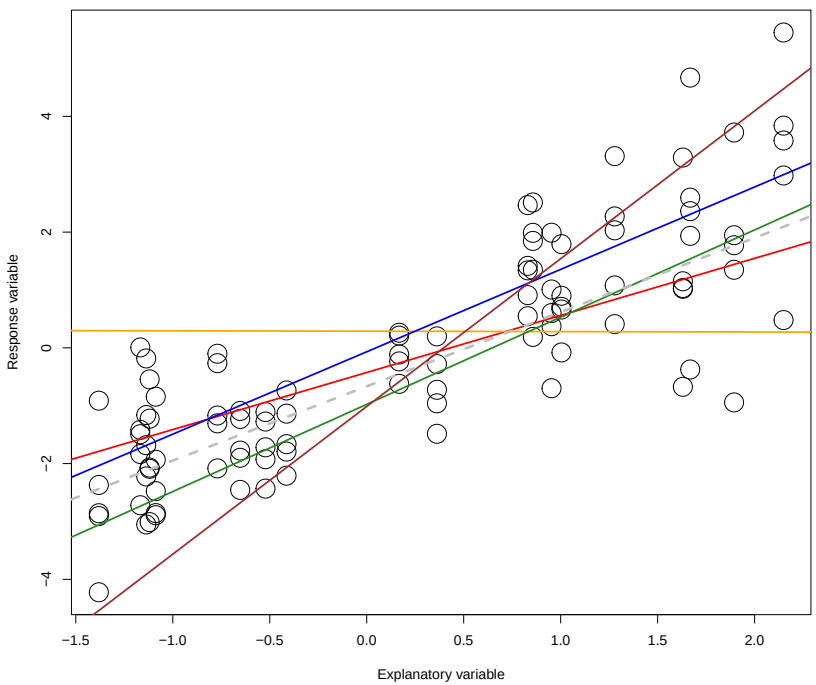


Correlation of mean abundance and enviromental responses

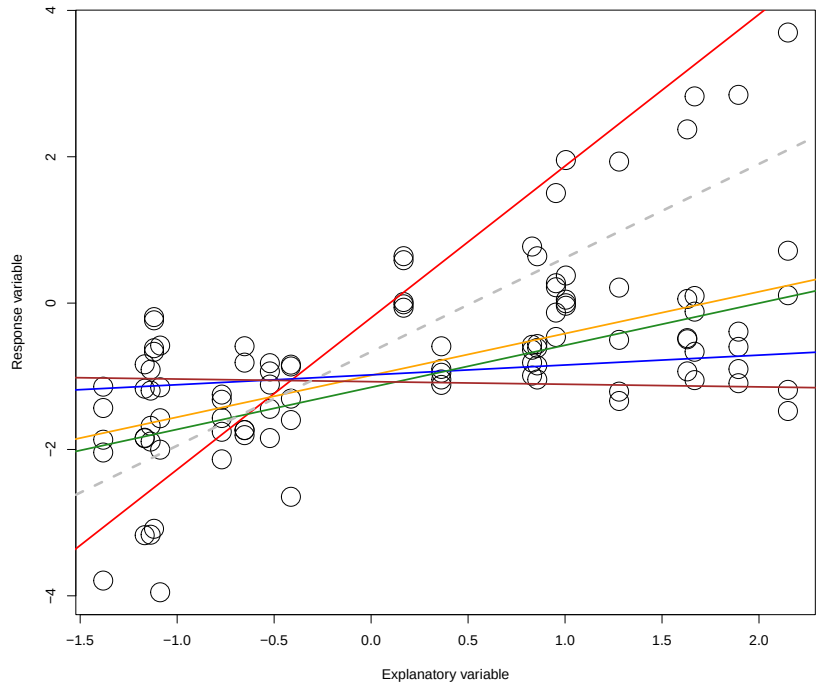
$$y_{ij} = \alpha_j + x_i\beta_j, \text{ with } \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \sigma_\alpha\sigma_\beta\rho \\ \sigma_\beta\sigma_\alpha\rho & \sigma_\beta^2 \end{pmatrix}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

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

ρ = −0.8



ρ = 0.8



Number of levels

In typical mixed-effects models, the rule of thumb is 5 levels to estimate a variance

- Here, the species are our "levels" (so we usually have many)
- Few species: not a good variance estimate
- Fewer species: not a good correlation estimate

Example 1

Remember: we have 14 species. Enough for variances, not for correlations.

In the previous presentation, we saw that NO3 had no statistically significant effect. We thus expect **shrinkage** (inverse related to variance - so low variance) from a random effects model.

Wetland macroinvertebrates

Always standardise (center, scale) your covariates with numerical optimisation, and for comparison of effects.

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

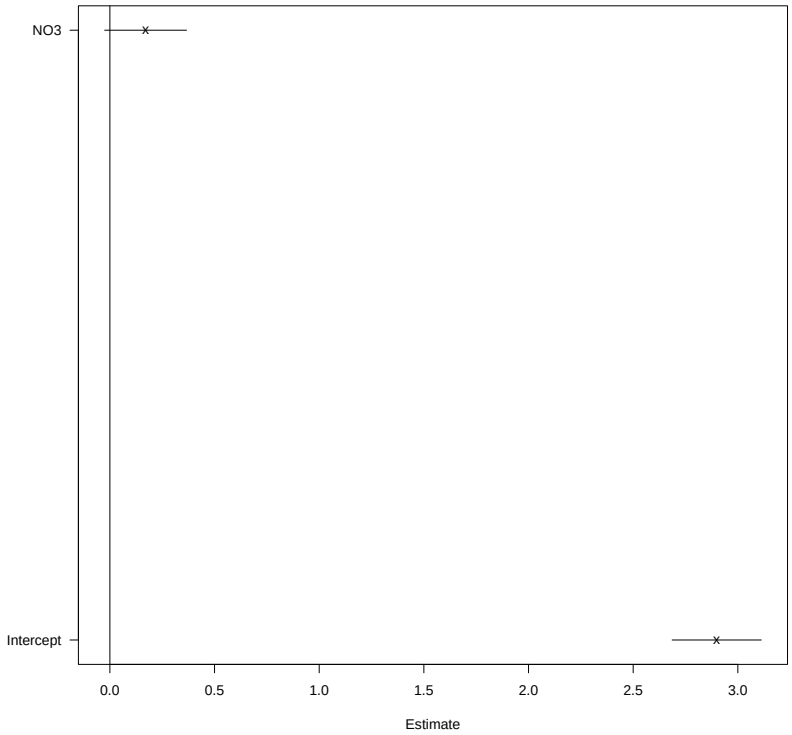
```
model1 <- gllvm(y, X = X, formula = ~NO3, beta0com = TRUE, family = "negative.binomial", num.lv = 0)
model2 <- gllvm(y, X = X, formula = ~(0+NO3|1), beta0com = TRUE, family = "negative.binomial", num.lv = 0)
```

Wetlands: summary

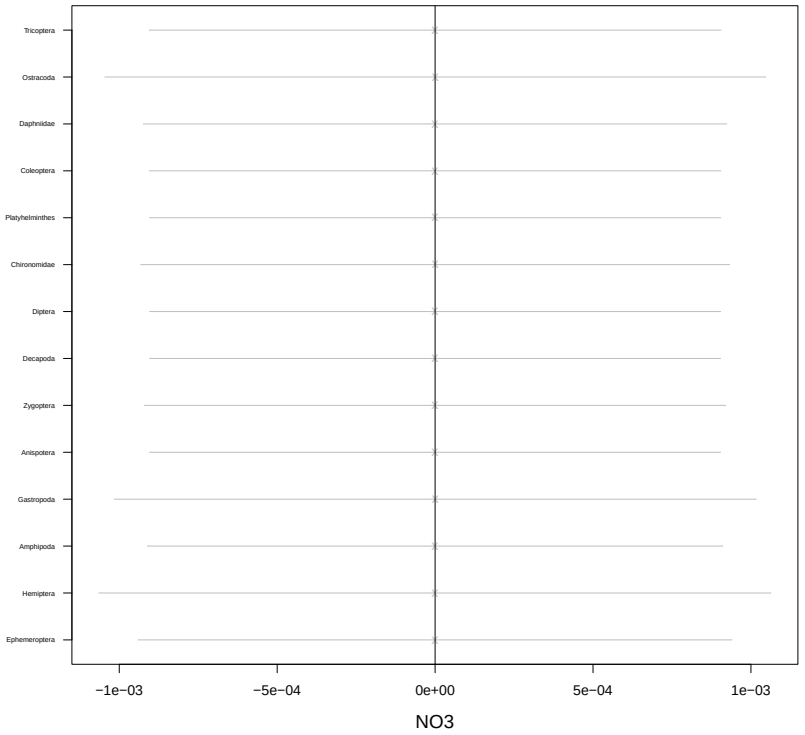
```
##
## Call:
## gllvm(y = y, X = X, formula = ~(0 + NO3 | 1), family = "negative.binomial",
##     num.lv = 0, beta0com = TRUE)
##
## Family:  negative.binomial
##
## AIC:  2791.998 AICc:  2793.222 BIC:  2864.247 LL:  -1379 df:  17
##
## Informed LVs:  0
## Constrained LVs:  0
## Unconstrained LVs:  0
##
## Formula:  ~(0 + NO3 | 1)
## LV formula:  ~ 0
## Row effect:  ~ 1
##
## Random effects:
##  Name Variance Std.Dev
##  NO3  5e-04    0.0215
##
## Coefficients predictors:
##           Estimate Std. Error z value Pr(>|z|)
## Intercept   2.8989     0.1294  22.398   <2e-16 ***
## NO3         0.1706     0.1189   1.435    0.151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results: plots

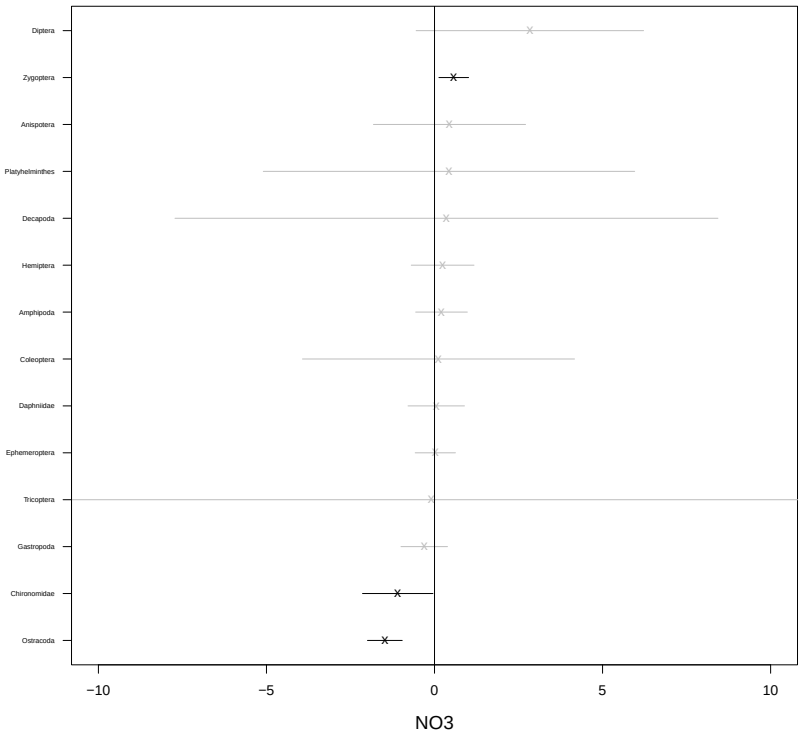
```
plot(summary(model2))
```

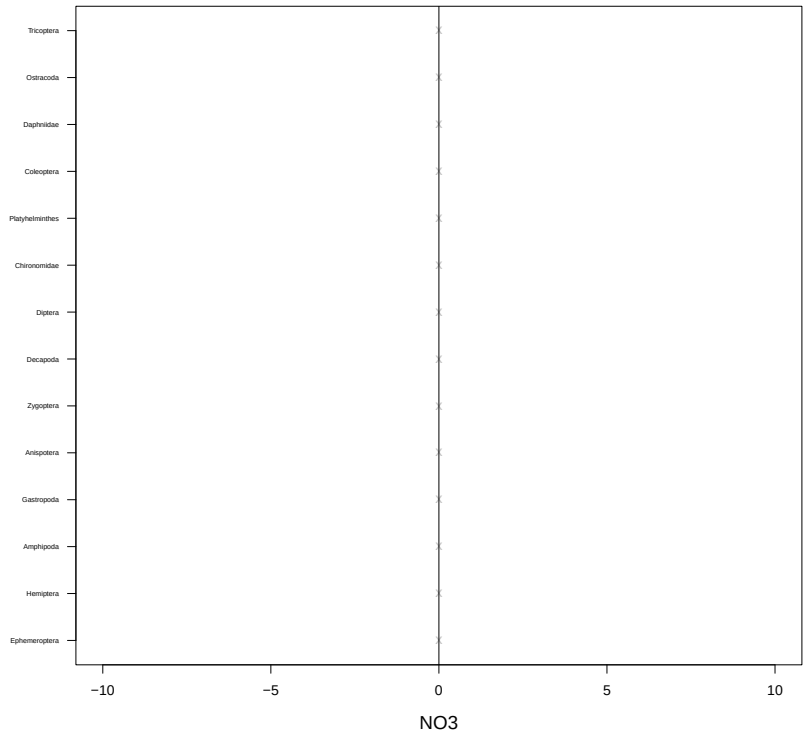



```
randomCoefplot(model2)
```



Results: comparison





We see: the "extreme" results (CI but also estimates) have been reigned in.

Wetlands: two effects

Let's add more effects

```
model3 <- gllvm(y, X = X, formula = ~(1|1)+(0+NO3|1)+(0+SO4|1),
                beta0com = TRUE, family = "negative.binomial", num.lv = 0)
```

Intraclass correlation

Here, we did not incorporate correlation. But, we can represent similarity of species in terms of (partial) intraclass correlation

$$ICC_p = \frac{\text{var(SO4)}}{\text{var}(I + NO3 + SO4)} \tag{5}$$

```
(sigma <- diag(model3$params$sigmaB))
```

```
## [1] 1.488128952 0.008644905 0.002589186
```

```
sigma[2]/sum(sigma)
```

```
## [1] 0.005765719
```

What this all means ecologically

We usually assume that an effect with higher variance, has a larger impact on the composition of a community.

What this all means ecologically

We usually assume that an effect with higher variance, has a larger impact on the composition of a community.

We can connect these statistical concepts to more familiar ecological concepts (pub. in prep.)

- Alpha diversity: (average) within-site variation
- Beta diversity: between-site variation
- Gamma diversity: total variation

And formulate our model accordingly.

Alpha diversity: within sites

$$y_{ij} = \alpha_j + x_i \beta_j$$

$$\begin{aligned} \mathsf{var}(\alpha_i + x_i \beta_j) &= \mathbb{E}(\alpha^2) + \mathbb{E}(\beta_j^2) \\ &= \sigma_\alpha^2 + x_i^2 \sigma_\beta^2 \end{aligned} \tag{6}$$

Gamma diversity: across sites

$$y_{ij} = \alpha_{ij} + x_i \beta_j$$

$$\begin{aligned} \mathsf{var}_{ij}(\alpha_j + x_i\beta_j) &= \mathbb{E}_j\{\mathsf{var}_i(\eta_{ij})\} + \mathsf{var}_j\{\mathbb{E}_i(\eta_{ij})\} \\ &= \sigma_\alpha^2 + \sigma_\beta^2\{\bar{x} + \mathsf{var}(x_i)\} \end{aligned} \tag{7}$$

Beta diversity: between sites

$$y_{ij} = \alpha_j + x_i\beta_j$$

$$\text{cov}(\alpha_j + x_i\beta_j, \alpha_j + x_k\beta_j) = \sigma^2_\alpha + x_k x_i \sigma^2_\beta$$

So, the change in beta diversity is:

$$\text{cov}(\alpha_j + x_i\beta_j, \alpha_j + x_k\beta_j) - \text{cov}(\alpha_j + x_i\beta_j, \alpha_j + x_l\beta_j) = x_i(x_k - x_l)\sigma^2_\beta$$

R^2_{GLMM} and repeatability

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013, **4**, 133–142 doi: 10.1111/j.2041-210x.2012.00261.x

A general and simple method for obtaining R^2 from generalized linear mixed-effects models

Shinichi Nakagawa[1,2]* and Holger Schielzeth[3]

BIOLOGICAL REVIEWS

Cambridge Philosophical Society

Biol. Rev. (2010), **85**, pp. 935–956.
doi: 10.1111/j.1469-185X.2010.00141.x

Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists

Shinichi Nakagawa[1]* and Holger Schielzeth[2,3]

Correlation and diversity

When calculating diversity, we need to take the correlation into consideration. For beta diversity:

$$\text{cov}(\alpha_j + \mathbf{x}_i^\top \boldsymbol{\beta}_j, \alpha_j + \mathbf{x}_k^\top \boldsymbol{\beta}_j) = \mathbf{x}_i^\top \Sigma \mathbf{x}_k \tag{8}$$

The correlation parameters represent parallel change in the community on gradients

and similarly for the change in beta diversity.

Example 2

For the Swiss birds, we have 56 species, so we can do a little more.

In the previous presentation, we saw that the effect of slope was a statistically significant. Treated as random, it is not.

Example 2: correlated effects

```
model6 <- gllvm(y, X = X, formula = ~(slp+asp|1),
                family = "binomial", num.lv = 0,
                beta0com = TRUE)
```

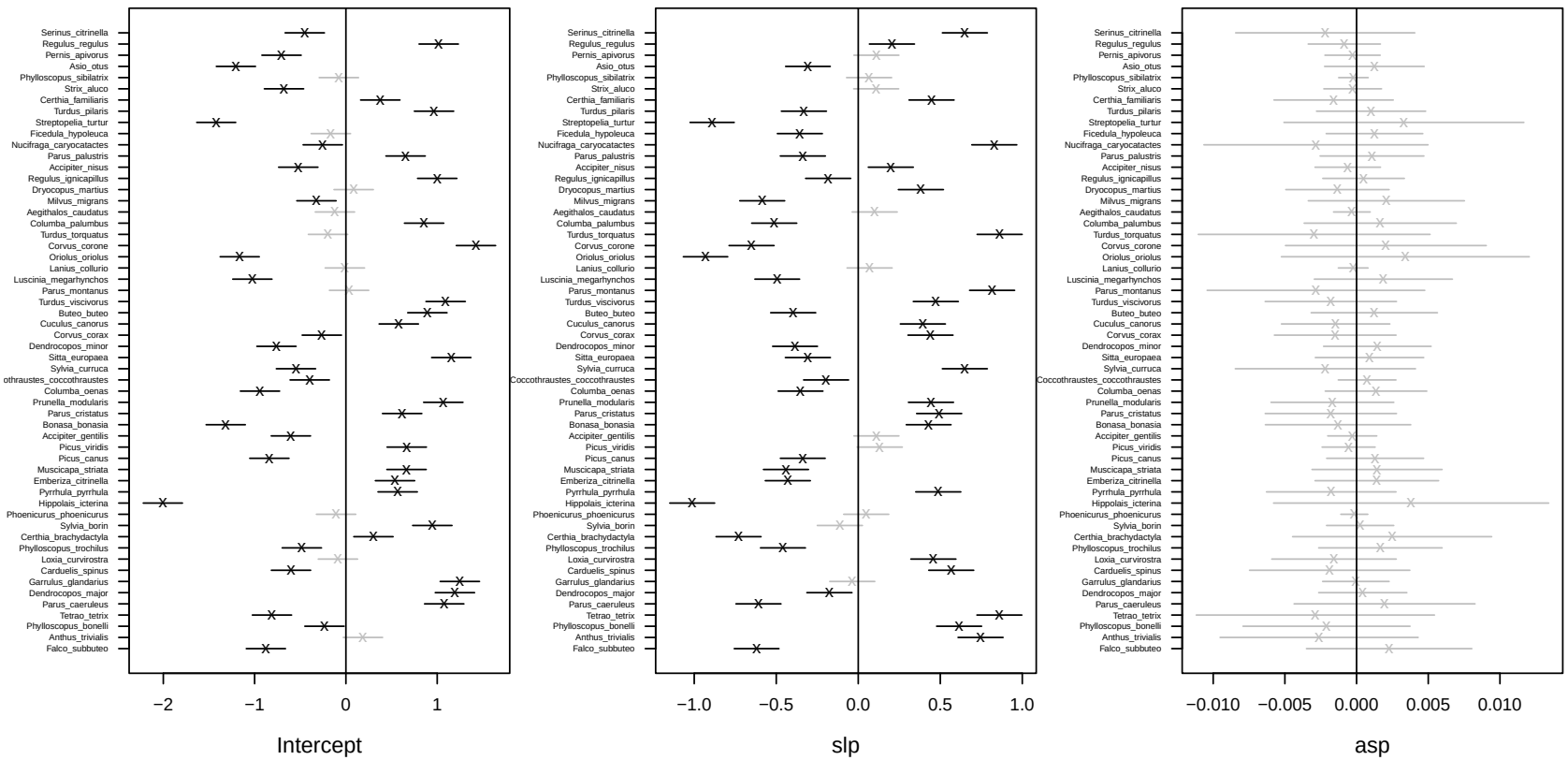
Perhaps, birds prefer flat areas in the sun.

Example 2: correlated effects

```
##
## Call:
## gllvm(y = y, X = X, formula = ~(slp + asp | 1), family = "binomial",
##     num.lv = 0, beta0com = TRUE)
##
## Family:  binomial
##
## AIC:  132304.1 AICc:  132304.1 BIC:  132392.9 LL:  -66143 df:  9
##
## Informed LVs:  0
## Constrained LVs:  0
## Unconstrained LVs:  0
##
## Formula:  ~(slp + asp | 1)
## LV formula:  ~ 0
## Row effect:  ~ 1
##
## Random effects:
##  Name      Variance Std.Dev Corr
##  Intercept 0.6541   0.8087
##  slp       0.2525   0.5025   0.0802
```

Example 2: correlated effects

```
randomCoefplot(model6)
```



Intraclass correlation

We can also do this with correlated effects. var(a+b) = var(a)+var(b) + 2cov(a+b)

$$ICC_p = \frac{\text{var(slope)}}{\sum \Sigma_\beta} \tag{9}$$

Intraclass correlation

We can also do this with correlated effects. var(a+b) = var(a)+var(b) + 2cov(a+b)

$$ICC_p = \frac{\text{var(slope)}}{\sum \Sigma_\beta} \tag{9}$$

```
Sigma <- model6$params$sigmaB
sigma <- diag(model6$params$sigmaB)
sigma[2]/sum(Sigma)
```

```
## [1] 0.2604024
```

The model predicts that 26% of the change in community composition is due to slope.

Steeper areas have lower species richness (due to the negative average effect).

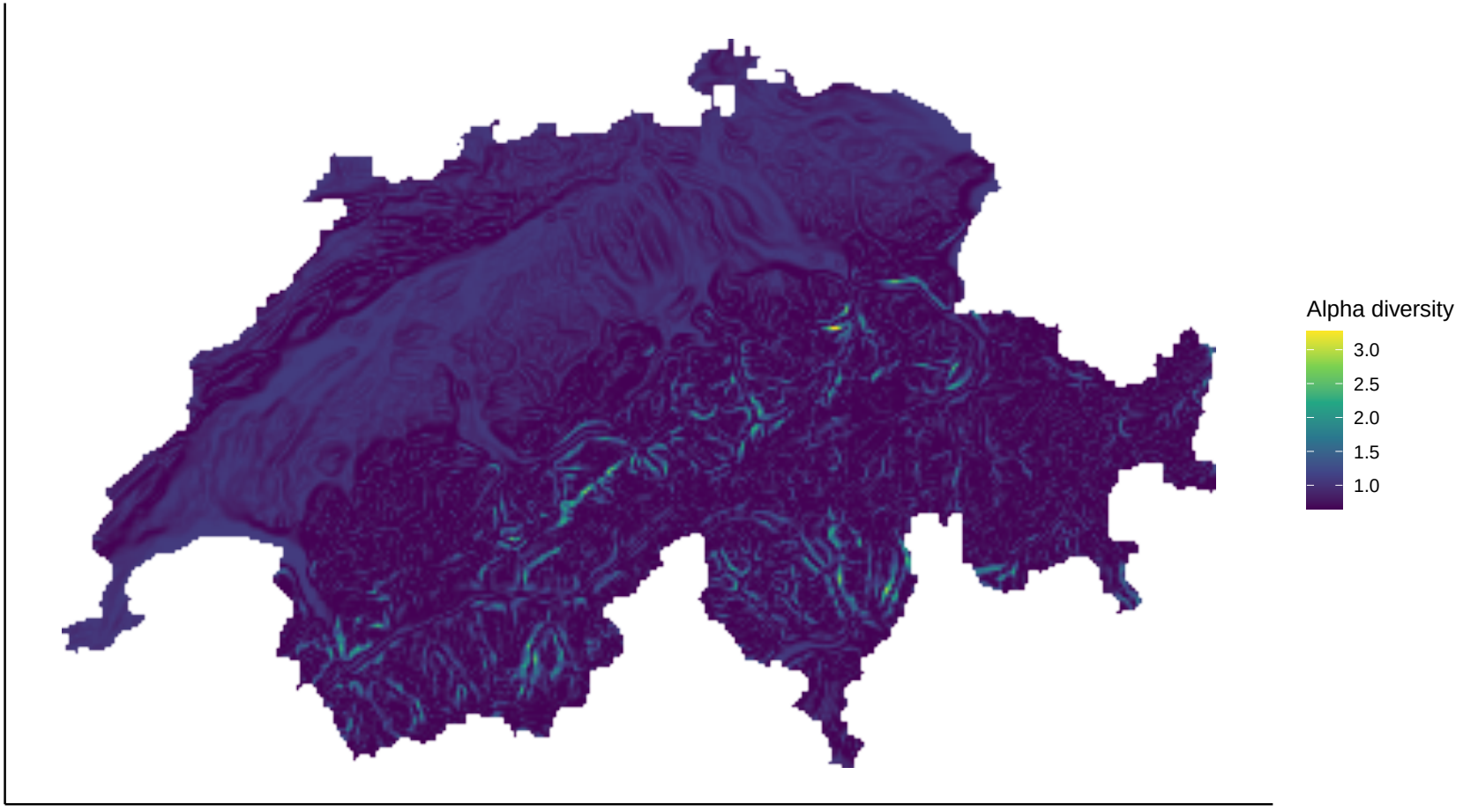
Predicting diversity

As we have explicit equations for diversity, we can predict is using the model.

We take the parameter estimates for the variances, plug-in new values for the covariates (account for scaling), and put it on a map.

Predicting alpha diversity

Bird alpha diversity (due to mean abundance, aspect and slope)



Alpha diversity

Alpha diversity can more flexibly be estimated as:

```
model7 <- gllvm(y, X = cbind(X, Site = factor(1:nrow(y))),
                formula = ~diag(1|Site)+(slp+asp|1),
                family = "binomial", num.lv = 0,
                beta0com = TRUE, Ab.struct = "diagonal",
                optim.method = "L-BFGS-B", sd.errors = FALSE)
```

but note that site-specific variance is not equal to alpha diversity.

Gamma diversity

Gamma diversity can be more flexibly estimated as:

```
model8 <- gllvm(y, X = cbind(X, Site = factor(1:nrow(y))),
                formula = ~(1|1) + diag(1|Site)+(slp+asp|1),
                family = "binomial", num.lv = 0, beta0com = TRUE,
                Ab.struct = "diagonal", optim.method = "L-BFGS-B",
                sd.errors = FALSE)
```

but also nicely corresponds to `row.eff = ~(1|site)` perhaps

Philosophy

So, our attitude becomes a little different. We can adjust the model, to formulate a certain measure of diversity.

Summary

Everything gets more difficult when we use mixed-effects models:

- Wald-statistic and p-values no longer (really) apply
- Hypothesis test does not always work well (on the boundary)
- Model selection does not usually work well
- Residuals are harder to define
- We should test extra (random effect) assumptions

Summary

Everything gets more difficult when we use mixed-effects models:

- Wald-statistic and p-values no longer (really) apply
- Hypothesis test does not always work well (on the boundary)
- Model selection does not usually work well
- Residuals are harder to define
- We should test extra (random effect) assumptions

And even more so for VGLMMs

Take away tips

No free lunch in statistics

- There are loads of fun things to do with VGLMMs
- Particularly for diversity calculations
- Keep your model as simple as possible, but not simpler
- Different packages have different benefits
  - `gllvm` vs. `glmmTMB` vs. `lme4`

There are many uses for random effects for community ecology

End