Multispecies mixed effects models

Jenni Niku

University of Jyväskylä

Mixed effects models

Mixed effects model = Consists of both fixed and random effects.

Fixed effects are common across observational units or species.

Random effects account variability between species, sites, units or any groups or clusters relevant to specific data.

- 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

Mixed effects models

GLMs assume independency, the mixed effect models can account dependent observations.

For instance dependency due:

- Hierarchical/nested sampling designs.
- Correlated/repeated measurements.
- Correlation between species.

The mixed effects model: simplified form

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

- g link function
- $E(\mathbf{y}|\mathbf{u})$ conditional mean
- $\mathbf{X}$ Fixed effects design matrix
- β Fixed effects parameter vector
- $\mathbf{Z}$ Random effects design matrix
- $\mathbf{u}$ Random effects parameter vector

Likelihood function

$$L(\mathbf{y}, \mathbf{\Sigma})\} = \int \prod_i f(y_i|\mathbf{u}) f(\mathbf{u}; \mathbf{\Sigma}) \quad (2)$$

- $f(y_i|\mathbf{u})$ responses distribution conditional to random effects $\mathbf{u}$
- $f(\mathbf{u}; \mathbf{\Sigma})$ distribution of random effects

Estimation

For non-normal data, integral has not closed form solution, so approximations provide solutions.

Possible approximate methods can be:

- Penalized quasi-likelihood methods
- Adaptive GH quadrature
- Laplace approximation (gllvm)
- Variational approximations (gllvm)

Or Bayesian MCMC as alternative to frequentist methods.

R packages

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

There are several R-packaces available. lme4 and glmmTMB are among most well known.

Here we focus on gllvm.

Mixed effects models in gllvm

- Package is designed (but not restricted to) for multivariate ecological data
- Allows random effects for responses/species or observational units/clusters of units
- Correlation between random effects and (some) within (spatial, temporal)
- Many supported distributions
- Estimation with ML approximation methods: Variational (VA), Laplace (LA) or Extended variational (EVA)
- Utilize automatic differentiation & C++ to enhance computation (TMB, Kristensen et al. 2015)

Random effects in gllvm

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

- formula: for species-specific fixed/random effects
- row.eff: for species community level fixed/random effects
- lv.formula: for effects in the ordination (this will be considerd later)

Species specific random effects in gllvm: formula

Model:

$$\eta_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta}_\mathbf{j} \tag{3}$$

Now, $\boldsymbol{\beta}_\mathbf{j}$ is a random effect (intercept or slope). Specifically, $\boldsymbol{\beta}_\mathbf{j} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- $\boldsymbol{\mu}$ the community level effect, mean for the random effects
- $\boldsymbol{\Sigma}$ variation in species specific random effects

Species specific random effects in gllvm: formula

Examples of how formula works in R. Generally:

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

(the 0 omits an intercept term)

"Nested":

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

"Crossed":

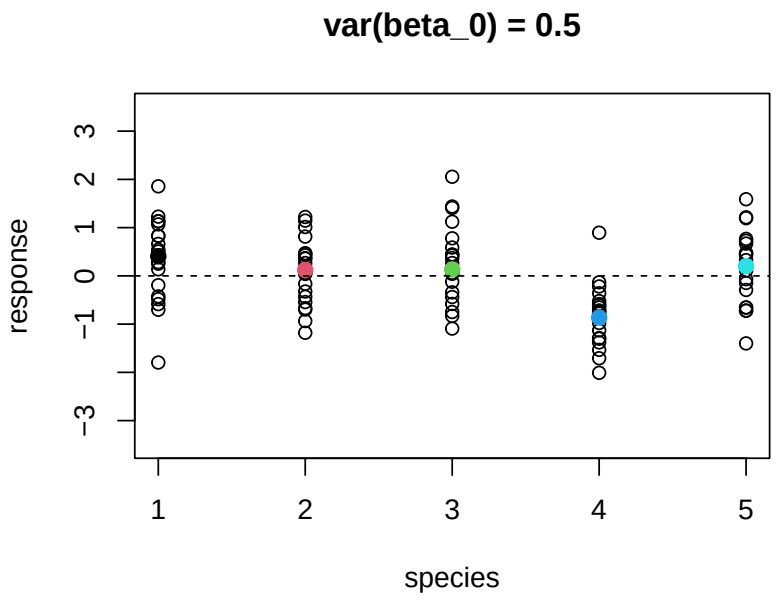
```
y ~ (1|a) + (1|b)
```

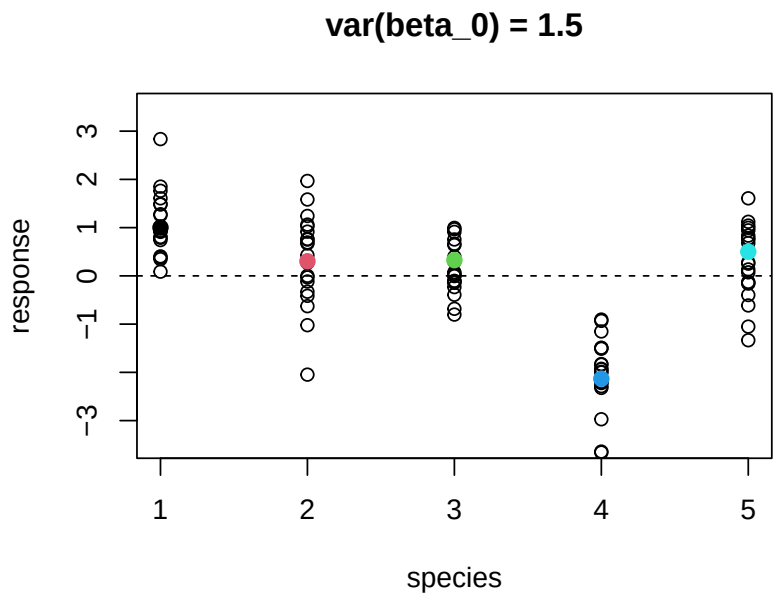
Effects within the same brackets are assumed to be correlated

formula: species specific mean abundance random

$$\eta_{ij} = \beta_{0j}, \text{ with } \beta_{0j} \sim N(\mu_0, \sigma_0^2) \quad (4)$$

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

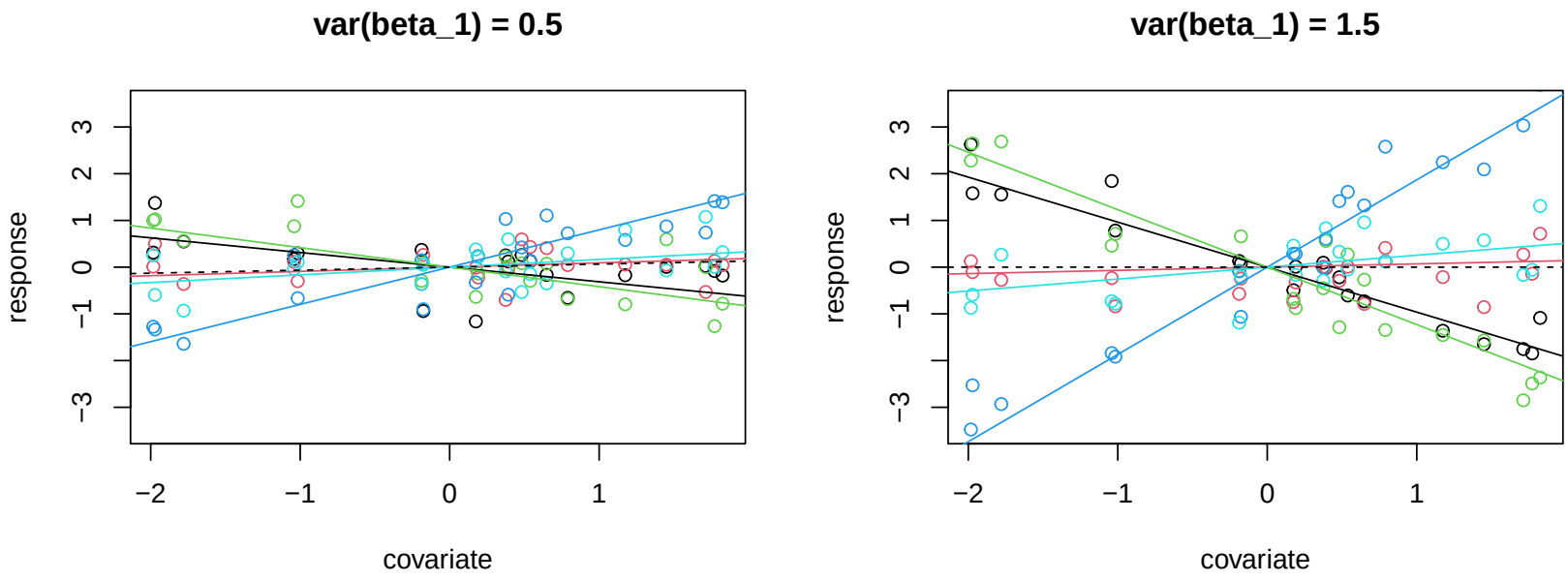




formula: environmental responses random

$$\eta_{ij} = \beta_0 + x_i^\top \beta_{1j}, \text{ with } \beta_{1j} \sim N(\mu_1, \sigma_1^2) \quad (5)$$

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

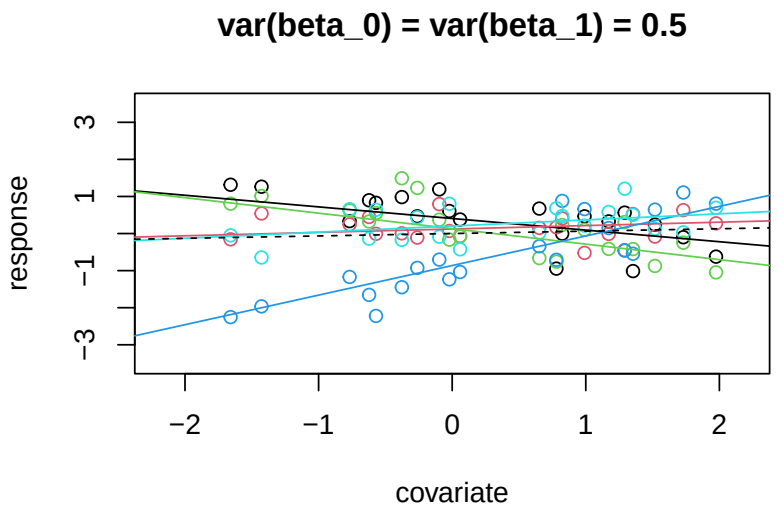


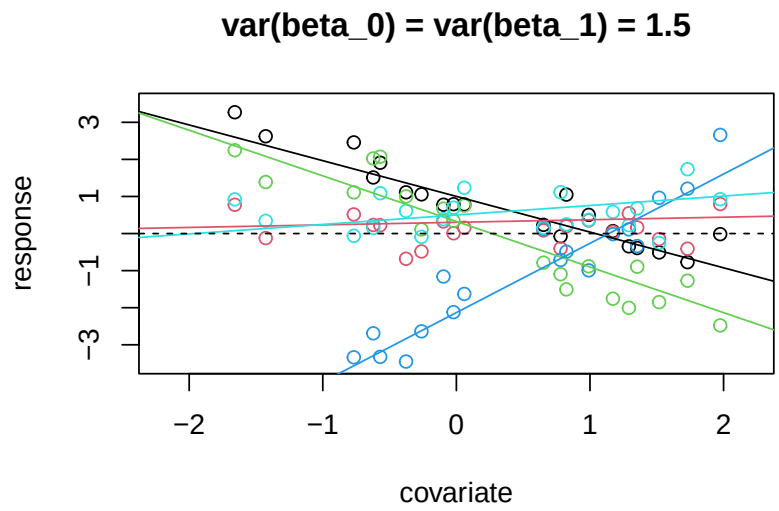
formula: mean and environmental random

$$\eta_{ij} = \beta_{0j} + x_i^\top \beta_{1j},$$

$$\text{with } \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} \right) \tag{6}$$

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



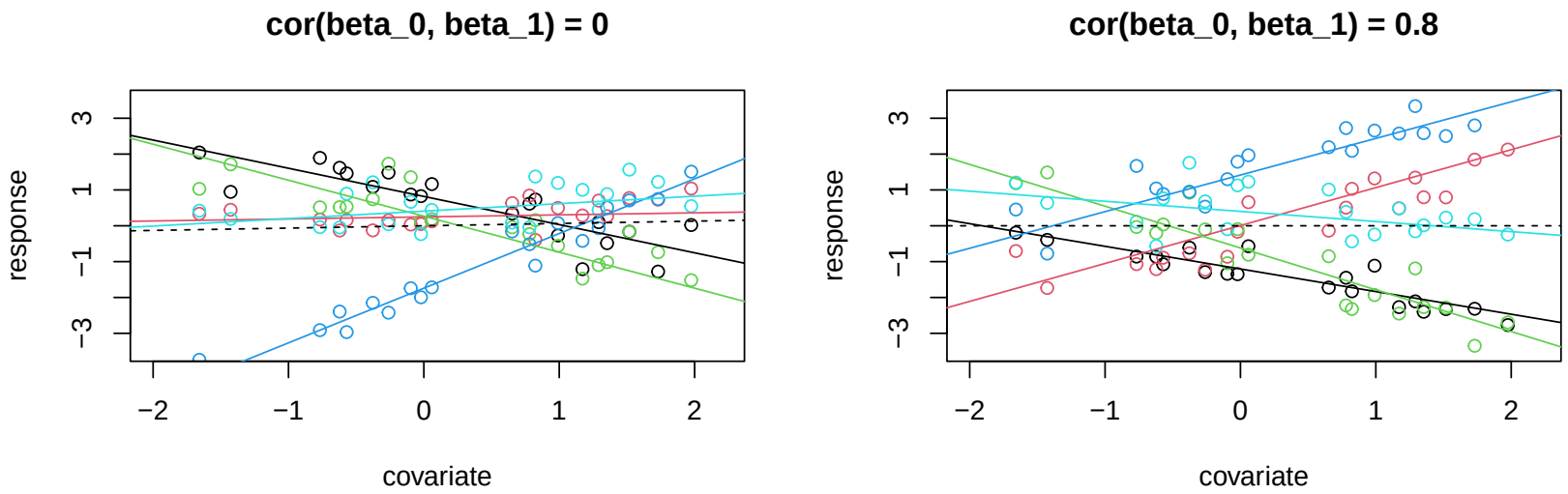


formula: both random and correlated

$$\eta_{ij} = \beta_{0j} + x_i^\top \beta_{1j},$$

$$\text{with } \begin{bmatrix}\beta_{0j}\\ \beta_{1j}\end{bmatrix} \sim N\left(\begin{bmatrix}\mu_0\\ \mu_1\end{bmatrix}, \begin{bmatrix}\sigma_0^2 & \sigma_0\sigma_1\rho\\ \sigma_0\sigma_1\rho & \sigma_1^2\end{bmatrix}\right) \quad (7)$$

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



Number of levels

In typical mixed-effects models, grouping variable should have at least 5 levels to estimate a variance. (enough information to deviate the group level effect from the population level mean)

In multivariate mixed effects model with `gllvm` and `formula`

- The species are the "levels"
- More species » better variance estimate for random effects

In multivariate mixed effects model with `gllvm` and 'row.eff"

- The groups are the "levels"

Example 1

Consider as an example ant data with 41 ant species from 30 sites.

Data has some species that are observed only a few times, thus some 'extreme' effects for environmental covariates can came up with fixed effects model.

Let's demonstrate the shrinking effect of the model for covariates.

Example 1: code for a model fit

```
data(antTraits, package = "mvabund")
y = antTraits$abund
X = scale(antTraits$env)
```

```
model_fixed <- gllvm(y, X, formula = ~Canopy.cover,
  family = "negative.binomial", num.lv = 0)
model_random <- gllvm(y, X, formula = ~(0 + Canopy.cover|1),
  family = "negative.binomial", num.lv = 0)
```

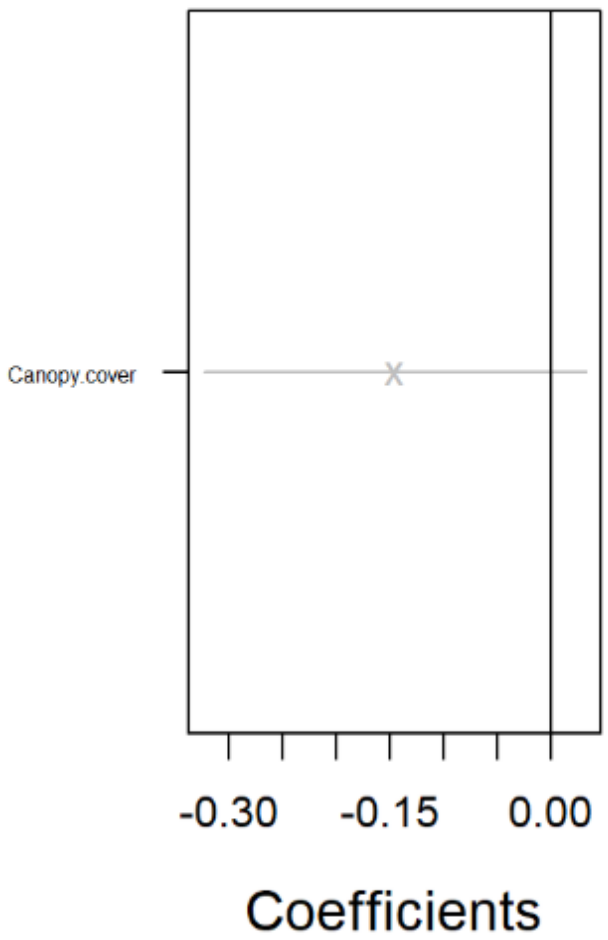
Example 1: summary

```
summary(model_random)
```

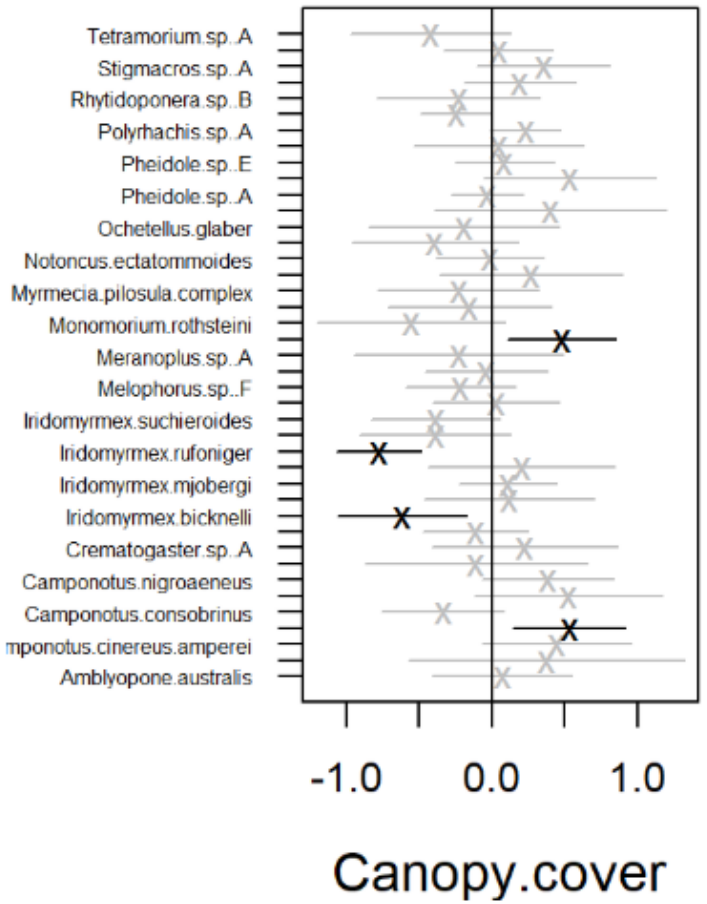
```
##
## Call:
## gllvm(y = y, X = X, formula = ~(0 + Canopy.cover | 1), family = "negative.binomial",
##     num.lv = 0)
##
## Family:  negative.binomial
##
## AIC:  4080.533 AICc:  4093.005 BIC:  4510.174 LL:  -1956.3 df:  84
##
## Informed LVs:  0
## Constrained LVs:  0
## Unconstrained LVs:  0
##
## Formula:  ~(0 + Canopy.cover | 1)
## LV formula:  ~ 0
## Row effect:  ~ 1
##
## Random effects:
##  Name         Variance Std.Dev
##  Canopy.cover 0.4051   0.6365
##
## Coefficients predictors:
##              Estimate Std. Error z value Pr(>|z|)
## Canopy.cover -0.14540    0.09051  -1.607    0.108
```

Example 1: Coefficients

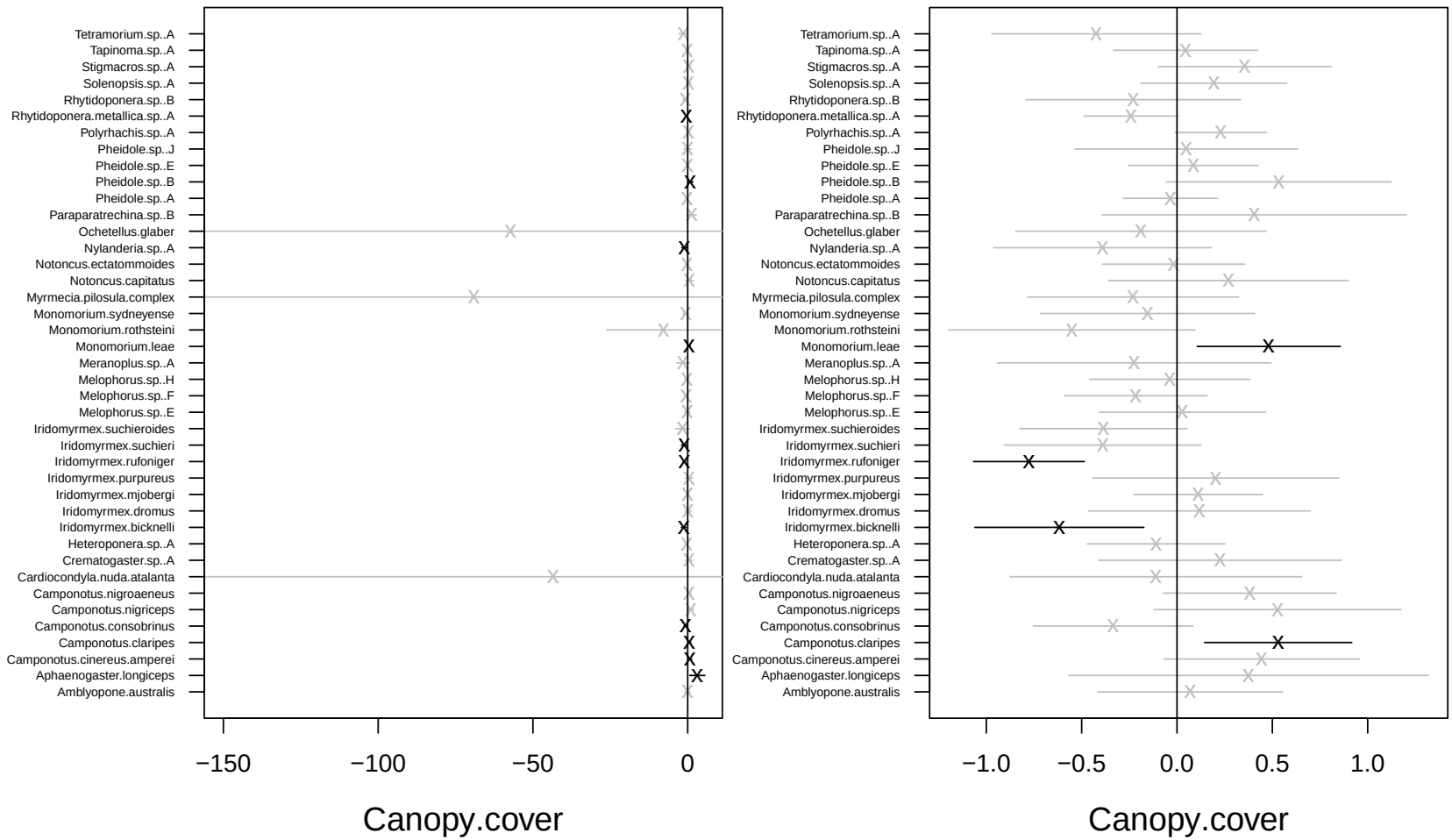
```
coefplot(model_random, order=FALSE)
```




```
randomCoefplot(model_random)
```



Example 1: Coefficient comparison



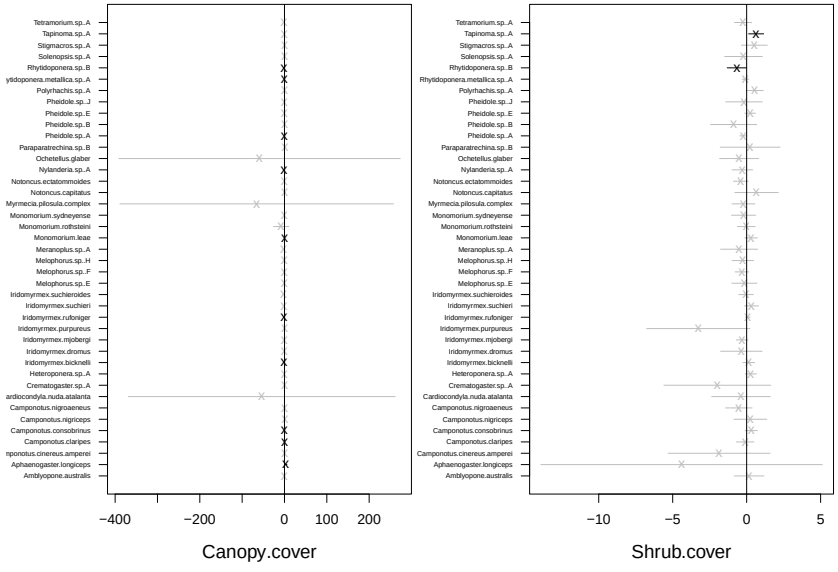
Extreme values shrinked towards zero.

Example 1: Two covariates

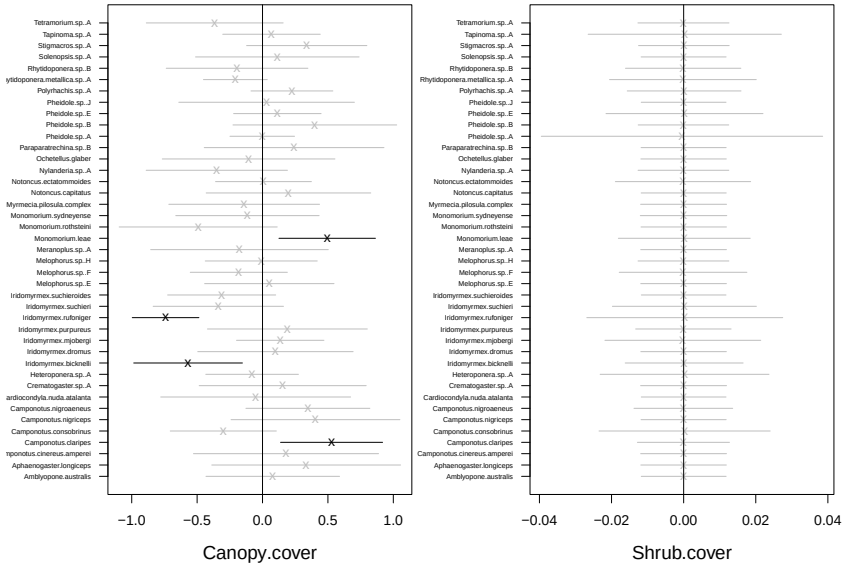
```
model_fixed <- gllvm(y, X, formula = ~Canopy.cover+Shrub.cover,
  family = "negative.binomial", num.lv = 0)
model_random <- gllvm(y, X, beta0com = TRUE,
  formula = ~(1|1) + (0 + Canopy.cover|1) + (0 + Shrub.cover|1),
  family = "negative.binomial", num.lv = 0)
```

Example 1: Coefficients comparison

```
coefplot(model_fixed, order=FALSE)
```




```
randomCoefplot(model_random,
  which.Xcoef = c("Canopy.cover", "Shrub.cover"))
```



One shrinked towards zero, other almost exactly zero as variable is not so “relevant”.

Example 1: Model comparison

Potentially shrinkage can improve model's predictive performance in situation like this.

```
goodnessOfFit(object = model_fixed, measure = c("cor", "RMSE"))
```

```
## $cor
## [1] 0.1485895

## $RMSE
## [1] 25.28328
```

```
goodnessOfFit(object = model_random, measure = c("cor", "RMSE"))
```

```
## $cor
## [1] 0.7086864
## $RMSE
## [1] 3.15118
```

Community level random effects in gllvm: row.eff

Model:

$$\eta_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta}_\mathbf{j} + \mathbf{Z}_i \boldsymbol{\alpha} \tag{8}$$

now, $\boldsymbol{\alpha}$ is a vector of community level random effects. $\mathbf{Z}$ is a design matrix for random effects and $\mathbf{Z}_i$ is i:th row of it. Specifically, $\boldsymbol{\alpha} \sim N(0, \boldsymbol{\sigma}^2)$

- $\boldsymbol{\mu}$ the community level effect, mean for the random effects
- $\boldsymbol{\Sigma}$ variation in species specific random effects

Community level random effects in gllvm: row.eff

Controlled with argument `row.eff`

- `row.eff` is a mixed-effects formula
- `row.eff = (1|group) + X1` is a random effect and a fixed effect
- `row.eff = "random"` quick call for incorporating row-specific random intercepts
- Can also incorporate spatial or temporal random effects, (we will get back to that later)

row.eff: community level random intercept

$$\eta_{ij} = \beta_{0j} + \alpha_i, \text{ with } \alpha_i \sim N(0, \sigma^2_\alpha), \tag{9}$$

where α_i is a row-specific community level random intercept.

```
row.eff = "random"
```

row.eff: multiple community level random intercepts

Row effect allows multiple structured community level random effects. This is useful for hierarchical sampling designs, for instance. Example: Hierarchical sampling design with sites and multiple plots within site.

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \alpha_{s(i)} + \alpha_{p(i)}, \\ \text{with } \alpha_{s(i)} &\sim N(0, \sigma_s^2), \\ \text{and } \alpha_{p(i)} &\sim N(0, \sigma_p^2) \end{aligned} \tag{10}$$

- Variation between sites: σ_s^2
- Variation between plots within site: σ_p^2
- $s(i) = s(i')$, if samples i and i' from same site

```
row.eff = ~ (1|site) + (1|plot), studyDesign = studyDesign,
```

`studyDesign` is a data frame with variables defining site and plot.

row.eff: correlation due random intercepts

This creates correlation structure within sites

$$\frac{\sigma_s^2}{\sigma_s^2+\sigma_p^2+\text{error term by distribution}}$$

Example 2

* Microbial community data consisting of abundances of bacteria species. * Collected from a total of 8 sites from three regions: Kilpisjärvi, Mayrhofen and Ny-Ålesund. * 4 to 8 soil samples from each site (Total of samples is 56). * Three continuous environmental variables (pH, phosphorous and soil organic matter).

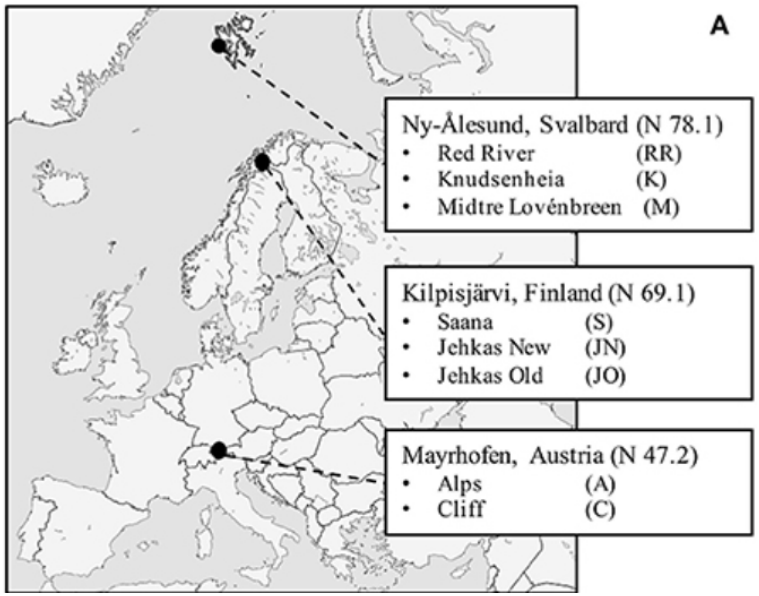
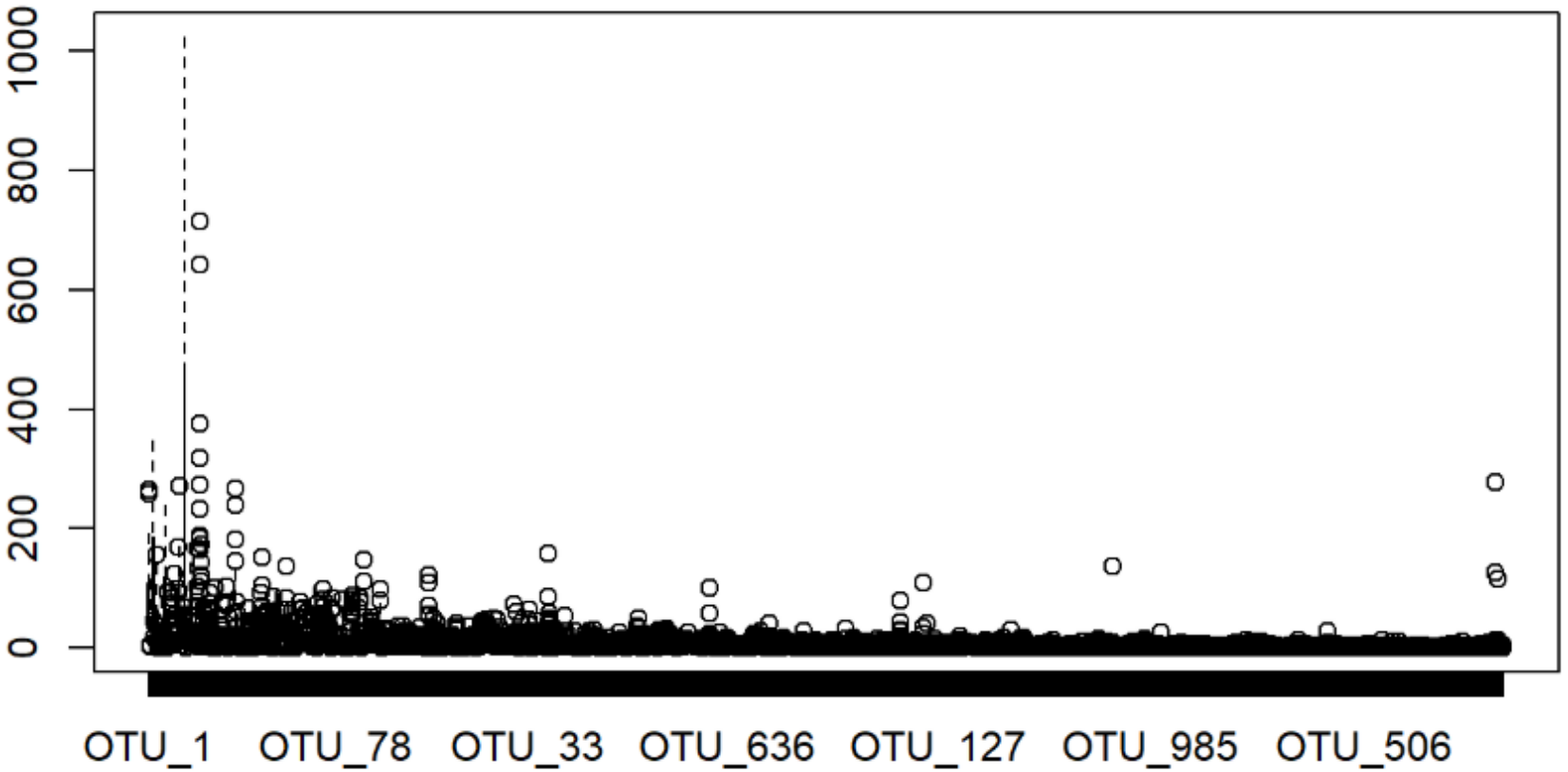


Figure: Kumar, M., Brader, G., Sessitsch, A., Mäki, A., van Elsas, J.D., and Nissinen, R. (2017). Plants Assemble Species Specific Bacterial Communities from Common Core Taxa in Three Arcto-Alpine Climate Zones. Frontiers in Microbiology, 8:12.

Example 2: Data

Community level random effects can be used to account hierarchical sampling design (Site - Sample).



Example 2: Model fit

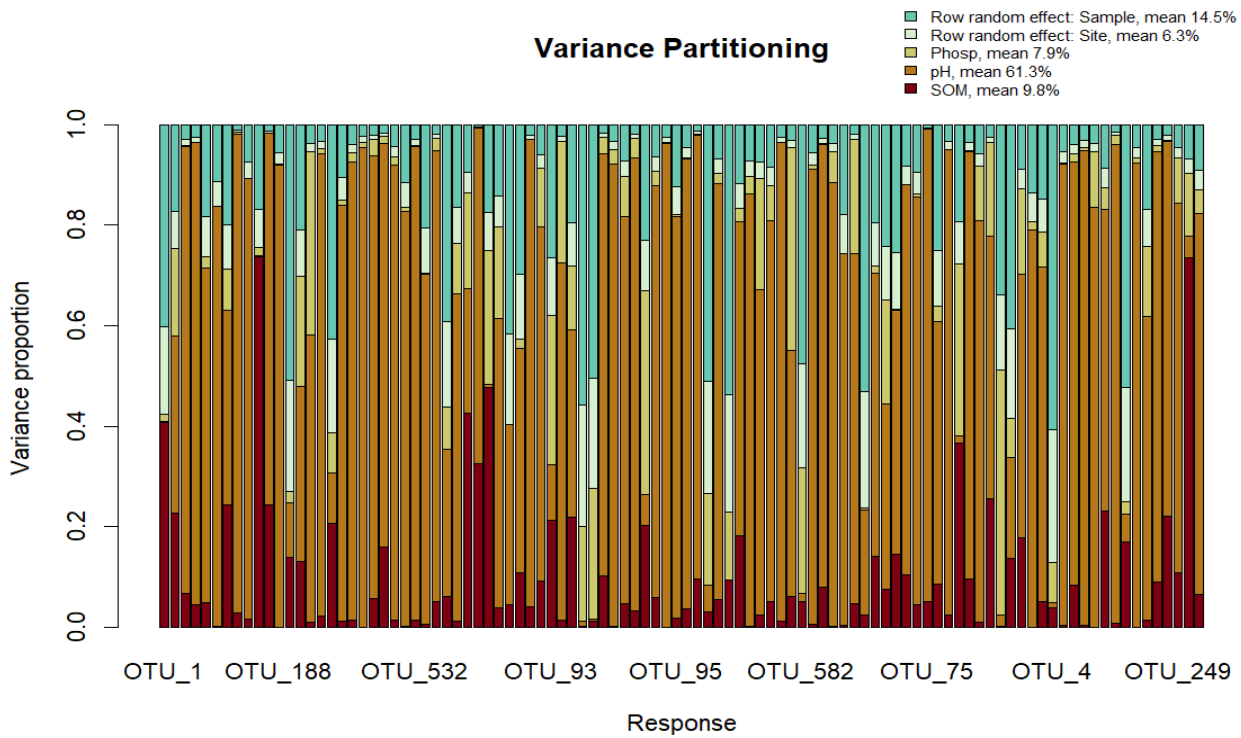
```
data("microbialdata")
y = microbialdata$Y[, colSums(microbialdata$Y>0)>20]
X <- scale(microbialdata$Xenv[, 1:3])
studyDesign <- data.frame(Site = microbialdata$Xenv$Site,
Sample = factor(rownames(microbialdata$Xenv)),  Region =
microbialdata$Xenv$Region)
```

```
modelX <- gllvm(y, X, formula = ~SOM + pH + Phosp,
row.eff = ~(1|Site)+(1|Sample), studyDesign = studyDesign,
                family = "negative.binomial", num.lv = 0)
```

Example 2: Variance partitioning

The proportional variance explained by the community level effects for sampling design and species specific effects of the continuous covariates is calculated using function `VP()`

```
plot(VP(modelX))
```



Example 2: Variation between Sites and Samples

Standard deviations of the random effects for Site and Sample:

```
modelX$params$sigma
```

```
##    Site|Site Sample|Sample
##    0.1791506     0.2472085
```

```
confint(modelX, parm = "sigma")
```

```
##                            cilow      ciup
## sigma.Site|Site     0.02096308 0.3373382
## sigma.Sample|Sample 0.19060109 0.3038158
```

Example 2: Correlation between samples

Correlation of same species between samples within sites on a linear predictor scale:

```
modelX$params$sigma[1]^2/sum(modelX$params$sigma^2)
```

```
## Site|Site
## 0.3443407
```

Connection to ecological diversity measures

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

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

And formulate our model accordingly.

Alpha diversity

$$y_{ij} = \beta_{0j} + x_i \beta_{1j} \tag{11}$$

$$\begin{aligned} var(\beta_{0j} + x_i \beta_{1j}) &= \mathbb{E}(\beta_{0j}^2) + x_i^2 \mathbb{E}(\beta_{1j}^2) \\ &= \sigma_0^2 + x_i^2 \sigma_1^2 \end{aligned} \tag{12}$$

Gamma diversity: across sites

$$y_{ij} = \beta_{0j} + x_i \beta_{1j} \tag{13}$$

$$\begin{aligned} var_{ij}(\beta_{0j} + x_i\beta_{1j}) &= \mathbb{E}_j\{var_i(\eta_{ij})\} + var_j\{\mathbb{E}_i(\eta_{ij})\} \\ &= \sigma_0^2 + \sigma_1^2\{\bar{x} + var(x_i)\} \end{aligned} \tag{14}$$

Beta diversity: between sites

$$y_{ij} = \beta_{0j} + x_i \beta_{1j} \tag{15}$$

$$\begin{aligned} cov(\beta_{0j} + x_i \beta_{1j}, \beta_{0j} + x_k \beta_{1j}) \\ = \sigma_0^2 + x_k x_i \sigma_1^2 \end{aligned} \tag{16}$$

So, the change in beta diversity is:

$$\begin{aligned} cov(\beta_{0j} + x_i \beta_{1j}, \beta_{0j} + x_k \beta_{1j}) - cov(\beta_{0j} + x_i \beta_{1j}, \beta_{0j} + x_l \beta_{1j}) \\ = x_i (x_k - x_l) \sigma_1^2 \end{aligned} \tag{17}$$

End

Thank you!