R package gllvm

Jenni Niku, University of Jyväskylä [jenni.m.e.niku@jyu.fi]

12/16/2020

R package gllvm

- **R** package **gllvm** fits Generalized linear latent variable models (GLLVM) for multivariate data (Niku et al., 2017).
- Developed by J. Niku, W.Brooks, R. Herliansyah, F.K.C. Hui, S. Taskinen, D.I. Warton, B. van der Veen.
- GitHub: https://github.com/JenniNiku/gllvm
- CRAN: https://cran.r-project.org/web/packages/gllvm/index.html

R package gllvm

- GLLVMs are computationally intensive to fit.
- Often such models have been fitted using MCMC approach, which is quite time consuming.
- **gllvm** package overcomes computational problems by applying closed form approximations to log-likelihood and using automatic differentiation in C++ to accelerate computation times (**TMB**).
- Estimation is performed using either variational approximation (VA), extended variational approximation (EVA) or Laplace approximation (LA) method implemented via **R** package **TMB**.
- VA method is faster and more accurate than LA, but not applicable for all distributions and link functions. In such cases EVA can be used.

R package gllvm

Using **gllvm** we can fit

- GLLVM without latent variables fits basic multivariate GLMs
- GLLVM without covariates gives model-based ordination and biplots, also constrained ordination is possible.
- GLLVM with environmental covariates for studying factors explaining species abundance
- Fourth corner models with latent variables for studying environmental-trait interactions

Additional tools: model checking, model selection, inference, visualization.

Distributions

Response	Distribution	Method	Link
Counts	Poisson	VA/LA	log
	NB	VA/LA	log
	ZIP	LA	log
Binary	Bernoulli	VA/LA	probit
		EVA/LA	logit
Ordinal	Ordinal	VA	probit
Normal	Gaussian	VA/LA	identity
Positive continuous	Gamma	VA/LA	log
non-negative continuous	Exponential	VA/LA	log
Biomass	Tweedie	EVA/LA	log
Cover	Beta	EVA/LA	probit

Data input

Main function of the **gllvm** package is `gllvm()`, which can be used to fit GLLVMs for multivariate data with the most important arguments listed in the following:

```
gllvm(y = NULL, X = NULL, TR = NULL, family, num.lv = NULL,
 formula = NULL, method = "VA", row.eff = FALSE, n.init=1, ...)
```

- y: matrix of abundances
- X: matrix or data.frame of environmental variables
- TR: matrix or data.frame of trait variables
- family: distribution for responses
- num.lv: number of latent variables
- method: approximation used “VA”, “EVA” or “LA”
- row.eff: type of row effects
- n.init: number of random starting points for latent variables

Example: Spiders

- Abundances of 12 hunting spider species measured as a count at 28 sites.
- Six environmental variables measured at each site.
    - `soil.dry`: Soil dry mass
    - `bare.sand`: cover of bare sand
    - `fallen.leaves`: cover of fallen leaves/twigs
    - `moss`: cover of moss
    - `herb.layer`: cover of herb layer
    - `reflection`: reflection of the soil surface with a cloudless sky

Data fitting

Fit GLLVM without covariates $g(E(y_{ij})) = \beta_{0j} + \boldsymbol{u}_i'\boldsymbol{\theta}_j$ with **gllvm**, as a default 2 latent variables are used:

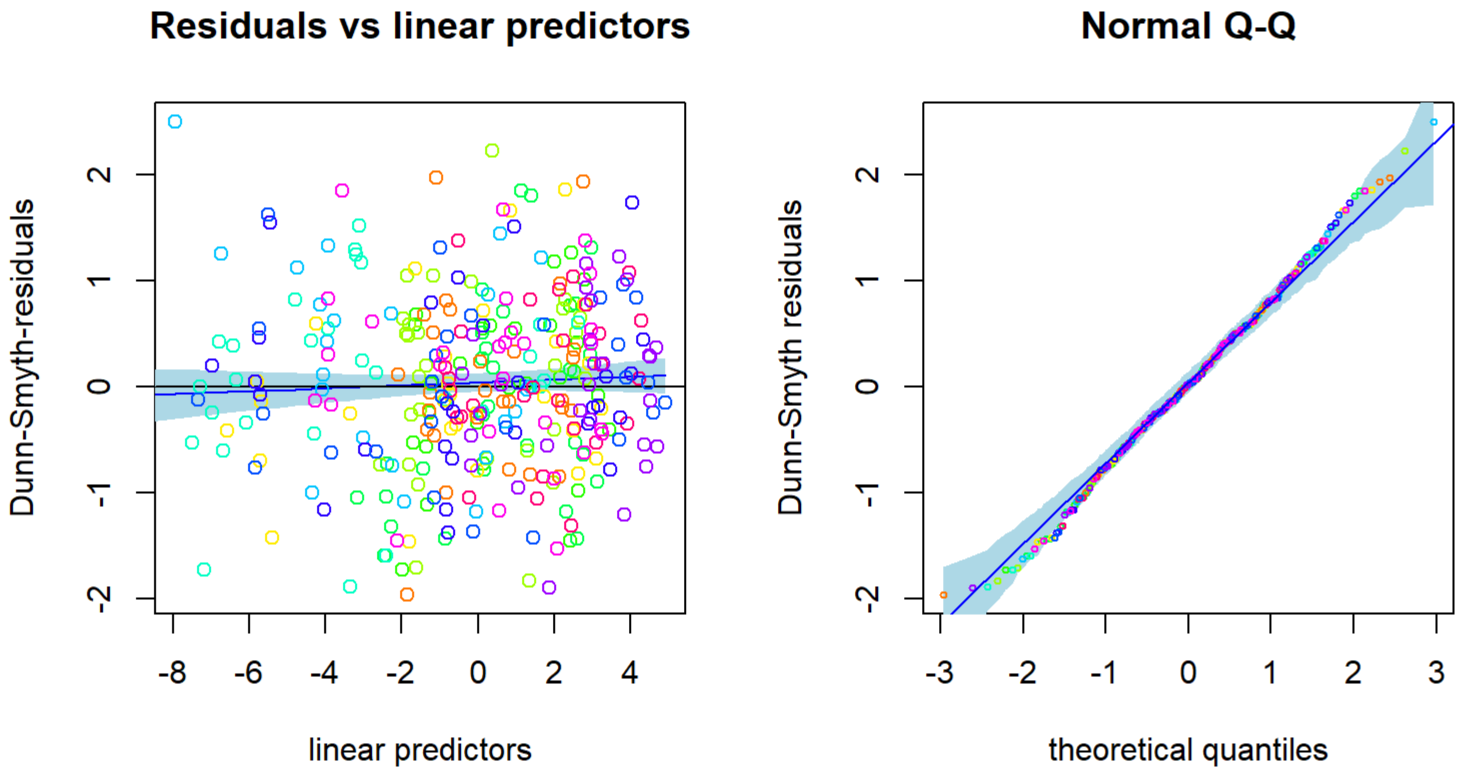
```
library(mvabund)
data("spider")
library(gllvm)
fitnb <- gllvm(y = spider$abund, family = "negative.binomial")
fitnb
## Call:
## gllvm(y = spider$abund, family = "negative.binomial")
## family:
## [1] "negative.binomial"
## method:
## [1] "VA"
##
## log-likelihood:  -733.6806
## Residual degrees of freedom:  289
## AIC:  1561.361
## AICc:  1577.028
## BIC:  1623.975
```

Residual analysis

- Residual analysis can be used to assess the appropriateness of the fitted model (eg. in terms of mean-variance relationship).
- Function `residuals()` calculates randomized quantile residuals for the model, and `plot()` function provides residual plots.
- Randomized quantile/Dunn-Smyth residuals are used in the package, as they provide standard normal distributed residuals, even for discrete responses, in the case of a proper model.

Residual analysis

```
par(mfrow = c(1,2))
plot(fitnb, which = 1:2)
```



Model selection

- Information criterias can be used for model selection.
- For example, compare distributions or choose suitable number of latent variables.

```r
fitp <- gllvm(y = spider$abund, family = poisson())
fitnb <- gllvm(y = spider$abund, family = "negative.binomial")
AIC(fitp)
## [1] 1761.655
AIC(fitnb)
## [1] 1561.361
```

Studying species associations

- Latent variables induce correlation across response variables, and so provide means of estimating correlation patterns across species, and the extent to which they can be explained by environmental variables.
- Information on correlation is stored in the LV loadings $\boldsymbol{\theta}_j$, so the residual covariance matrix, storing information on species co-occurrence that is not explained by environmental variables, can be calculated as $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top$, where $\boldsymbol{\Gamma} = [\boldsymbol{\theta}_1 \ldots \boldsymbol{\theta}_m]'$.
- `getResidualCor` function can be used to estimate the correlation matrix of the linear predictor across species.

Studying species associations

- Let’s consider first the correlation matrix based on a model without predictors: $g(E(y_{ij})) = \beta_{0j} + \boldsymbol{u}_i'\boldsymbol{\theta}_j$

```
fitnb <- gllvm(spider$abund, family = "negative.binomial")
```

- The obtained correlation matrix then does not take into account the environmental conditions driving species abundances at sites, and reflects only what has been observed.

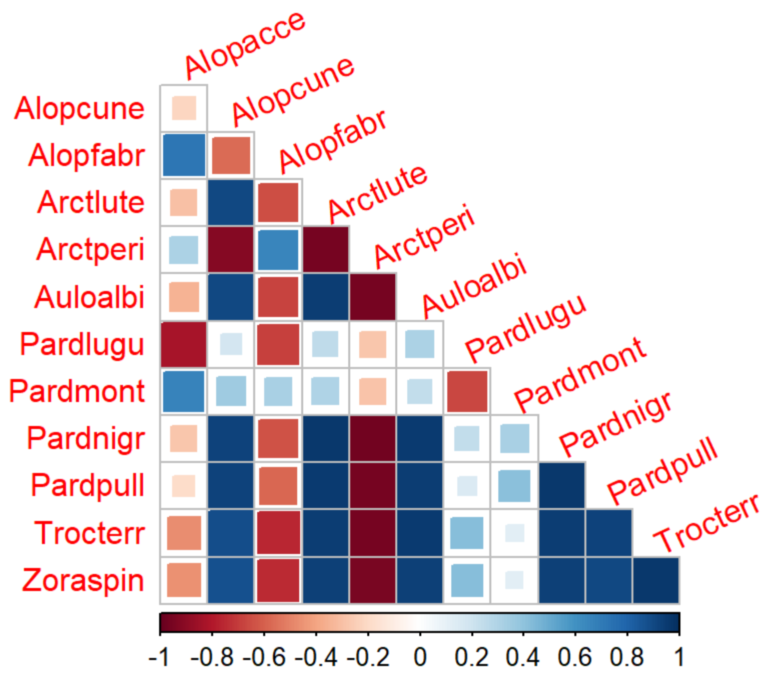
Correlation matrix

The residual correlation matrix can be visualized using, eg., a `corrplot()` function from `corrplot` package:

```
cr <- getResidualCor(fitnb)
library(corrplot);
```

```
corrplot(cr, diag = FALSE, type = "lower", method = "square", tl.srt = 25)
```

Correlation matrix



GLLVM as a model based ordination method

- GLLVMs can be used as a model-based approach to unconstrained ordination by including two latent variables in the model: $g(E(y_{ij})) = \beta_{0j} + \boldsymbol{u}_i'\boldsymbol{\theta}_j$

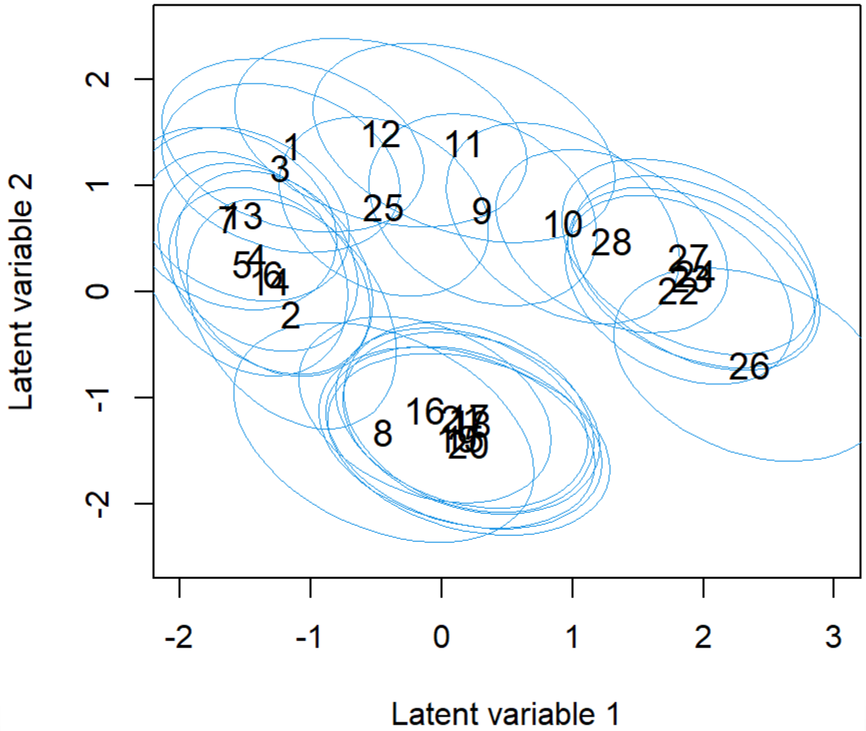
- The latent variable term try to capture the underlying factors driving species abundances at sites.

- Predictions for the two latent variables, $\hat{\boldsymbol{u}}_i = (\hat{u}_{i1}, \hat{u}_{i2})$, then provide coordinates for sites in the ordination plot and then provides a graphical representation of which sites are similar in terms of their species composition.

Ordination plot

- `ordiplot()` produces ordination plots based on fitted GLLVMs.
- Uncertainty of the ordination points in model based ordination can be assessed with prediction errors of latent variables.

```
ordiplot(fitnb, predict.region = TRUE, ylim=c(-2.5,2.5), xlim=c(-2,3))
```

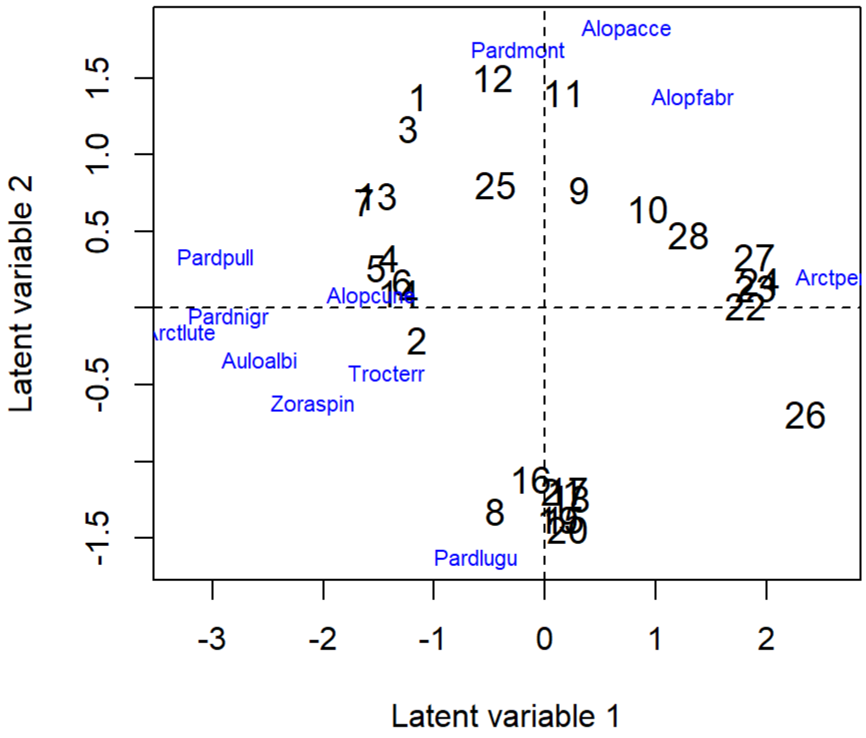


Biplot

- The ordination can also be used for visualizing the species associations by producing a biplot, (argument `biplot = TRUE` in `ordiplot()`), that is, by adding latent variable loadings $\boldsymbol{\theta}_j$ to the ordination of sites.
- In a biplot latent variables and their loadings are rotated so that the LV loadings of the species are in the same direction with the sites where they are most abundant.
- The biplots can be used for finding groups of correlated species or finding indicator species common at specific sites.

Biplot

```
ordiplot(fitnb, biplot = TRUE)
abline(h = 0, v = 0, lty=2)
```

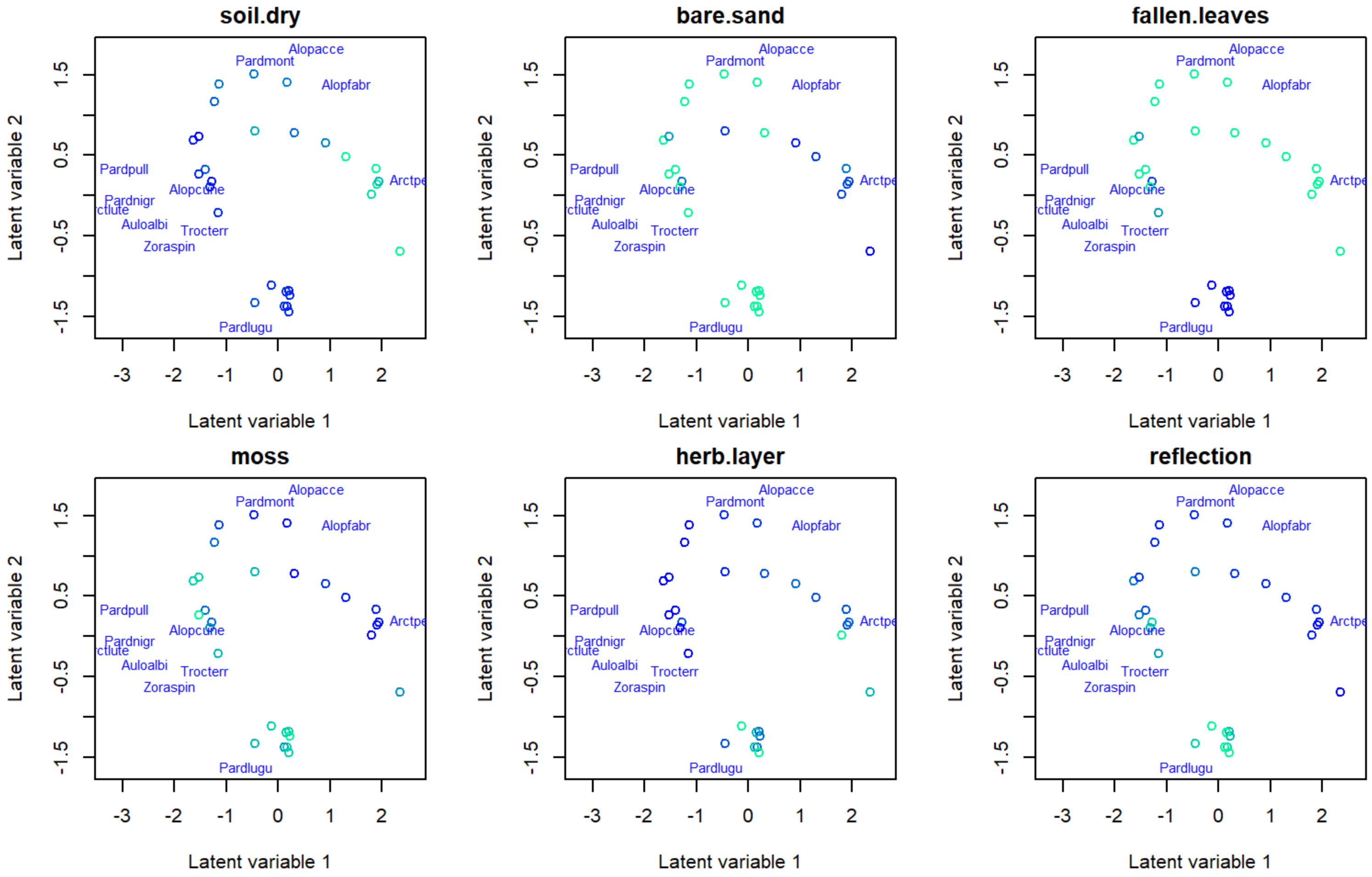


Environmental gradients

The potential impact of environmental variables on species communities can be viewed by coloring ordination points according to the variables.

```
# Arbitrary color palette, a vector length of 20
rbPal <- c("#00FA9A", "#00EC9F", "#00DFA4", "#00D2A9", "#00C5AF", ...)
X <- spider$x
par(mfrow = c(2,3), mar=c(4,4,2,2))
for(i in 1:ncol(X)){
Col <- rbPal[as.numeric(cut(X[,i], breaks = 20))]
ordiplot(fitnb, symbols = T, s.colors = Col, main = colnames(X)[i],
         biplot = TRUE)
}
```

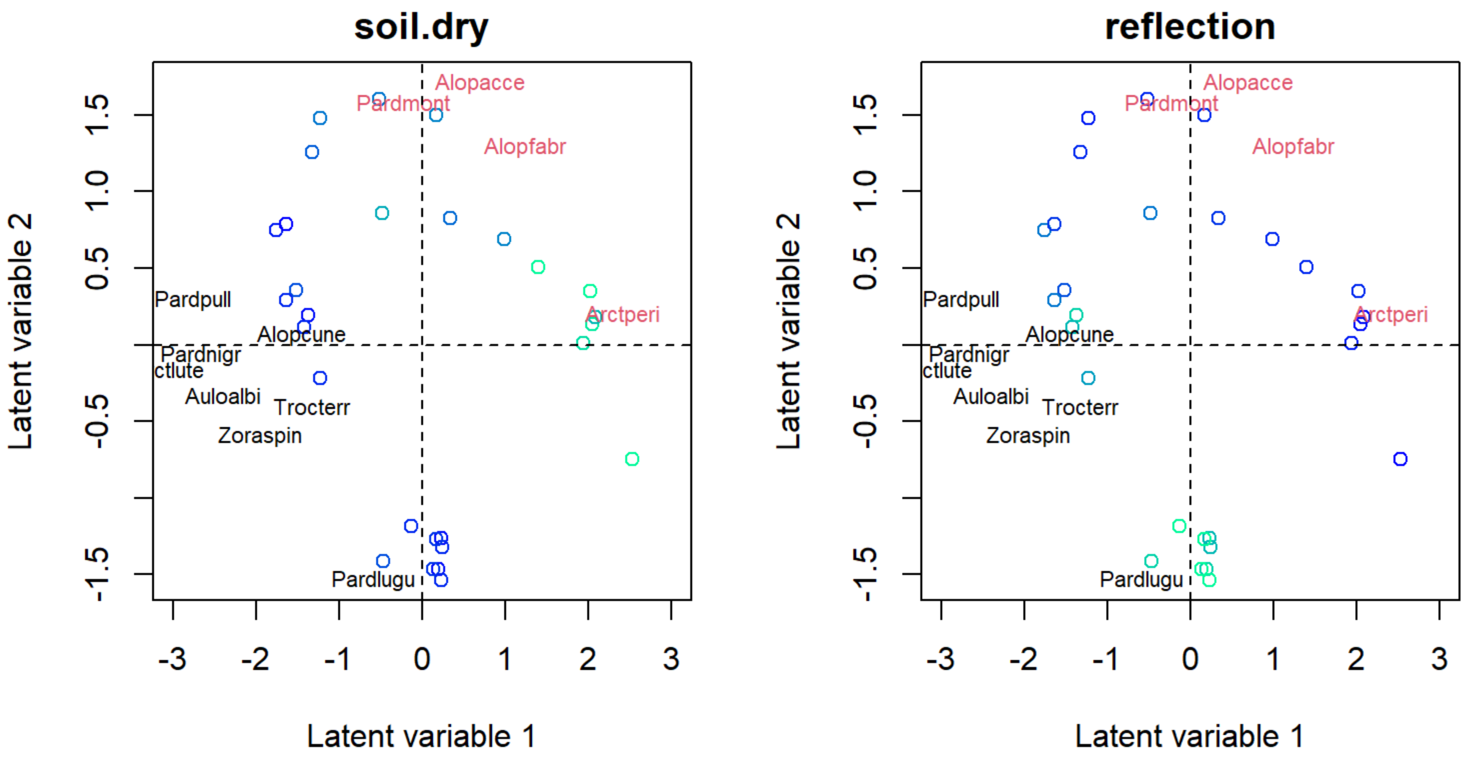
Environmental gradients



Studying effects of environmental variables

- The effects of environmental variables on species can be studied by including environmental variables $\boldsymbol{x}_i$ to GLLVM: $g(E(y_{ij})) = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{u}_i'\boldsymbol{\theta}_j\,.$
- $\boldsymbol{\beta}_j$ is a vector of species specific coefficients for environmental variables.
- Next consider for example two environmental variables, `soil.dry` (soil dry mass) and `reflection` (reflection of the soil surface with a cloudless sky), which shows different environmental gradients in ordination:

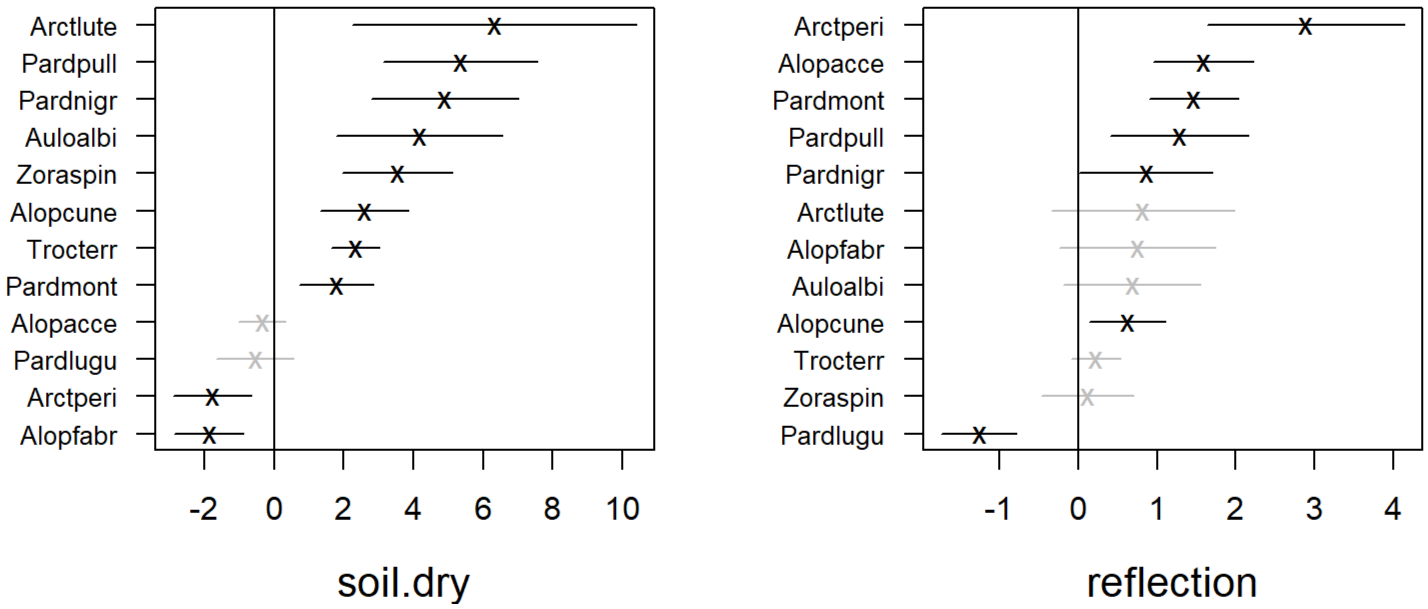
Environmental gradients



Coefficient plot

`coefplot()` plots point estimates of the species specific environmental coefficients $\boldsymbol{\beta}_j$ with confidence intervals.

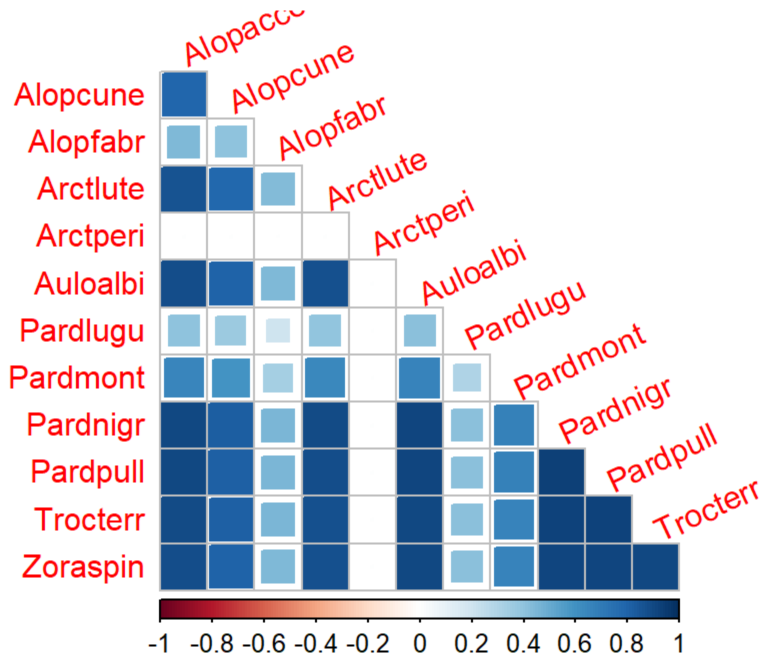
```
fitx1 <- gllvm(spider$abund, X, formula = ~soil.dry + reflection, family = "negative.binomi
coefplot(fitx1, mfrow = c(1,2), cex.ylab = 0.8)
```



Correlation matrix

Correlation matrix for model with predictors shows correlation patterns between species when the effect of the predictors are taken into account.

```
crx <- getResidualCor(fitx1)
corrplot(crx, diag = FALSE, type = "lower", method = "square", tl.srt = 25)
```



Fourth corner models

- If species trait variables $\boldsymbol{t}_j$, measuring eg. species behaviour or physical appearance, would be available, fourth corner models should be considered: $g(E(y_{ij})) = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{x}_i'\boldsymbol{B}_I\boldsymbol{t}_j + \boldsymbol{u}_i'\boldsymbol{\theta}_j$
- Such models can also be fitted with `gllvm()` function by including a matrix of traits with argument `TR`.
- Examples can be found in the **gllvm** package's vignettes.

More

More information and examples about the usage of the package can be found from the **gllvm** package's website: https://jenniniku.github.io/gllvm/

Github repository of the workshop:
https://github.com/BertvanderVeen/IRSAE2020GLLVMworkshop