gllvm 2.0: Fast fitting of advanced ordination methods and joint species distribution models

Pekka Korhonen

Brief intro to `gllvm`

What?

- R package for joint species distribution modeling, employing generalized linear latent variable models (GLLVM)
- Emphasis on model-based ordination and co-occurence
- Fast estimation based on approximate marginal likelihood

When?

- Initial version accompanied Niku et al. (2017)
- First article dedicated on the software in Niku et al. (2019)
- Lots of functionalities added since then

Installing the `gllvm` package

The CRAN version of the package can be installed using:

```r
install.packages("gllvm")
```

For the more up to date, developmental version, instead use:

```r
devtools::install_github("JenniNiku/gllvm")
```

Abundance/community data

Example datasets included in `gllvm`:

```r
data("beetle")
head(beetle$Y, c(5,8))
```

```
     agonfuli agonmuel amaraene amarapri amarauli amarbifo amarcomm amareury
gs21        0       51       25        0        0        0        1        0
gs22        0       59        2        0        0        0        2        0
eg11        0        0        0        0        0        0        3        0
eg12        0        1        1        0        0        0        1        0
eg13        0       29        1        0        0        0        2        0
```

```r
data("kelpforest")
head(kelpforest$Y, c(5,15))
```

```
  AB AL AMZO   ANSP AR ARUD     AS ATM     AU BA BAEL BCAL     BF BLD     BN
1  0  0    0 0.0125  0    0 0.1125   0 0.0125  0    0    0 0.0000   0 0.0125
2  0  0    0 0.0000  0    0 0.0000   0 0.0000  0    0    0 0.0125   0 0.0000
3  0  0    0 0.0000  0    0 0.0750   0 0.0000  0    0    0 0.0000   0 0.0000
4  0  0    0 0.0000  0    0 0.0125   0 0.0000  0    0    0 0.0125   0 0.0000
5  0  0    0 0.0000  0    0 0.0000   0 0.0000  0    0    0 0.0000   0 0.0000
```

“Standard” GLLVM[1]

Let y_{ij} denote the record for response (species, OTUs, etc.) $j = 1, \ldots, m$, recorded at sample $i = 1, \ldots, n$, e.g., study sites. Additionally, we may have records for environmental variables $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})^\top$ for each sample.

In a GLLVM, the mean $\mathbb{E}[y_{ij}] = \mu_{ij}$ is regressed against $\boldsymbol{x}_i$ and a set of d **latent variables** $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{id})^\top$ via:

$$g(\mu_{ij}) = \eta_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{u}_i^\top \boldsymbol{\gamma}_j,$$

for each pair i, j.

$$g(\mu_{ij}) = \eta_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{u}_i^\top \boldsymbol{\gamma}_j$$

- $g(\cdot)$ is a link function, e.g., logit, probit
- α_i is a row/sample effect (fixed or random)
- $\beta_{0j}, \boldsymbol{\beta}_j$ are species-specific regression coefficients
- $\boldsymbol{u}_i \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\mathcal{I}})$ i.i.d.
- $\boldsymbol{\gamma}_j = (\gamma_{j1}, \dots, \gamma_{jd})^\top$ are LV **loadings**

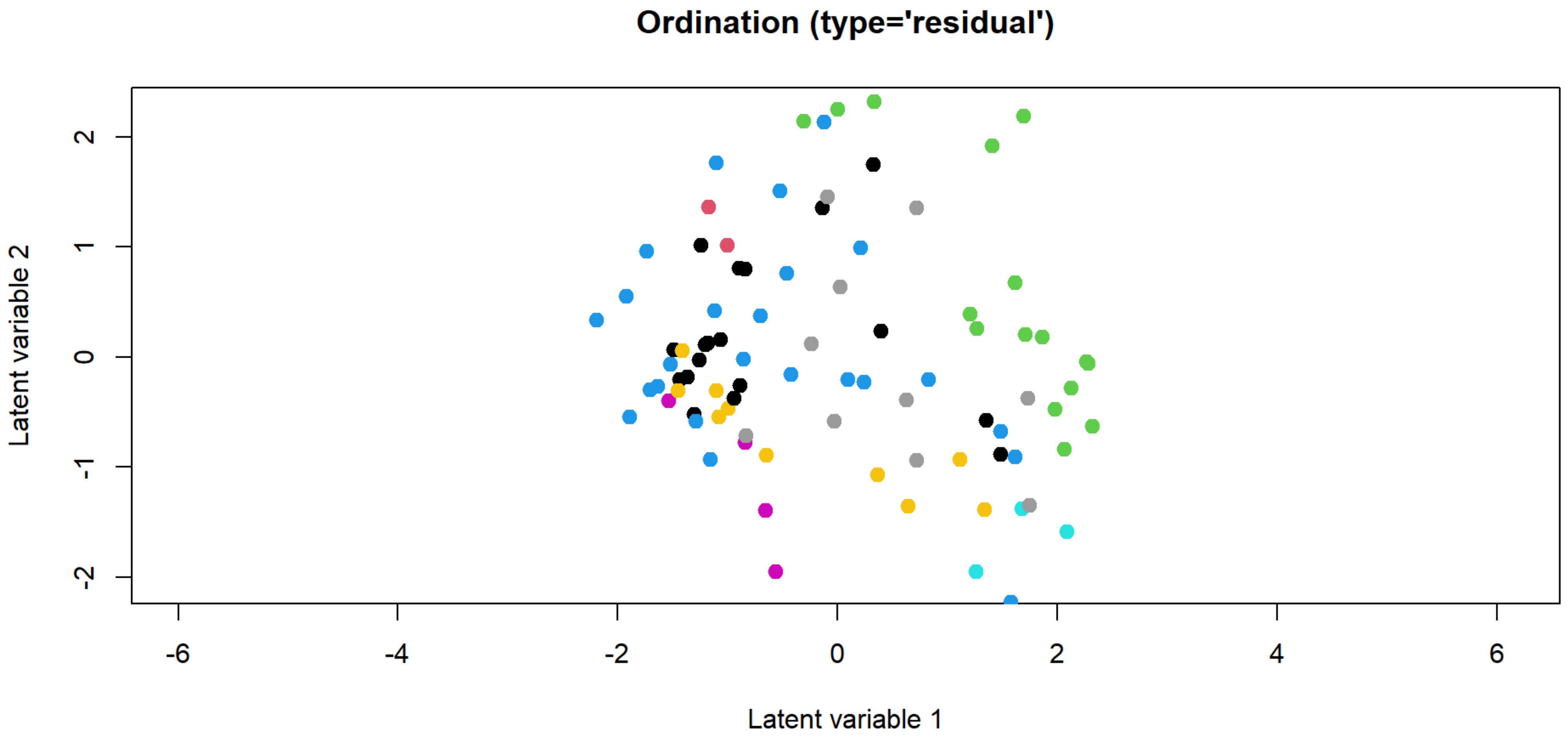
Furthermore, in a fouth-corner GLLVM:

$$\boldsymbol{\beta}_j = \boldsymbol{\beta}_e + \boldsymbol{B}_{et}\boldsymbol{t}_j + \boldsymbol{b}_j$$

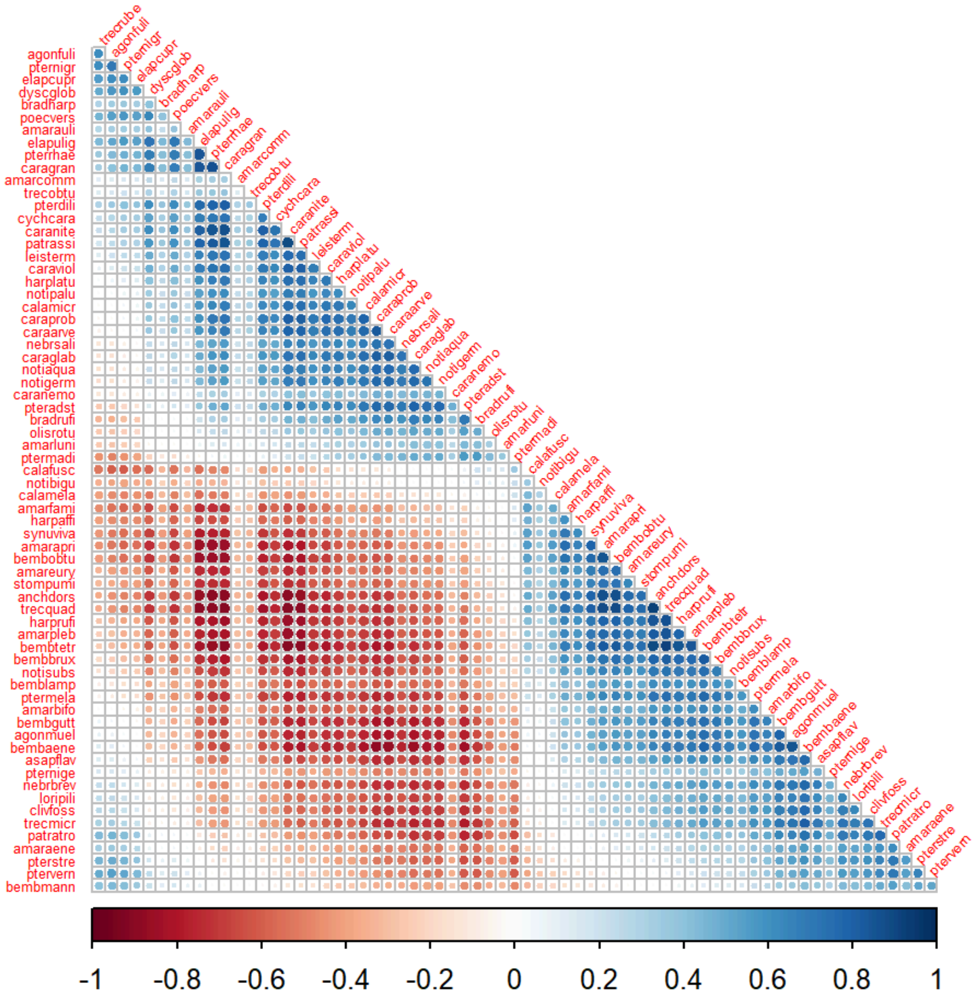
Ordination and interspecies correlations

- With the choice $d = 2$, GLLVMs can be used for unconstrained (or residual) ordination of observations/samples, by plotting the predicted LV scores $\boldsymbol{u}_i = (u_{i1}, u_{i2})^\top$. Additionally, biplots displaying also the effects of species can be constructed easily.
- Co-occurance patterns between species can be inspected via the residual correlation matrix $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$, where $\boldsymbol{\Lambda} = [\boldsymbol{\gamma}_1 \ldots \boldsymbol{\gamma}_m]^\top$ holds the species' loadings.

```r
m0 = gllvm(beetle$Y, num.lv=2, family="negative.binomial")
ordiplot(m0, symbols = TRUE, s.colors=beetle$X$Area, pch=19, jitter=TRUE)
```



```r
cr = getResidualCor(m0); library(corrplot)
corrplot(cr, diag=FALSE, type="lower", tl.cex=0.4, order="AOE", tl.srt=45)
```



gllvm 2.0: original featureset (left) vs. newer additions (right)

Model type	Linear, i.i.d. LVs;
	Fourth-corner GLLVM
		Correlated LVs;
		Quadratic LVs;
		Informed LVs;
		(Phylogenetic) REs;
		Reduced-rank regression

Response type	Presence-absence;
	(Overdispersed) counts;
	Ordinal;
	(Non-negative) continuous
		Zero-inflated counts;
		Positive continuous;
		Percent cover (with 0% and 100% records)

Community-level row effects	Single fixed/random	Multiple fixed/random; Correlated/structured effects
Ordination analysis	Unconstrained; Residual	Constrained; Concurrent
Species association	Residual correlation	Environmental correlation

Inference	Analysis of deviance;
	CIs for parameters
		Fixed-effects covariances;
		Prediction intervals;
		Missing data (MAR)

Visualization	Ordination (bi-)plots;
	Estimated fixed effects
		Uncertainty in ordination;
		Predicted random effects;
		Variance partitioning plot

Model fitting methods	Laplace approximation; Variational approximations	Extended VA; Parallel computation

Examples

SBC LTER kelp forest data [1]

- Comprises of percent cover records of $m = 130$ species of macroalgae and sessile invertebrates
- Collected between 2000–2020 along 44 permanent transects across 11 observational sites
    - Some of the sites were located on islands, others on coast
- Around 88% of the covers were recorded as zeros
- Seabed rockiness and giant kelp frond density were also considered as environmental covariates

Structured row effects and LVs

To accommodate the hierarchical study design in the SBC LTER dataset, the `row.eff` argument now accepts formulas, e.g.:

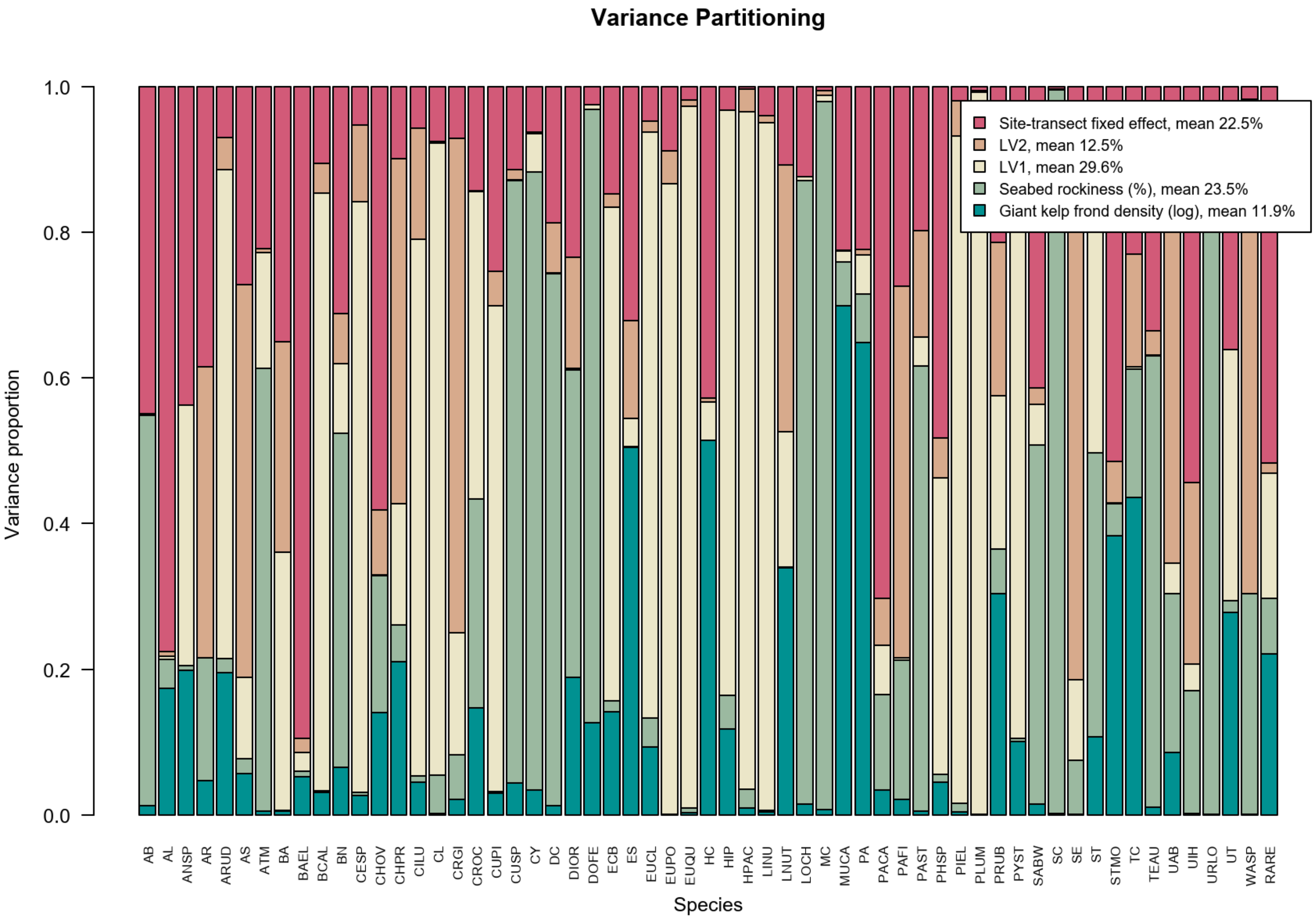
```r
gllvm(y=Y, X=X, family="orderedBeta", row.eff=~(1|SITE/TRANSECT) + YEAR,
      studyDesign=Z[,c("SITE","TRANSECT","YEAR")])
```

which would fit a model with a fixed effect corresponding to sampling year, and a random effect for each transect nested within the observation sites. Similarly, for LVs, e.g.:

```r
gllvm(y=Y, X=X, family="orderedBeta", num.lv=2, row.eff=~(1|SITE/TRANSECT),
      lvCor=~(1|YEAR), studyDesign=Z[,c("SITE","TRANSECT","YEAR")])
# or for autoregressive correlation:
gllvm(y=Y, X=X, family="orderedBeta", num.lv=2, row.eff=~(1|SITE/TRANSECT),
      lvCor=~corAR1(1|YEAR), studyDesign=Z[,c("SITE","TRANSECT","YEAR")])
```

with other options including `corExp`, `corMatern`, and `corCS`.

The functions `varPartitioning` and `plotVP` are useful for models with (nested) row effects, LVs, and covariate effects:



Phylogenetic random effect model

If the data include functional traits for the species, a fourth-corner GLLVM

$$\boldsymbol{\beta}_j = \boldsymbol{\beta}_e + \boldsymbol{B}_{et}\boldsymbol{t}_j + \boldsymbol{b}_j$$

can be fitted with

```r
gllvm(y=Y, X=X, TR=Traits, family="orderedBeta", num.lv=2,
      formula=~KELP_FRONDS + PERCENT_ROCKY
        + (KELP_FRONDS + PERCENT_ROCKY) : (GROUP),
        randomX=~KELP_FRONDS + PERCENT_ROCKY, n.init=5)
```

In the standard case, $\boldsymbol{b}_j$ are independent for $j = 1, \ldots, m$. This can be relaxed if phylogenetic information is available.

Assuming a phylogenetic covariance matrix $\boldsymbol{C}$, in a phylogenetic GLLVM[1], for covariate $l = 1, \ldots, k$, we have

$$(b_{1l}, \ldots, b_{ml})^\top \sim \mathcal{N}_m\left(\mathbf{0}, \sigma_l^2[\boldsymbol{C}\rho_l + (1-\rho_l)\boldsymbol{\mathcal{I}}]\right),$$

where $\rho_l \in [0, 1]$ is the *phylogenetic signal parameter*, which can also be shared between the covariates in $\boldsymbol{X}$, i.e., $\rho_l = \rho$ for each $l = 1, \ldots, k$.

As such a model can be very demanding computationally, w.r.t. m, in `gllvm` we have adopted the nearest neighbour Gaussian process appoximation of Datta et al. (2016).

Model fitting and visualization

Phylogenetic random effect model can then be fitted with:

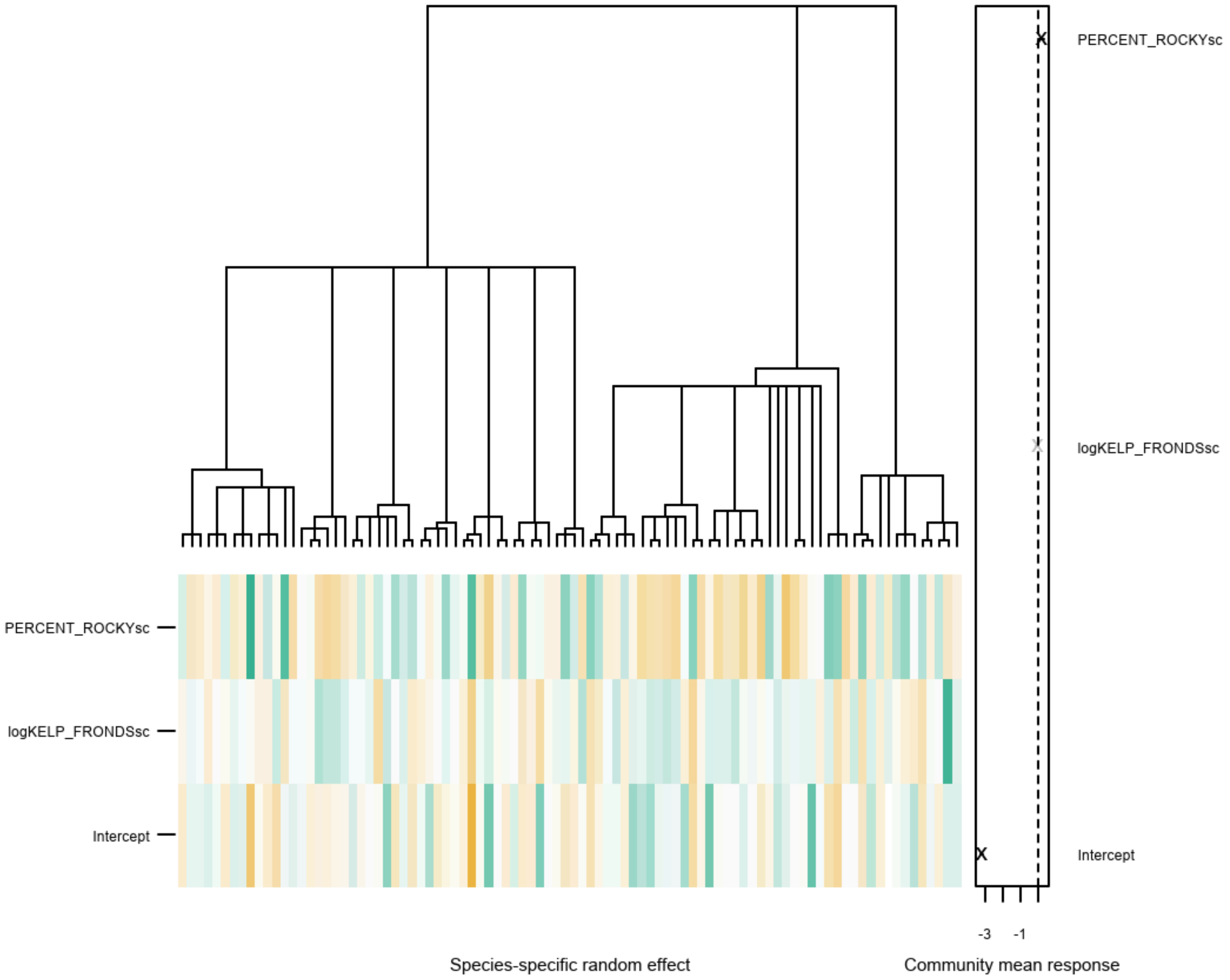
```
ftPhylo = gllvm(y=Y[,order], X=X, TR=Traits[order,],
            formula=˜(KELP_FRONDSsc + PERCENT_ROCKYsc) +
              (KELP_FRONDSsc + PERCENT_ROCKYsc) : (GROUP),
            randomX=~KELP_FRONDSsc + PERCENT_ROCKYsc,
            colMat=list(colMat[order,order], dist=dist[order,order]),
            colMat.rho.struct="term", nn.colMat=10, family="orderedBeta",
            n.init=5, optim.method="L-BFGS-B", num.lv=2)
```

Signal parameter common to all covariates can be specified instead with the argument `colMat.rho.struct="single"`

Results, together with the phylogenetic tree (constructed with the ape package), can be visualized with:

```
phyloplot(ftPhylo, tree)
```

phyloplot


PERCENT_ROCKYsc
logKELP_FRONDSsc
Intercept
-3
-1
Species-specific random effect
Community mean response

Scottish ground beetle dataset[1]

- Counts from $m = 68$ species of beetles, on $n = 87$ sites
- Notably, the data include $k = 17$ primary covariates
  - Among them, e.g.; organic content, soil pH, moisture, canopy height, stem density, biomass, elevation, etc.
  - In a typical GLM, $\eta_{ij} = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j$, this would mean 1224 regression parameters
- Risks of overfitting can be alleviated via reduced-rank regression, or *constrained ordination*

Constrained ordination / RRR

Let $\boldsymbol{B}$ denote a $k \times d$ matrix of *canonical coefficients*. Then, we impose the following structure for the coefficients $\boldsymbol{\beta}_j$:

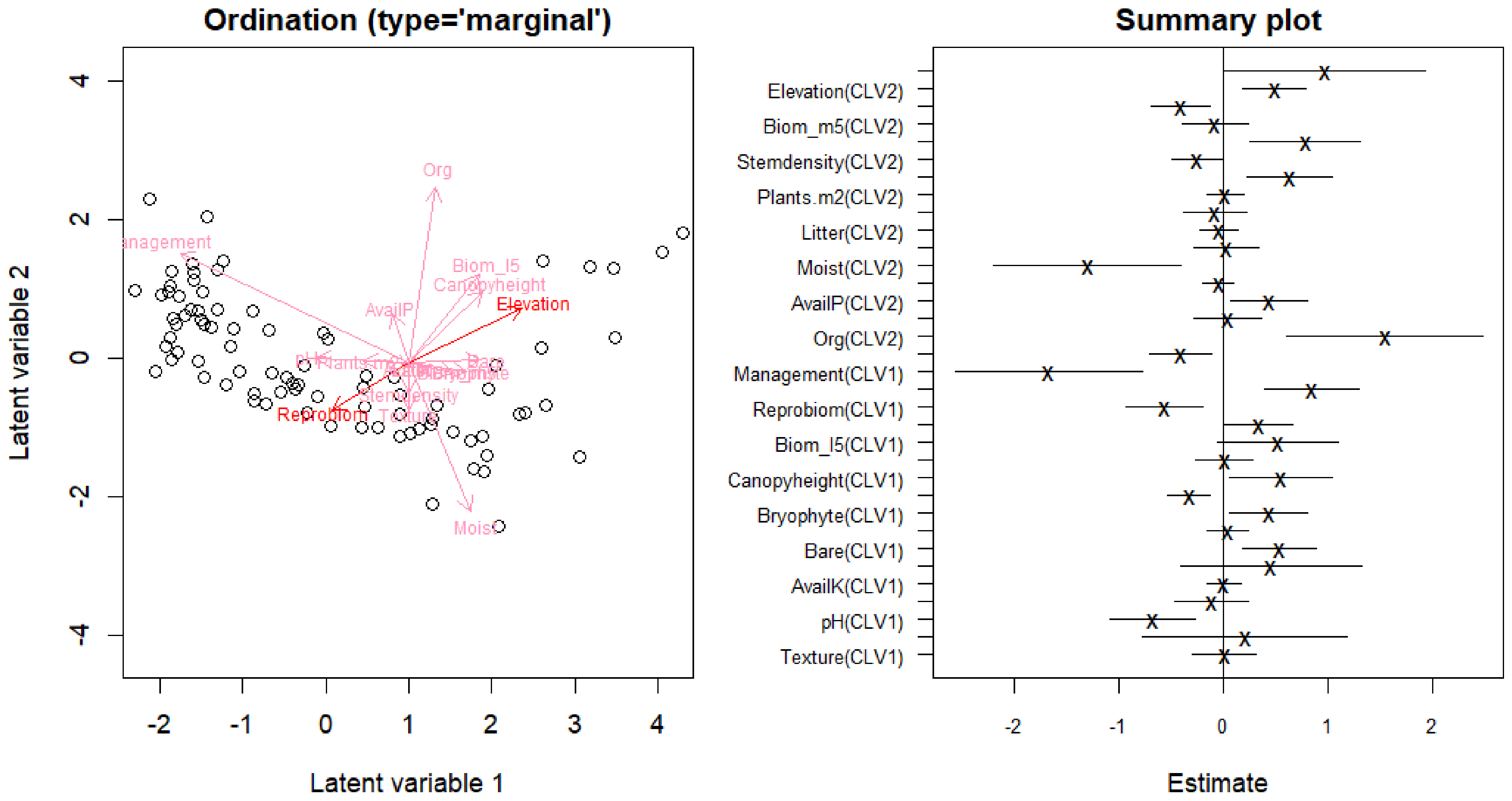
$$\boldsymbol{\beta}_j = \boldsymbol{B}\boldsymbol{\gamma}_j \implies \eta_{ij} = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{B}\boldsymbol{\gamma}_j.$$

Alternatively, the resulting model can be seen as a GLLVM with latent variables in the from of

$$\boldsymbol{u}_i = \boldsymbol{B}^\top \boldsymbol{x}_i.$$

Note, that $\boldsymbol{B}$ can be formulated as either fixed or random.

```r
ftConstOrd = gllvm(y=beetle, X=X, family="negative.binomial", num.RR=2)
# 'randomB' argument could be used here to specify B as random
par(mfrow=c(1,2))
ordiplot(ftConstOrd, symbols=TRUE)
plot(summary(ftConstOrd), cex.axis=.75, main="Summary plot")
```



Concurrent ordination[1]

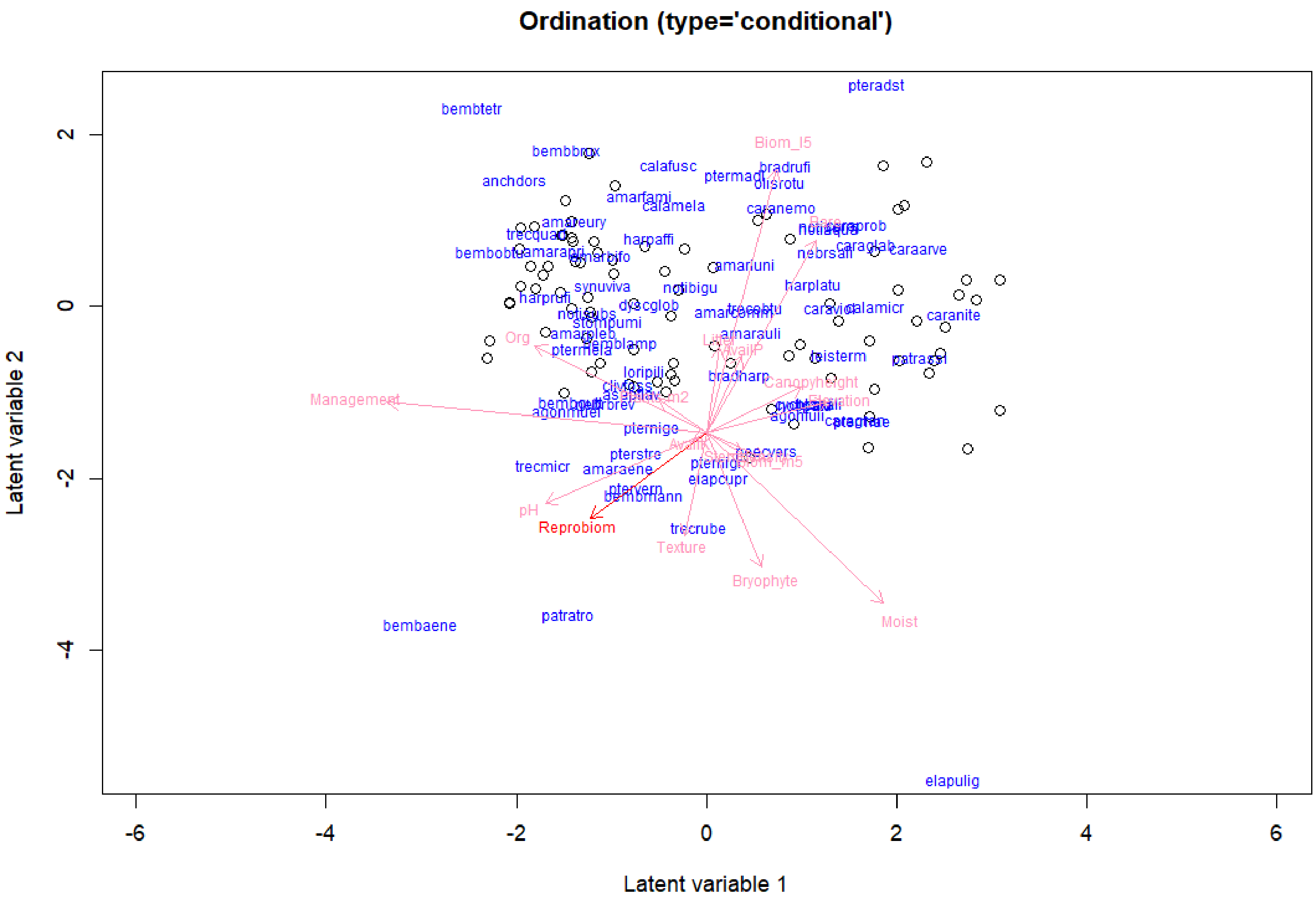
When $\boldsymbol{u}_i = \boldsymbol{B}^\top \boldsymbol{x}_i$, the LVs are governed solely by the covariates observed—often unrealistic in practice. An unique advantage of the GLLVM framework is the capacity to incorporate “residual” LVs:

$$\eta_{ij} = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{B} \boldsymbol{\gamma}_j + \boldsymbol{\epsilon}_i^\top \boldsymbol{\gamma}_j,$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma}^2)$.

This specification allows for simultaneous unconstrained and constrained ordination, hence, *concurrent* ordination.

```r
ftConcOrd = gllvm(y=beetle, X=X, family="ZINB", num.lv.c=2, n.init=5)
ordiplot(ftConcOrd, symbols=TRUE, biplot=TRUE)
```



Future outlook

- Polya-Gamma augmentation for logistic models[1]
- Faster estimation for spatially (and spatio-temporally) correlated LVs utilizing e.g., NNGPs, SPDEs, etc.
- More processes for trait evolution in the phylogenetic model
- Mixed response types, distributions for compositional data
- Further parallelization, e.g., GPU-based computing
- Regularization, stochastic/mini-batch gradient descent

Thank you!

References

Datta, Abhirup, Sudipto Banerjee, Andrew O. Finley, and Alan E. Gelfand. 2016. "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets." *Journal of the American Statistical Association* 111 (514): 800–812. https://doi.org/10.1080/01621459.2015.1044091.

Niku, J., F. K. C. Hui, S. Taskinen, and D. I. Warton. 2019. "Gllvm – Fast Analysis of Multivariate Abundance Data with Generalized Linear Latent Variable Models in ." *Methods in Ecology and Evolution* 10 (12): 2173–82.

Niku, J., D. I. Warton, F. K. C. Hui, and S. Taskinen. 2017. "Generalized Linear Latent Variable Models for Multivariate Count and Biomass Data in Ecology." *Journal of Agricultural, Biological, and Environmental Statistics* 22: 498–522. https://doi.org/10.1007/s13253-017-0304-7.

Polson, Nicholas G., James G. Scott, and Jesse Windle. 2013. "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables." *Journal of the American Statistical Association* 108 (504): 1339–49. https://doi.org/10.1080/01621459.2013.829001.

Reed, Daniel C., and Robert J. Miller. 2023. "SBC LTER: Reef: Kelp Forest Community Dynamics: Cover of Sessile Organisms, Uniform Point Contact. LTER Network

Member Node." https://doi.org/10.6073/pasta/0af1a5b0d9dde5b4e5915c0012ccf99c.

Ribera, Ignacio, Sylvain Doledec, Iain S. Downie, and Garth N. Foster. 2001. "Effect of Land Disturbance and Stress on Species Traits of Ground Beetle Assemblages." *Ecology* 82: 1112–29. https://doi.org/10.1890/0012-9658(2001)082[1112:EOLDAS]2.0.CO;2.

Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton: Chapman & Hall.

van der Veen, Bert, Francis K C Hui, Knut A Hovstad, and Robert B O'Hara. 2023. "Concurrent ordination: Simultaneous unconstrained and constrained latent variable modelling." *Methods in Ecology and Evolution* 14 (2): 683–95.

van der Veen, Bert, and Robert Brian O'Hara. 2024. "Fast Fitting of Phylogenetic Mixed Effects Models." *arXiv*. https://arxiv.org/abs/2408.05333.