Joint species distribution modelling with the R-package Hmsc

Gleb Tikhonov, Øystein Opedal, Nerea Abrego, Aleksi Lehikoinen , Melinda de Jonge, Jari Oksanen & Otso Ovaskainen

Appendix S1. Technical specification of the Hmsc implementation

Contents

1. Overview of HMSC structure, input data and notation

The structure of the core HMSC model is shown in Figure S1 and the notation is specified in Tables S1-S3. The input data are illustrated graphically in Figure S2.

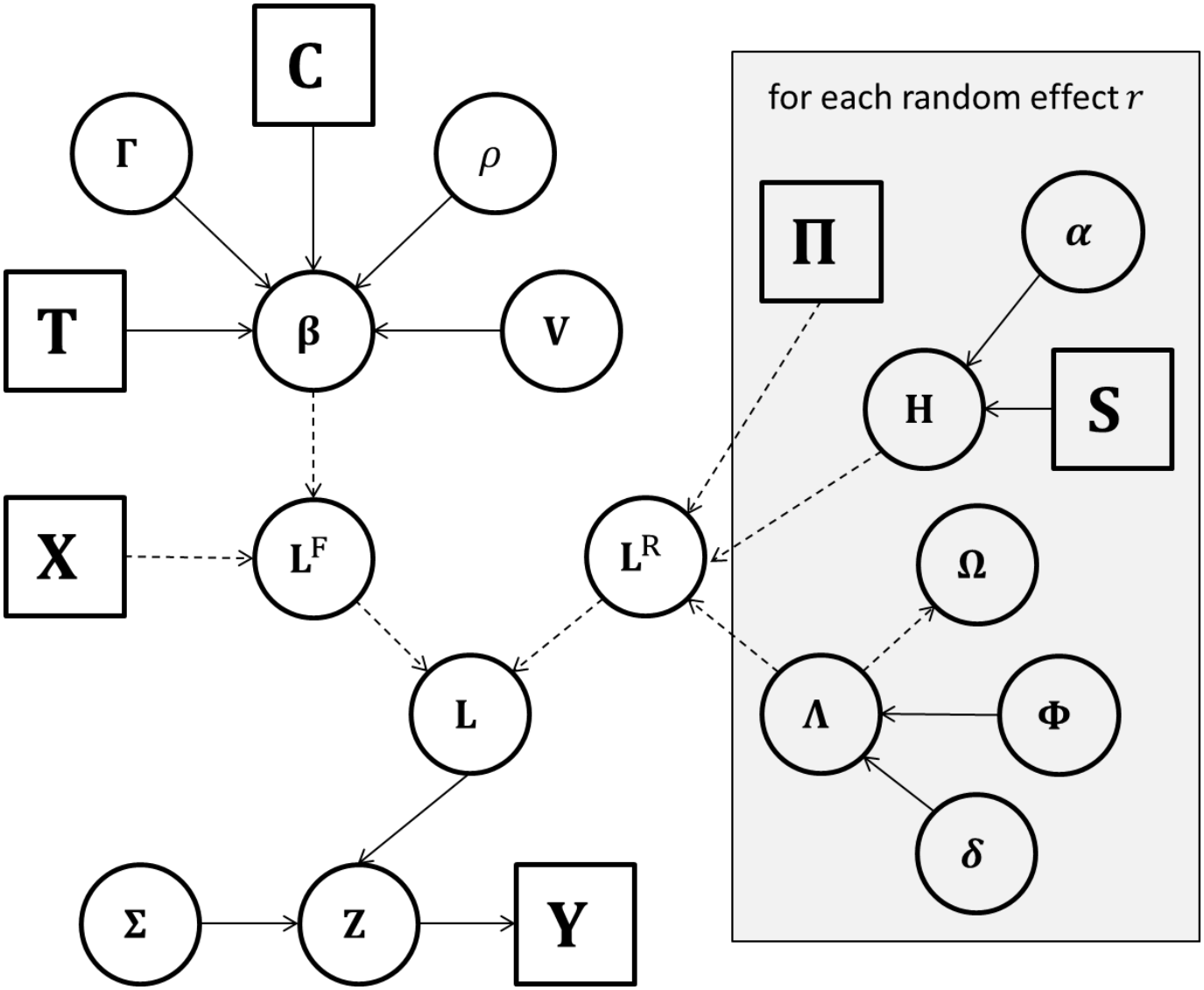

Figure S1. The Directed Acyclic Graph of the core HMSC model. The input data are represented as rectangles and the estimated parameters of the model as ellipses. The continuous arrows depict stochastic links modelled through probabilistic relationships, and the dashed arrows deterministic links. The gray box illustrates a potentially repeated structure, so that there can be multiple random effects in the same model.

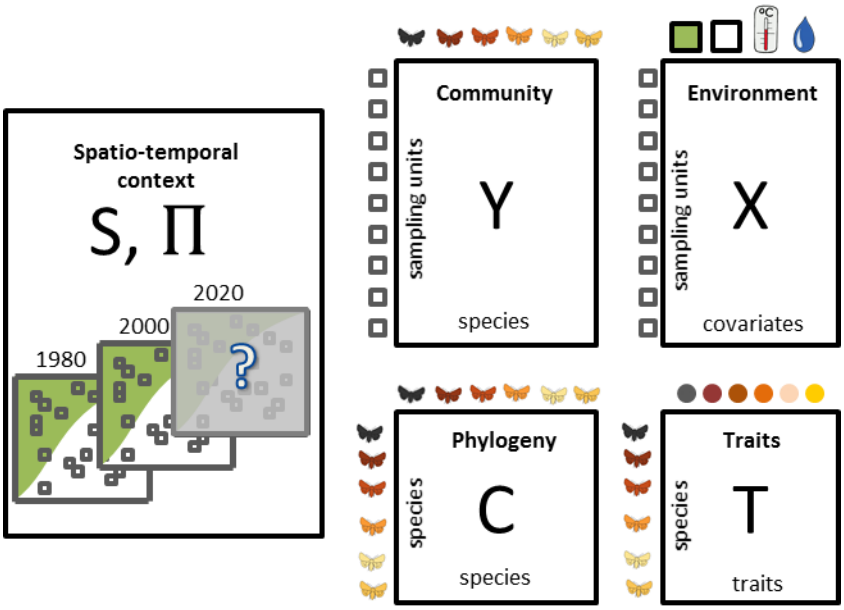

Figure S2. Input data matrices of HMSC. The community data (denoted as the **Y** matrix) include the occurrences or abundances of the species recorded in a set of temporal and/or spatial sampling units. The environmental data (denoted as the **X** matrix) consist of the environmental covariates measured over the sampling units. The trait data (denoted as the **T** matrix) consist of a set of traits measured for the species present in the **Y** matrix. The phylogenetic correlations matrix (denoted as the **C** matrix) quantifies phylogenetic relatedness among all species pairs, and it can be derived e.g. from a phylogenetic tree. The spatiotemporal context includes location and time information about the samples, coded as hierarchical levels that aggregates the sampling units (denoted as the **Π** matrix), and the spatial or temporal coordinates of the units included at each hierarchical level (denoted as the **S** matrix).

Table S1. Indices and their ranges in the core HMSC model.

Index and its range	Refers to
$i = 1, \ldots, n$	sampling unit
$j = 1, \ldots, n_s$	species
$k = 1, \ldots, n_c$	environmental covariate
$l = 1, \ldots, n_t$	species trait
$h = 1, \ldots, n_f$	latent factor
$u = 1, \ldots, n_u$	hierarchical unit
$q = 1, \ldots, d$	coordinate in $\mathbb{R}^d$ (e.g. spatial or temporal)
$r = 1, \ldots, n_r$	random effect

Table S2. Data matrices and their dimensions in the core HMSC model. The spatial coordinates are defined separately for each random effect r.

Data matrix	Dimension	Refers to
Y, element y_{ij}	$n \times n_s$	Community data
X, element x_{ik}	$n \times n_c$	Environmental data

T, element t_{jl}	$n_s \times n_t$	Species trait data
C, element $c_{j_1 J_2}$	$n_s \times n_s$	Phylogenetic relatedness
Π, element π_{iu}	$n \times n_u$	Study design
S, element s_{uq}	$n_u \times d$	Spatial coordinates

Table S3. Parameters and their interpretations in the core HMSC model. The column category indicates whether the parameter is related to the fixed effect (F), random effect (R), or data model (D) part of HMSC. The parameters of the random effect part are defined separately for each random effect r.

Category	Parameter	Type	Interpretation
F	$\mathbf{L}^{\mathrm{F}}$, element L^{F}_{ij}	$n \times n_s$ matrix	Linear predictor of fixed effects
F	**B**, element β_{kj}	$n_c \times n_s$ matrix	Species niches
F	**M**, element μ_{kj}	$n_c \times n_s$ matrix	Expected species niches based on traits
F	ρ	Scalar	Phylogenetic signal in species niches
F	**Γ**, element γ_{kl}	$n_c \times n_t$ matrix	Influence of traits on niches
F	**V**, element $V_{k_1 k_2}$	$n_c \times n_c$ matrix	Residual covariance of species niches
R	$\mathbf{L}^{\mathrm{R}}$, element L^{R}_{ij}	$n \times n_s$ matrix	Linear predictor of random effects
R	**H**, element η_{uh}	$n_u \times n_f$ matrix	Site loadings
R	$\boldsymbol{\alpha}$, element α_h	vector of length n_f	Spatial scale of site loadings
R	**Λ**, element λ_{hj}	$n_f \times n_s$ matrix	Species loadings
R	**Ω**, element $\Omega_{j_1 j_2}$	$n_s \times n_s$ matrix	Species associations
R	**Φ**, element ϕ_{hj}	$n_f \times n_s$ matrix	Local shrinkage of species loadings
R	$\boldsymbol{\delta}$, element δ_l	vector of length n_f	Global shrinkage of species loadings
D	**L**, element L_{ij}	$n \times n_s$ matrix	Linear predictor
D	**Σ**, element σ_j^2	$n_s \times n_s$ diagonal matrix	Residual variance
D	**Z**, element z_{ij}	$n \times n_s$ matrix	Latent variable

1.1. The mathematical structure of HMSC

Here we specify the structure of the HMSC model shown in Figue S1 with the help of mathematical equations. HMSC is a multivariate and hierarchical generalized linear mixed model. The matrix of linear predictors **L** is a sum of fixed and random effects, so that $\mathbf{L} = \mathbf{L}^{\mathrm{F}} + \mathbf{L}^{\mathrm{R}}$.

In the notation of Tables S1-S3, the fixed effect are modelled as $\mathbf{L}^{\mathrm{F}} = \mathbf{XB}$, where the species-specific regression parameters are further modelled as a function of traits and phylogenetic relationships as $\mathrm{vec}(\mathbf{B}) \sim N\left(\mathrm{vec}(\mathbf{\Gamma T}^{\mathrm{T}}), \left[\rho\mathbf{C} + (1-\rho)\mathbf{I}_{\mathrm{n_s}}\right] \otimes \mathbf{V}\right)$. The random effects are modelled as $\mathbf{L}^{\mathrm{R}} = \mathbf{\Pi H \Lambda}$. The data are modelled as $y_{ij} \sim \boldsymbol{D}_j(z_{ij})$, $\mathrm{vec}(\mathbf{Z}) \sim N(\mathrm{vec}(\mathbf{L}), \mathbf{\Sigma} \otimes \mathbf{I}_{\mathrm{n}})$, where $\boldsymbol{D}_j$ refers to the statistical distribution for the data model related to the j-th species, which may vary among different species. The current implementation

of Hmsc includes normal, probit, Poisson (with log link function) and lognormal Poisson models. For more details, especially on the ecological reasoning and interpretation of these equations, we refer to Ovaskainen et al. (2017).

2. The prior distribution of HMSC

In the notation introduced above, the primary parameters of Hmsc can be presented as the vector $\boldsymbol{\theta} = (\mathbf{B}, \rho, \boldsymbol{\Gamma}, \mathbf{V}, \mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\delta}, \boldsymbol{\Sigma})$. The prior density $p(\boldsymbol{\theta}) = p(\mathbf{B}, \rho, \boldsymbol{\Gamma}, \mathbf{V}, \mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\delta}, \boldsymbol{\Sigma})$ decomposes to a product of three parts (corresponding to the fixed effects, random effects, and data models) as $p(\boldsymbol{\theta}) = p(\mathbf{B}, \rho, \boldsymbol{\Gamma}, \mathbf{V})p(\mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\delta})p(\boldsymbol{\Sigma})$.

2.1. Prior distribution of fixed effects

The prior of fixed effects decomposes as $p(\mathbf{B}, \rho, \boldsymbol{\Gamma}, \mathbf{V}) = p(\mathbf{B}|\rho, \boldsymbol{\Gamma}, \mathbf{V})p(\rho)p(\boldsymbol{\Gamma})p(\mathbf{V})$. Here the first part corresponds to modelling species niches as a function of traits and phylogetic relationships as $\mathrm{vec}(\mathbf{B}) \sim N\left(\mathrm{vec}(\boldsymbol{\Gamma}\mathbf{T}^{\mathrm{T}}), \left[\rho\mathbf{C} + (1-\rho)\mathbf{I}_{n_s}\right] \otimes \mathbf{V}\right)$.

- The prior distribution for the influence of traits on species niches is given by $\mathrm{vec}(\boldsymbol{\Gamma}) \sim N\left(\boldsymbol{\mu}_\gamma, \mathbf{U}_\gamma\right)$. The default values of the prior $\boldsymbol{\mu}_\gamma$ (vector of length $n_t n_c$) is the zero vector. The default value for the prior parameter $\mathbf{U}_\gamma$ (variance-covariance matrix of dimension $n_t n_c \times n_t n_c$) is the identity matrix. This means that the default prior for each element γ_{lk} of the matrix $\boldsymbol{\Gamma}$ is $\gamma_{lk} \sim N(0,1)$, and different elements of the matrix $\boldsymbol{\Gamma}$ are a priori independent of each other.
- The default prior of HMSC for the phylogenetic signal parameter ρ is

$$\begin{cases} \text{with probability } 0.5 & \rho = 0, \\ \text{with probability } \dfrac{1}{2n_\rho} & \alpha_h = m\dfrac{\alpha^*}{n_\rho},\ m = 1 \ldots n_\rho \end{cases}$$

  Here n_ρ is the number of points in the descrete grid prior. As the $n_\rho \to +\infty$, the prior distribution of ρ approaches the limit of a spike-and-slab prior, where half of the probability mass is allocated to zero and the other half is distributed as $\rho \sim \mathrm{Uniform}(0,1)$. This choice for the prior is done to avoid forcing a phylogenetic signal in cases where the evidence for such as signal in the data is weak.

- The prior distribution of the parameter $\mathbf{V}$ is the Inverse Wishart distribution $\mathbf{V} \sim W^{-1}(\mathbf{V}_0, f_0)$. The default prior for the scale matrix $\mathbf{V}_0$ ($n_c \times n_c$ matrix) is the identity matrix, and the default value for the degrees of freedom parameter f_0 is $f_0 = n_c + 1$.

We note that the default choices of the prior parameters described above can be considered as reasonable default priors if the user applies the default scaling of the data matrices. As default in the current implementation, prior to MCMC sampling, both the environmental covariate data and the species' trait data are scaled to have zero mean and unit variance. After MCMC sampling, the estimated paramerers are back-transformed so that they are on the scale of the original data, and thus the scaling remains "invisible" for the user. Alternatively, the user can switch off the automatic scaling and transform the data explicitly if needed.

2.2. Prior distribution of random effects

The prior of the random effecs decomposes as $p(\mathbf{H}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\delta}) = p(\mathbf{H}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\delta})$. There can be multiple random effects in the same HMSC model, we describe here the prior for one random effect.

- In the case of a spatial random effect, the prior for the spatial scale α_h of each factor h is

$$\begin{cases} \text{with probability } 0.5 & \alpha_h = 0, \\ \text{with probability } \dfrac{1}{2n_\alpha} & \alpha_h = m\dfrac{\alpha^*}{n_\alpha},\ m = 1 \ldots n_\alpha \end{cases}$$

  Here α^* is set to the maximum distance within the enclosing rectangle of all sampling units and n_α is the number of points in the descrete grid prior. As $n_\alpha \to +\infty$, the prior distribution of α_h approaches the limit of a spike-and-slab prior, where half of the probability mass is allocated to zero and the other half is distributed as $\alpha_h \sim \text{Uniform}(0, \alpha^*)$.

  In spatial modes, the prior for the site loadings **H** is a Gaussian process prior with exponential covariance function and unit marginal variance. This prior implies that $\boldsymbol{\eta}_{\cdot h} \sim \boldsymbol{\Sigma}_h)$, where the variance-covariance matrix $\boldsymbol{\Sigma}_h$is a correlation matrix with elements $\boldsymbol{\Sigma}_{hu_1u_2} = \exp\left(-\frac{d(u_1,u_2)}{\alpha_h}\right)$, and $d(u_1, u_2)$ is the distance between units $u_1$and u_2. In non-spatial models, $\boldsymbol{\Sigma}_h$ is set to an identity matrix.

- The prior for the species loadings is the multiplicative gamma process shrinking prior that Bhattacharya and Dunson (2011) proposed for modelling of high-dimensional covariance matrices, such as the **Ω** matrix for a large number of species. The density of the multiplicative gamma process shrinking prior decomposes as $p(\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\delta}) = p(\boldsymbol{\Lambda}|\boldsymbol{\Phi}, \boldsymbol{\delta})p(\boldsymbol{\Phi})p(\boldsymbol{\delta})$. The matrix **Φ** with elements ϕ_{hj} has the same dimensions $n_f \times n_s$ as the matrix **Λ**, and it models local shrinkage of species loadings. The parameter $\boldsymbol{\delta}$ with elements δ_h is a vector of length n_f, and it models global shrinkage of species loadings. The prior is defined as

$$\lambda_{hj}|\phi_{hj}, \delta \sim N\left(0, \phi_{hj}^{-1}\tau_h^{-1}\right),\ \ \tau_h = \Pi_{l=1}^{h}\,\delta_l,$$

$$\phi_{hj}|\upsilon \sim \text{Ga}(\upsilon/2, \upsilon/2),$$

$$\delta_1|a, b \sim \text{Ga}(a_1, b_1), \qquad \delta_l|a, b \sim \text{Ga}(a_2, b_2) \text{ for } l \geq 2.$$

  The default values of the prior parameters are $\upsilon = 3$, $a = (50{,}50)$ and $b = (1{,}1)$. Increasing the value of the a_1 parameter increases shrinkage in general, whereas increasing the value of the a_2 parameter increasese how much additional shrinkage is applied to factor $h + 1$ compared to factor h (see Battacharya and Dunson 2011 for more details). The deterministic relationship between species loadings and the association matrix $\boldsymbol{\Omega} = \boldsymbol{\Lambda}^{\mathbf{T}}\boldsymbol{\Lambda}$ is illustrated in the DAG as a dashed arrow.

2.3. Prior distribution of data models

In the probit and Poisson models, the likelihood of the data depends solely on the linear predictor without any additional parameters. In the probit model $y_{ij} \sim \text{Bernoulli}\left(\Phi\left(L_{ij}\right)\right) = 1\left(z_{ij} > 0\right)$, where $z_{ij} = L_{ij} + \varepsilon_{ij}$, and $\varepsilon_{ij} \sim N(0{,}1)$. In the Poisson model $y_{ij} \sim \text{Poisson}\left(\exp\left(L_{ij}\right)\right)$, a distribution approximated due to

implementation specifics by Poisson$\left(\exp\left(z_{ij}\right)\right)$, where $z_{ij} = L_{ij} + \varepsilon_{ij}$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, and σ_ε^2 is a small positive constant.

The normal and lognormal Poisson models have the additional non-fixed variance parameter σ_j^2, as in the normal model $y_{ij} = z_{ij}$ and in the lognormal Poisson model $y_{ij} \sim \text{Poisson}\left(\exp\left(z_{ij}\right)\right)$, where in both models $z_{ij} = L_{ij} + \varepsilon_{ij}$ and $\varepsilon_{ij} \sim N(0, \sigma_j^2)$. The prior distribution for each σ_j^2 is the inverse-gamma distribution with shape parameter $a_{\sigma j}$ and rate parameter $b_{\sigma j}$, where the prior parameters $a_{\sigma j}$ and $b_{\sigma j}$ are to be decided by the user. As default, Hmsc assumes that $a_{\sigma j} = 1$ and $b_{\sigma j} = 5$ for each species j.

The diagonal elements of the matrix **Σ** consist of all the residual variance parameters σ_j^2, both non-fixed and fixed.

2.4. Which part of the prior distribution is most important from the user's point of view?

As described above, the prior distribution of HMSC is relatively complex, simply for the reason that HMSC is a hierarchical multivariate model that includes many parameters for which a prior distribution needs to be defined. Thus, a relevant question to ask is how sensitive the HMSC results will be with respect to the choice of the prior, and when and how the user should adjust the prior distribution.

We recommend the user generally to assume the default prior distribution, unless the user is a true expert in Bayesian analyses and has a good reason for choosing another prior. The default prior can be expected to be a reasonable choice for most cases, assuming that the user also follows e.g. the default scaling of the **X** and **T** matrices, which are specifically made for the reason of making the default prior generally applicable. Adjustring the prior without having a good understanding of how and why to do it is likely to lead to undesired results.

If there are much data, the results are likely to be relatively unsensitive with respect to the prior choice, with one exception. The exception is the prior for the species loadings, in particular the values of the prior parameters $a = (a_1, a_2)$ for which the default values are $a = (50{,}50)$. As described bove, these two values determine the amount of shrinkage for specis loadings. If decreasing the values of these parameters e.g. to $a = (5{,}5)$, less shrinkage will be applied. With less shrinkage, HMSC will estimate a stronger and structurally richer species association network, but this comes with the risk of overfitting, i.e. capturing noise rather than signal. Conversely, if increasing the shrinkage, e.g. with $a = (500{,}500)$, the risk of overfitting decreases, but at the same time the risk of loosing the signal in species associations increases. The extent to which the estimated association networks model signal vs. noise can be examined through conditional cross-validation, where the occurrences of the focal species are predicted conditional on the known occurrences of the other species. The ideal level of shrinkage is likely to depend also on the number of species involved in the study, more shrinkage needed for large species communities than for small species communities.

3. Posterior sampling in HMSC

The HMSC package employs the/a conditional block Gibbs scheme for estimation of the posterior distribution of the model parameters via MCMC sampling. This scheme iteratively updates each of the model parameters from its conditional distribution, conditional on other parameters being fixed to their current values. However, unlike a standard Gibbs scheme such as those implemented in generic MCMC software (e.g. JAGS or BUGS), our sampler updates simultaneously not only scalar parameters, but entire

vectors and matrices of parameters. This enables higher computational efficiency due to the elimination of potential autocorrelation in the MCMC between the jointly updated parametrs, while retaing the same asymptotical numerical complexity. Additionally, this scheme often allows exploiting vectorized operations, which enables further speed-up of the implementation. To enable such advantages, wherever possible, the HMSC is designed in such a way that the conditional distributions have analytically tractable forms. For those few parameters that do not allow such analytically tractability, we employ discrete grid sampling, based on the prior-weighted conditional likelihood.

Below, we list the internal functions of the Hmsc package that perform the Gibbs steps during posterior sampling. These are presented in two groups. The first group includes the core samplers, which are most universal with regard to the model structure and therefore applied in all types of models. These samplers primarily follow the structure of Figure S1, so that each vector or matrix of parameters represented by an ellipse in Fig S1 is sampled conditional on the other parameters. The second group includes the additional samplers, which exploit partial analytical marginalization to enable sampling parameters represented by several ellipses in Fig. 1 at once. Similarly to block vs non-block Gibbs schemes, this enhancement allows further reducing autocorrelation in MCMC. However, in this case, such marginalization can lead to an increase of asymptotic numerical complexity, and computationally efficient formulas for the required sampling may exist only for certain cases. Therefore, these samplers do not properly cover the whole spectrum of model variants that HMSC supports, but nevertheless try to cover the most common cases. Furthermore, in some cases the increased computational cost associated with the additional samplers of the second group can overweight the benefits of the decreased autocorrelation in MCMC. Thus, the final decision of whether to use these samplers or not is ultimately left for the person applying the model.

3.1. Group 1: Core samplers

3.1.1. The function updateBetaLambda

This function updates the $\mathbf{B}$ and $\mathbf{\Lambda}$ parameters jointly conditional on the $\mathbf{Z}$, $\mathbf{\Gamma}$, $\mathbf{H}$, $\mathbf{V}$, $\mathbf{\Sigma}$ and ρ parameters. We denote $\mathbf{Q} = \rho\mathbf{C} + (1-\rho)\mathbf{I}_{n_s}$ and $\boldsymbol{\tau} = \left[\tau_1, \ldots, \tau_{n_f}\right]$, $\tau_h = \prod_{h'=1}^{h} \delta_{h'}$. Then the joint conditional distribution for vector $\boldsymbol{w} = [\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T]^T$ is

$$\boldsymbol{w} \sim N(\boldsymbol{m}_{\boldsymbol{w}}, \mathbf{\Sigma}_{\boldsymbol{w}})$$

$$\mathbf{\Sigma}_{\boldsymbol{w}}^{-1} = \begin{bmatrix} \mathbf{V}^{-1} \otimes \mathbf{Q}^{-1} & 0 \\ 0 & \mathrm{diag}\big(\mathrm{vec}(\mathrm{diag}(\boldsymbol{\tau})\mathbf{\Psi})\big)^{-1} \end{bmatrix} + \begin{bmatrix} \mathbf{X} \\ \mathbf{H} \end{bmatrix}^T \begin{bmatrix} \mathbf{X} \\ \mathbf{H} \end{bmatrix} \otimes \mathbf{\Sigma}^{-1}$$

$$\boldsymbol{m}_{\boldsymbol{w}} = \mathbf{\Sigma}_{\boldsymbol{w}} \left(\begin{bmatrix} \mathbf{V}^{-1} \otimes \mathbf{Q}^{-1} & 0 \\ 0 & \mathrm{diag}\big(\mathrm{vec}(\mathrm{diag}(\boldsymbol{\tau})\mathbf{\Psi})\big)^{-1} \end{bmatrix} \begin{bmatrix} \mathrm{vec}(\mathbf{\Gamma}\mathbf{T}^T) \\ \mathbf{0}_{n_f n_s \times 1} \end{bmatrix} + vec\left(\begin{bmatrix} \mathbf{X} \\ \mathbf{H} \end{bmatrix}^T \mathbf{Z}\mathbf{\Sigma}^{-1} \right) \right)$$

This formula is derived from the fact that conditional on $\mathbf{H}$, the problem can be expressed as a linear regression. In the special case of $\rho = 0$ and thus $\mathbf{Q} = \mathbf{I}_{n_s}$, the joint density above factorizes across species $j = 1 \ldots n_s$, so that it can be more efficiently sampled species-wise

$$\left[\mathbf{B}_{\cdot j}^T, \mathbf{\Lambda}_{\cdot j}^T\right]^T = \boldsymbol{w}_{\boldsymbol{j}} \sim N\left(\boldsymbol{m}_{\boldsymbol{w}_j}, \mathbf{\Sigma}_{\boldsymbol{w}_j}\right)$$

$$\mathbf{\Sigma}_{\boldsymbol{w}_j}^{-1} = \begin{bmatrix} \mathbf{V}^{-1} & 0 \\ 0 & \mathrm{diag}\left(\mathrm{vec}\big(\mathrm{diag}(\boldsymbol{\tau})\mathbf{\Psi}_{\cdot j}\big)\right)^{-1} \end{bmatrix} + \sigma_j^{-2} \begin{bmatrix} \mathbf{X} \\ \mathbf{H} \end{bmatrix}^T \begin{bmatrix} \mathbf{X} \\ \mathbf{H} \end{bmatrix}$$

$$\boldsymbol{m}_{\boldsymbol{w}_j} = \boldsymbol{\Sigma}_{\boldsymbol{w}_j}\left(\begin{bmatrix}\mathbf{V}^{-1} & 0\\ 0 & \operatorname{diag}\left(\operatorname{vec}\left(\operatorname{diag}(\boldsymbol{\tau})\boldsymbol{\Psi}_{\cdot j}\right)\right)^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Gamma}\mathbf{T}_{j\cdot}^{T}\\ \mathbf{0}_{n_f\times 1}\end{bmatrix} + \sigma_j^{-2}\begin{bmatrix}\mathbf{X}\\ \mathbf{H}\end{bmatrix}^{T}\mathbf{Z}_{\cdot j}\right)$$

3.1.2. The function updateGammaV

This function first updates the $\mathbf{V}$ parameter conditional on the $\mathbf{B}$,$\boldsymbol{\Gamma}$ and ρ parameters and then the $\boldsymbol{\Gamma}$ parameter conditional on the $\mathbf{B}$,$\mathbf{V}$ and ρ parameters. We denote $\mathbf{Q} = \rho\mathbf{C} + (1-\rho)\mathbf{I}_{n_s}$. Then $\mathbf{V}$ is sampled according to

$$\mathbf{V} \sim W^{-1}(\mathbf{V}_0 + (\mathbf{B} - \boldsymbol{\Gamma}\mathbf{T}^T)\mathbf{Q}^{-1}(\mathbf{B} - \boldsymbol{\Gamma}\mathbf{T}^T)^T, f_0 + n_s).$$

Denoting $\mathbf{U} = \mathbf{U}_\gamma^{-1} + \mathbf{T}^T\mathbf{Q}^{-1}\mathbf{T} \otimes \mathbf{V}^{-1}$, $\boldsymbol{\Gamma}$ is sampled according to

$$\boldsymbol{\Gamma} \sim N\left(\mathbf{U}^{-1}\left(\mathbf{U}_\gamma^{-1}\boldsymbol{\mu}_\gamma + (\mathbf{T}^{\mathrm{T}}\mathbf{Q}^{-1} \otimes \mathbf{V}^{-1})\boldsymbol{\beta}\right), \mathbf{U}^{-1}\right).$$

The derivation of the formula for $\mathbf{V}$ follows the proof of the Inverse-Wishart distribution being conjugate for multivariate normal with known mean. The derivation of the latter formula is done by applying the conditional multivariate Gaussian to the joint distribution of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$.

3.1.3. The function updateRho

This function updates the ρ parameter conditional on the $\mathbf{B}$, $\boldsymbol{\Gamma}$ and $\mathbf{V}$ parameters. ρ is sampled from the grid of prior values $\frac{m}{n_\rho}, m = 0 \ldots n_\rho$ proportional to the conditional likelihood of the grid values

$$p(\mathbf{B}|\boldsymbol{\Gamma}, \mathbf{V}, \rho) = N\left(\boldsymbol{\beta}|\operatorname{vec}(\boldsymbol{\Gamma}\mathbf{T}^T), \left[\rho\mathbf{C} + (1-\rho)\mathbf{I}_{n_s}\right]\otimes\mathbf{V}\right)$$

weighted by the prior distribution of ρ.

3.1.4. The fuction updateLambdaPriors

This function first updates the $\boldsymbol{\Psi}$ parameters conditional on the $\boldsymbol{\Lambda}$, and $\boldsymbol{\delta}$ parameters. It then updates the $\boldsymbol{\delta}$ parameters conditional on the $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ parameters. We denote $\tau_h = \prod_l^h \delta_l$. $\boldsymbol{\Psi}$ is sampled according to

$$\psi_{hj} \sim \operatorname{Gamma}\left(\frac{\nu}{2} + 0.5, \frac{\nu}{2} + \lambda_{hj}^2\tau_h\right)$$

Then the $\boldsymbol{\delta}$ parameters are sampled sequentially as

$$\delta_1 \sim \operatorname{Gamma}\left(a_1 + \frac{n_s n_f}{2}, b_1 + \delta_1^{-1}\sum_{l=1}^{n_f}\left(\tau_l\sum_{j=1}^{n_s}\lambda_{lj}^2\psi_{lj}\right)\right)$$

$$\delta_h \sim \operatorname{Gamma}\left(a_2 + \frac{n_s(n_f - h + 1)}{2}, b_2 + \delta_h^{-1}\sum_{l=h}^{n_f}\left(\tau_l\sum_{j=1}^{n_s}\lambda_{lj}^2\psi_{lj}\right)\right)$$

A detailed derivation of these formulas is provided in Bhattacharya and Dunson (2011).

3.1.5. The function updateEta

This function updates the $\mathbf{H}$ parameters conditional on the $\mathbf{Z}$, $\mathbf{B}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}$ parameters. We denote by $\mathbf{S} = \mathbf{Z} - \mathbf{XB}$. Then

$$\boldsymbol{\eta} \sim N(\mathbf{U}\operatorname{vec}(\boldsymbol{\Pi}^T\mathbf{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}^T), \mathbf{U})$$

$$\mathbf{U}^{-1} = \mathrm{diag}\left(\mathbf{K}(\alpha_1), \ldots, \mathbf{K}\left(\alpha_{n_f}\right)\right) + \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \otimes \boldsymbol{\Pi}^T\boldsymbol{\Pi}$$

In case of non-spatial latent factors, each $\mathbf{K}(\alpha_h) = \mathbf{I}_{n_u}$ and the distribution of different rows of $\mathbf{H}$ are conditionally independent, allowing for a simplification of the general formula above:

$$\mathbf{H}_{u\cdot} \sim N\left(\mathbf{U}_u \mathrm{vec}\left(\widehat{\boldsymbol{S}}_{u\cdot}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}^T\right), \mathbf{U}_u\right)$$
$$\mathbf{U}_u^{-1} = \mathbf{I}_{n_f} + (\boldsymbol{\Pi}^T\boldsymbol{\Pi})_{uu} \cdot \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$$

Where $\widehat{\boldsymbol{S}} = \boldsymbol{\Pi}^T\mathbf{S}$ and $(\boldsymbol{\Pi}^T\boldsymbol{\Pi})_{uu}$ is the number of rows of $\mathbf{Y}$ that correspond to the unit u of the latent factor. The derivation of these formulas is based on conditional multivariate Gaussian of the joint distribution of $\boldsymbol{\eta}$ and $\boldsymbol{z}$.

3.1.6. The function updateAlpha

This function updates the $\boldsymbol{\alpha}$ parameters for each spatial random effect level conditional on the matrix of latent factors $\mathbf{H}$ and the structure of that spatial random level, e.g. the coordinates of the spatial units. The conditional distribution of each α_h is sampled from the grid of prior values $m\frac{\alpha^*}{n_\alpha}, m = 0 \ldots n_\alpha$ proportional to the conditional likelihood of the grid values $p(\mathbf{H}_{\cdot h}|\alpha_h) = N\left(\mathbf{H}_{\cdot h}|\mathbf{0}_{n_y\times 1}, \mathbf{K}(\alpha_h)\right)$ weighted by the prior distribution of α_h, where $K_{uu'}(\alpha_h) = \exp(-|s_{u\cdot}, s_{u'\cdot}|/\alpha_h)$.

3.1.7. The function updateInvSigma

This function updates the $\boldsymbol{\sigma}$ parameters conditional on the $\mathbf{Z}$,$\mathbf{B}$,$\mathbf{H}$ and $\boldsymbol{\Lambda}$ parameters. Denoting $\mathbf{E} = \mathbf{Z} - \mathbf{XB} - \boldsymbol{\Pi}\mathbf{H}\boldsymbol{\Lambda}$, each σ_j is sampled as

$$\sigma_j \sim \mathrm{Ga}\left(a_{\sigma j} + 0.5\sum_{i=1}^{n} 1\left(y_{ij} \neq \mathrm{NA}\right), b_{\sigma j} + 0.5\sum_{i=1}^{n} \epsilon_{ij}^2 \cdot 1\left(y_{ij} \neq \mathrm{NA}\right)\right)$$

The derivation of this sampling formula follows the proof of conjugancy of Gaussian likelihood with known mean.

3.1.8. The function updateZ

This function updates the $\mathbf{Z}$ parameters conditional on the $\mathbf{B}$, $\mathbf{H}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}$ parameters. We denote by $\mathbf{E} = \mathbf{Z} - \mathbf{XB} - \boldsymbol{\Pi}\mathbf{H}\boldsymbol{\Lambda}$. The sampling formula depends on what type of data model is assumed for each species:

a. Normal. $z_{ij} = y_{ij}$

b. Probit. $z_{ij} \sim N_{[l,u]}\left(e_{ij}, \sigma_j^2\right)$, where $N_{[l,u]}(\mu, \sigma^2)$ is a truncated Normal distribution with mean μ variance σ^2, and the truncation interval $[l, u]$ that depends on y_{ij}: $\begin{cases} y_{ij} = 0 \Longrightarrow l = -\infty, u = 0 \\ y_{ij} = 1 \Longrightarrow l = 0, u = +\infty \end{cases}$

c. Poisson and overdispersed Poisson.

$$z_{ij} \sim N\left(\frac{\left(\sigma_j^{-2} + w\right)^{-1}\left(y_{ij} - r\right)}{2} + \sigma_j^{-2}\left(e_{ij} - \log r\right) + \log r\,, \left(\sigma_j^{-2} + w\right)^{-1}\right)$$

where w is a sample from Polya-Gamma distribution $w \sim PG\left(y_{ij} + r, z_{ij} - \log r\right)$ and $r = 1000$ – a large constant that determines how the Poisson distribution is approximated with Negative-

Binomial as the number of failures increases. To enable a non-overdispersed Poisson distribution with this algorithm, the residual variance σ_j is artificially fixed to a small positive value $\sigma_j = 0.01$.

d. Missing observation. $z_{ij} \sim N(e_{ij}, \sigma_j^2)$

These formulas represent the known data augmentation scheme for different types of data. The proofs can be found in Albert and Chib (1993) for binary data and Zhou et al. (2012) for count data.

3.1.9. The function updateNf

This function adaptively adjusts the number of latent factors n_f. On the M-th step of the MCMC sampling scheme, the function computes with probability $e^{-(c_0+c_1 M)}$ the proportion of elements for each row of $\mathbf{\Lambda}$ that are by absolute value less than ϵ. If none of the proportions is greater than Δ, a new latent factor and the corresponding latent loading are added, both of which are initialized by sampling from the prior. Otherwise, the latent factors that correspond to proportions exceeding Δ are removed. Default values for c_0, c_1, ϵ and Δ are 1, 0.0005, 0.001 and 0.995 correspondingly. More details on this adaptation algorithm are given in Bhattacharya and Dunson (2011).

3.2. Group 2: Additional samplers

3.2.1. The function updateGamma2

This fuction updates the $\mathbf{\Gamma}$ parameters conditional on the $\mathbf{Z}$,$\mathbf{V}$,$\mathbf{\Sigma}$,$\mathbf{H}$,$\mathbf{\Lambda}$ and ρ parameters. This updater combines elements from the functions updateBetaLambda and updateGammaV, as it integrates over the $\mathbf{B}$ parameters in order to decrease the autoregression of the Gibbs sampler due to conditional dependencies in the sampling. Currently this is implemented only for models without phylogenetic information, e.g. when $\mathbf{C} = \mathbf{I}_{n_s}$, $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \mathbf{0}_{n_c n_t}$ and $\mathbf{U}_\gamma = \mathbf{I}_{n_t} \otimes \widehat{\mathbf{U}}_{\boldsymbol{\gamma}}$, since this is a common particular case that enables specifically efficient numerical implementation.

We denote by $\mathbf{S} = \mathbf{Z} - \mathbf{\Pi H \Lambda}$, then

$$\boldsymbol{\gamma} \sim N(\boldsymbol{m}_{\boldsymbol{\gamma}}, \mathbf{\Sigma}_{\boldsymbol{\gamma}})$$

$$\boldsymbol{m}_{\boldsymbol{\gamma}} = vec\left(\widehat{\mathbf{U}}_{\boldsymbol{\gamma}}(\mathbf{X}^T\mathbf{S}\mathbf{T} - \mathbf{X}^T\mathbf{X}(\mathbf{V}^{-1} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{S}\mathbf{T})\right)$$
$$- \left(\mathbf{T}^T\mathbf{T} \otimes \widehat{\mathbf{U}}_{\boldsymbol{\gamma}}\mathbf{X}^T\mathbf{X}(\mathbf{V}^{-1} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{V}^{-1}\right) \cdot \left(\mathbf{I}_{n_t} \otimes \widehat{\mathbf{U}}_{\boldsymbol{\gamma}}^{-1} + \mathbf{T}^T\mathbf{T} \otimes (\mathbf{V}^{-1} - \mathbf{V}^{-1}(\mathbf{V}^{-1} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{V}^{-1})\right)^{-1}$$
$$\cdot vec(\mathbf{V}^{-1}(\mathbf{V}^{-1} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{S}\mathbf{T})$$

$$\mathbf{\Sigma}_{\boldsymbol{\gamma}} = \mathbf{I}_{n_t} \otimes \widehat{\mathbf{U}}_{\boldsymbol{\gamma}} - \mathbf{T}^T\mathbf{T} \otimes \left(\widehat{\mathbf{U}}_{\boldsymbol{\gamma}}\mathbf{X}^T\mathbf{X}\widehat{\mathbf{U}}_{\boldsymbol{\gamma}} - \widehat{\mathbf{U}}_{\boldsymbol{\gamma}}\mathbf{X}^T\mathbf{X}(\mathbf{V}^{-1} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\widehat{\mathbf{U}}_{\boldsymbol{\gamma}}\right)$$
$$+ \left(\mathbf{T}^T\mathbf{T}\right.$$
$$\otimes \widehat{\mathbf{U}}_{\boldsymbol{\gamma}}\mathbf{X}^T\mathbf{X}(\mathbf{V}^{-1} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{V}^{-1}\left)\right(\mathbf{I}_{n_t} \otimes \widehat{\mathbf{U}}_{\boldsymbol{\gamma}}^{-1} + \mathbf{T}^T\mathbf{T}$$
$$\left.\otimes (\mathbf{V}^{-1} - \mathbf{V}^{-1}(\mathbf{V}^{-1} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{V}^{-1})\right)^{-1}\left(\mathbf{T}^T\mathbf{T} \otimes \widehat{\mathbf{U}}_{\boldsymbol{\gamma}}\mathbf{X}^T\mathbf{X}(\mathbf{V}^{-1} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{V}^{-1}\right)^T$$

This quite long formula could be derived from the expressions for conditional multivariate Gaussian distribution – particularly from the joint distribution of $\begin{bmatrix}\boldsymbol{\gamma} \\ vec(\boldsymbol{S})\end{bmatrix}$, conditioning on the $vec(\boldsymbol{S})$.

3.2.2. The function updateGammaEta

This function updates the **Γ** and **H** parameters conditional on the **Z**,**V**,**Σ**,**Λ**,$\boldsymbol{\alpha}$ and ρ parameters. It basically extends the ideas of the updateGamma2 function, detailed in the previous subsection, but additionally attempts to marginalize over the **H** parameters. This sampler is currenly implemented only for the case $\boldsymbol{\mu_\gamma} = \mathbf{0}_{n_c n_t}$. The multivariate Gaussian sampling formula is derived from the joint multivariate Gaussian distribution of

$$\begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\eta} \\ vec(\boldsymbol{S}) \end{bmatrix} \sim N\left(\boldsymbol{m}_{\boldsymbol{\gamma\eta S}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma\eta S}}\right)$$

by conditioning on the $vec(\boldsymbol{S})$. While this expression is srtraightforwardly analytically tractable, in practice it turns out that the directly involved expressions are quite comprehensive to simplify in order to acheive computationally scalable results. On the other hand, it is appears to be easier to work with an augmented multivariate Gaussian distribution

$$\begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\beta} \\ \boldsymbol{\eta} \\ vec(\boldsymbol{S}) \end{bmatrix} \sim N\left(\boldsymbol{m}_{\boldsymbol{\gamma\beta\eta S}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma\beta\eta S}}\right)$$

First, we can obtain the conditional distrubution for $\boldsymbol{\beta}$ from

$$\begin{bmatrix} \boldsymbol{\beta} \\ vec(\boldsymbol{S}) \end{bmatrix} \sim N\left(\boldsymbol{m}_{\boldsymbol{\beta S}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta S}}\right)$$

And then due to conditional dependence structure, obtain the distributions of $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ independently of each other conditional on the $vec(\boldsymbol{S})$ and sampled $\boldsymbol{\beta}$. These are done in line with the standard updaters described above. However, we specifically stress that unlike a sequential application of the standard updaters for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$, this joint updater samples directly from the joint conditional distribution of these three parameters.

The relevant formulas are quite cumbersome and involve quite lengthy linear algebra exercises. Furthermore, they vary depending on the particular model's structure. We intentionally replaced them here with a verbal conceptual description of how the sampling is handled. For those readers who are specifically interested in this updater, we redirect to its source code available in the HMSC github repo.

4. Dependency on other R-packages

The dependency of Hmsc on other packages can be explored from NAMESPACE, which lists which functions from other R-packages are imported. The R-packages from which functions are imported are BayesLogit, FNN, MASS, MCMCpack, Matrix, abind, ape, coda, fields, ggplot2, grDevices, graphics, methods, mvtnorm, nnet, pROC, parallel, pdist, phytools, statmod, stats and truncnorm.

5. References

Albert, J. H., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association **88**:669-679.

Bhattacharya, A., and D. B. Dunson. 2011. Sparse Bayesian infinite factor models. Biometrika **98**:291-306.

Ovaskainen, O., G. Tikhonov, A. Norberg, F. G. Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology Letters **20**:561-576.

Vanhatalo, J., M. Hartmann, and L. Veneranta. 2018. Joint species distribution modeling with additive multivariate Gaussian process priors and heteregenous data. ArXiv e-prints.
Zhou, M., L. Li, D. Dunson, and L. Carin. 2012. Lognormal and gamma mixed negative binomial regression. Proceedings of the International Conference on Machine Learning **2012**:1343-1350.