Joint Species Distribution Modeling

GJAM Gibbs sampling implementation in RCpp

Students: Contini Matteo, Fiorello Lorenzo, Pasquale Angelo

Supervisors: Poggiato Giovanni[a], Dr. Corradin Riccardo

February 19, 2020

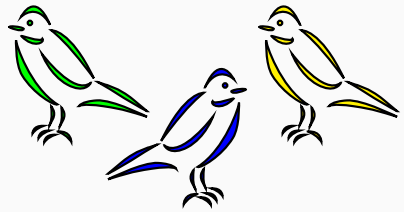
Master of Science in Mathematical Engineering - Bayesian Statistics

[a]Laboratoire d'Ecologie Alpine, Grenoble.

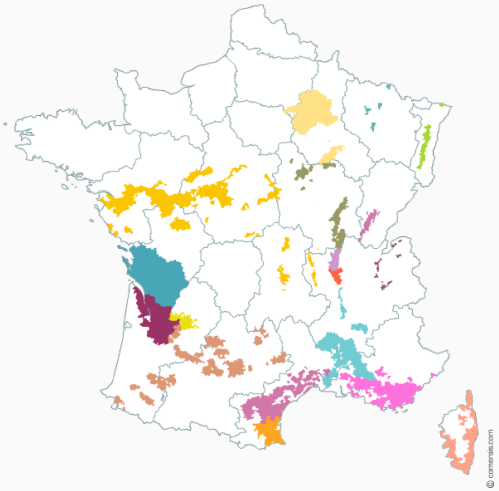
Summing up

Species Distribution Models

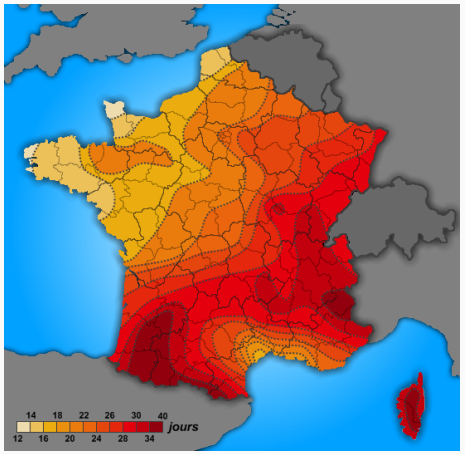
Species



Species distributions



Environmental covariates



The model

Our variables

- Subscript notation:
  - sites: $i = 1 \ldots, n$;
  - species: $j = 1 \ldots, S$;
  - covariates: $k = 2, ..., K$.
- Response variable $\boldsymbol{Y} \in \{0, 1\}^{n \times S}$:

$$Y_{ij} = \begin{cases} 1 & \text{if species } j \text{ is present at site } i \\ 0 & \text{otherwise} \end{cases}$$

- Latent variable $\boldsymbol{V} \sim \mathcal{N}_{n,S}(\mathbf{mean}, \mathbf{var})$: $Y_{ij} = 1$ iff $V_{ij} > 0$.
- Three models:
  - JSDM: Joint Species Distribution Models;
  - HMSC: Hierarchical Model of Species Communities;
  - GJAM: Generalized Joint Additive Model.

Core JSDM model

Let $\boldsymbol{V} \sim \mathcal{N}_{n,S}(\mathbf{mean}, \mathbf{var})$ be a *latent variable* related to $\boldsymbol{Y}$ ($Y_{ij} = 1$ iff $V_{ij} > 0$).

The model:

$$V_{ij} = \boldsymbol{X_{i.}}\, \boldsymbol{B_{j.}}^T + e_{ij}$$
$$B_{jk} \overset{\text{ind}}{\sim} \mathcal{N}(\omega_k, \sigma_k)$$
$$\boldsymbol{e_{i.}} \overset{\text{iid}}{\sim} \mathcal{MVN}(\mathbf{0}, \boldsymbol{R})$$

where

- $\boldsymbol{X}$: $n \times K$ matrix of measured covariates (we set $X_{i1} = 1$ for the intercept);
- $\boldsymbol{B}$: $S \times K$ matrix of regression parameters (B_{jk} is the response of species j to the covariate k).

Core JSDM model

The model:

$$V_{ij} = \boldsymbol{X_{i.}}\,\boldsymbol{B_{j.}}^{T} + e_{ij}$$
$$B_{jk} \overset{\text{ind}}{\sim} \mathcal{N}(\omega_k, \sigma_k)$$
$$\boldsymbol{e_{i.}} \overset{\text{iid}}{\sim} \mathcal{MVN}(\mathbf{0}, \boldsymbol{R})$$

where

- $\boldsymbol{R} \longrightarrow$ is the residual covariance matrix that captures interspecies dependencies

Core JSDM model

The model:

$$V_{ij} = \boldsymbol{X_{i.}}\,\boldsymbol{B_{j.}}^{T} + e_{ij}$$
$$B_{jk} \overset{\text{ind}}{\sim} \mathcal{N}(\omega_k, \sigma_k)$$
$$\boldsymbol{e_{i.}} \overset{\text{iid}}{\sim} \mathcal{MVN}(\mathbf{0}, \boldsymbol{R})$$

where

- $\boldsymbol{R} \longrightarrow$ is the residual covariance matrix that captures interspecies dependencies
- $\boldsymbol{B} \longrightarrow$ each line describes the dependence of species j on the environmental covariates

Core JSDM model

The model:

$$V_{ij} = \boldsymbol{X_{i.}}\,\boldsymbol{B_{j.}}^{T} + e_{ij}$$
$$B_{jk} \overset{\text{ind}}{\sim} \mathcal{N}(\omega_k, \sigma_k)$$
$$\boldsymbol{e_{i.}} \overset{\text{iid}}{\sim} \mathcal{MVN}(\mathbf{0}, \boldsymbol{R})$$

Numerically challenging since $\boldsymbol{R}$ full rank matrix mapping to a space of $\frac{S(S+1)}{2}$ unrestricted parameters.

Core JSDM model

The model:

$$V_{ij} = \boldsymbol{X_{i.}}\, \boldsymbol{B_{j.}}^T + e_{ij}$$
$$B_{jk} \overset{\text{ind}}{\sim} \mathcal{N}(\omega_k, \sigma_k)$$
$$\boldsymbol{e_{i.}} \overset{\text{iid}}{\sim} \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{R})$$

Numerically challenging since $\boldsymbol{R}$ full rank matrix mapping to a space of $\frac{S(S+1)}{2}$ unrestricted parameters.

$$\boldsymbol{R} \overset{\text{HMSC}}{\longmapsto} \boldsymbol{R} = \boldsymbol{A}^T\, \boldsymbol{A} + \sigma_\varepsilon^2\, \boldsymbol{I_S}$$

where $\boldsymbol{A}$ is a $S \times n_f$ matrix.
After this transformation we'll have $\mathcal{O}(S)$ parameters

An extension: the GJAM model

Small S (Core Model)

$$
\begin{aligned}
Y_{ij} &= 1(V_{ij} > 0) \\
V_{ij} &= \boldsymbol{X_{i.}}\,\boldsymbol{B_{j.}}^T + e_{ij} \\
B_{jk} &\overset{\text{ind}}{\sim} \mathcal{N}(\omega_k, \sigma_k) \\
\boldsymbol{e_{i.}} &\overset{\text{iid}}{\sim} \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{R}) \\
\omega_k &\overset{\text{iid}}{\sim} \mathcal{N}(0, 100) \\
\sigma_k &\overset{\text{iid}}{\sim} \mathcal{U}(0, 100) \\
\boldsymbol{R} &\sim \mathcal{IW}(S+1, \boldsymbol{I})
\end{aligned}
$$

Big S (HMSC)

$$
\begin{aligned}
Y_{ij} &= 1(V_{ij} > 0) \\
V_{ij} &= \boldsymbol{X_{i.}}\,\boldsymbol{B_{j.}}^T + e_{ij} \\
e_{ij} &= \boldsymbol{W_{i.}}\,\boldsymbol{A_{j.}}^T + \varepsilon_{ij} \\
\boldsymbol{W_{i.}} &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{n_f}) \\
\boldsymbol{\varepsilon_{i.}} &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma_\varepsilon^2\, \boldsymbol{I}_S) \\
\sigma_\varepsilon^2 &\sim \mathcal{IG}(0.01, 0.01)
\end{aligned}
$$

- $\boldsymbol{W}$: $n \times n_f$ matrix of latent factors (for all sites);
- $\boldsymbol{A}$: $S \times n_f$ matrix of latent loadings (for all species).

Dimension Reduction Using Dirichlet Processes

The residual correlation depends on the $S \times n_f$ matrix $\boldsymbol{A}$:

$$\boldsymbol{\Sigma} = \boldsymbol{A}^T \boldsymbol{A} + \sigma_\varepsilon^2 \boldsymbol{I}_S$$

What if some species share the same behaviour with respect to other species?

Dimension Reduction Using Dirichlet Processes

The residual correlation depends on the $S \times n_f$ matrix $\boldsymbol{A}$:

$$\boldsymbol{\Sigma} = \boldsymbol{A}^T \boldsymbol{A} + \sigma_\varepsilon^2 \boldsymbol{I}_S$$

What if some species share the same behaviour with respect to other species?

Cluster interpretation

Species that share the same rows of $\boldsymbol{A}$ (i.e. are in the same clusters) are those that **share the same residual correlation with respect to other species**.

Dimension Reduction Using Dirichlet Processes

The residual correlation depends on the $S \times n_f$ matrix $\boldsymbol{A}$:

$$\boldsymbol{\Sigma} = \boldsymbol{A}^T \boldsymbol{A} + \sigma_\varepsilon^2 \boldsymbol{I}_S$$

What if some species share the same behaviour with respect to other species?

Cluster interpretation

Species that share the same rows of $\boldsymbol{A}$ (i.e. are in the same clusters) are those that **share the same residual correlation with respect to other species**.

How?

$\Longrightarrow$ exploit the clustering properties of the *Dirichlet Process* (Bayesian non-parametric prior).

GJAM with stick-breaking approximation

[Taylor-Rodríguez et al. (2017)]

$$\boldsymbol{V}_{i\cdot} \mid \boldsymbol{k}, \boldsymbol{Z}, \boldsymbol{w}_i, \boldsymbol{B}, \sigma_\varepsilon^2 \overset{\text{iid}}{\sim} \mathcal{MVN}(\boldsymbol{B}\,\boldsymbol{x}_i + \boldsymbol{Q}(\boldsymbol{k})\,\boldsymbol{Z}\,\boldsymbol{w}_i, \sigma_\epsilon^2\,\boldsymbol{I}_S) \quad \text{for } i = 1, \ldots, n$$

$$(\boldsymbol{B}, \sigma_\epsilon^2) \propto \frac{1}{\sigma_\epsilon^2}$$

$$\boldsymbol{w}_i \overset{\text{iid}}{\sim} \mathcal{MVN}(0, \boldsymbol{I}_{n_f})$$

$$k_l \mid \boldsymbol{p} \overset{\text{iid}}{\sim} \sum_{j=1}^{N} p_j \delta_j(k_l) \quad \text{for } l = 1, \ldots, S$$

$$\boldsymbol{Z}_j \mid \boldsymbol{D}_z \overset{\text{iid}}{\sim} \mathcal{MVN}(0, \boldsymbol{D}_z) \quad \text{for } j = 1, \ldots, N$$

$$\boldsymbol{p} \sim \mathcal{GD}_N(a_\alpha, b_\alpha) \quad \text{with} \quad \begin{cases} a_\alpha = \left(\frac{\alpha}{N}, \ldots, \frac{\alpha}{N}\right) \\ b_\alpha = \left(\frac{\alpha(N-1)}{N}, \frac{\alpha(N-2)}{N}, \ldots, \frac{\alpha}{N}\right) \end{cases}$$

$$\boldsymbol{D}_z \sim \mathcal{IW}\left(2 + n_f - 1, 4\,\text{diag}\left(\frac{1}{\eta_1}, \ldots, \frac{1}{\eta_{n_f}}\right)\right)$$

$$\eta_h \overset{\text{iid}}{\sim} \mathcal{IG}(0.5, 10^{-4}) \quad \text{for } h = 1, \ldots, n_f$$

Sampling strategy

Given the too complicated expression of the posterior joint density we implement a *Gibbs sampler*, based on the following full conditional densities:

1. $[\boldsymbol{Z} \mid \boldsymbol{D_z}, \boldsymbol{B}, \boldsymbol{W}, \sigma^2_\varepsilon, \boldsymbol{V}]$
2. $[\boldsymbol{W} \mid \boldsymbol{B}, \boldsymbol{A}, \sigma^2_\varepsilon, \boldsymbol{V}]$
3. $[\boldsymbol{k} \mid \boldsymbol{p}, \boldsymbol{B}, \boldsymbol{Z}, \sigma^2_\varepsilon, \boldsymbol{V}]$
4. $[\boldsymbol{p} \mid \boldsymbol{k}]$
5. $[\sigma^2_\varepsilon \mid \boldsymbol{A}, \boldsymbol{W}, \boldsymbol{B}, \boldsymbol{V}]$
6. $[\boldsymbol{D_z} \mid \boldsymbol{Z}]$
7. $[V_i \mid \boldsymbol{B}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Y}]$
8. $[\boldsymbol{B} \mid \boldsymbol{X}, \boldsymbol{V}]$

Full conditionals

- $[\boldsymbol{Z} \mid \boldsymbol{D_z}, \boldsymbol{B}, \boldsymbol{W}, \sigma_\varepsilon^2, \boldsymbol{V}]$: The posterior for each row of $\boldsymbol{Z}$ depends on whether or not the row considered was chosen to be at least one row from A. That is, for $j = 1, \ldots N$
  - If $j \notin k$, sample
  $$\boldsymbol{Z}_j \sim \mathcal{N}_r(\boldsymbol{0}, \boldsymbol{D_z}) \tag{1.1}$$
  - If $j \in k$, let $S_j = \{l = 1, \ldots S \text{ s.t. } k_l = j\}$ and let
  $$\boldsymbol{Z} \mid \boldsymbol{D_z}, \boldsymbol{B}, \boldsymbol{W}, \sigma_\varepsilon^2, \boldsymbol{V} \overset{\text{ind}}{\sim} \mathcal{N}_r(\mu_{z_l}, \Sigma_{\boldsymbol{Z}_j}) \tag{1.2}$$
  where $\Sigma_{\boldsymbol{Z}_j} = \left(\frac{|S_j|}{\sigma_\epsilon^2} \boldsymbol{W}^T \boldsymbol{W} + \boldsymbol{D_z}^{-1} \right)^{-1}$,
  $\mu_{\boldsymbol{Z}_j} = \Sigma_{\boldsymbol{Z}_j} \boldsymbol{W}^T \frac{1}{\sigma_\varepsilon^2} \sum_{l \in S_j} (\boldsymbol{V}^{(l)} - \boldsymbol{X}\, \beta_l)$, and finally, $\boldsymbol{V}^{(l)}$ and β_l are the l-th column of the matrix $\boldsymbol{V}$ and the l-th row of $\boldsymbol{B}$, respectively.
- $[\boldsymbol{W} \mid \boldsymbol{B}, \boldsymbol{A}, \sigma_\varepsilon^2, \boldsymbol{V}]$
$$\boldsymbol{w}_i \mid \boldsymbol{B}, \boldsymbol{A}, \sigma_\epsilon^2, \boldsymbol{V}_i \sim \mathcal{N}_r \left(\Sigma_{\boldsymbol{W}}\, \boldsymbol{A}\, \frac{1}{\sigma_\varepsilon^2} (\boldsymbol{V}_i - \boldsymbol{B}\, x_i), \Sigma_{\boldsymbol{W}} \right) \tag{2}$$
where $\Sigma_{\boldsymbol{W}} = (\frac{1}{\sigma_\epsilon^2} \boldsymbol{A}^T \boldsymbol{A} + I_r)^{-1}$

Full conditionals

- $[\boldsymbol{k} \,|\, \boldsymbol{p}, \boldsymbol{B}, \boldsymbol{Z}, \sigma^2_\varepsilon, \boldsymbol{V}]$

$$[\boldsymbol{k} \,|p, \boldsymbol{B}, \boldsymbol{Z}, \sigma^2_\epsilon, \boldsymbol{V}] = \prod_{l=1}^{q} \left\{ \sum_{j=1}^{N} p_{lj} \delta_j(k_l) \right\} \tag{3}$$

with $p_{lj} \propto p_j \times \exp[-\frac{1}{2\sigma^2_\epsilon} || \, \boldsymbol{V}^{(l)} - \boldsymbol{X} \, \beta_l - \boldsymbol{W} \, Z_j ||^2]$

- $[\boldsymbol{p} \,|\, \boldsymbol{k}]$ The full conditional posterior for $\boldsymbol{p}$, given conjugancy of the $\mathcal{GD}$ distribution with multinomial sampling, the draws of $\boldsymbol{p}$ are

$$\begin{aligned} p_1 &= \xi_1 \\ p_j &= (1 - \xi_1) \ldots (1 - \xi_{j-1}) \, \xi_j \;\text{ for }\; j = 2, 3, \ldots, N-1 \\ p_N &= 1 - \sum_{j=1}^{N-1} p_j \end{aligned} \tag{4}$$

with

$$\xi_j \overset{\text{ind}}{\sim} Beta\left(\tfrac{\alpha}{N} + \textstyle\sum_{l=1}^{S} I_{(k_l = j)}, \tfrac{N-j}{N} \alpha + \sum_{s=j+1}^{N} \sum_{l=1}^{S} I_{(k_l = s)} \right)$$

Full conditionals

- $[\sigma_\varepsilon^2 | \boldsymbol{A}, \boldsymbol{W}, \boldsymbol{B}, \boldsymbol{V}]$

$$\sigma_\epsilon^2 | \boldsymbol{A}, \boldsymbol{W}, \boldsymbol{B}, \boldsymbol{V} \sim \mathcal{IG}\left(\frac{nS + \nu}{2} + 1, \frac{\sum_{i=1}^{n} || \boldsymbol{V}_i - \boldsymbol{B}\, x_i - \boldsymbol{A}\, w_i||^2}{2} + \frac{\nu}{G^2} \right) \tag{5}$$

- $[\boldsymbol{D_z} \,|\, \boldsymbol{Z}]$: By conjugacy of the prior $\boldsymbol{D_z}$ with the normal prior for $\boldsymbol{Z}$, the full conditional for $\boldsymbol{D_z}$ is

$$\boldsymbol{D_z} \,|\, \boldsymbol{Z} \sim \mathcal{IW}\left(\boldsymbol{D_z} \,|2 + r + N - 1, \boldsymbol{Z}'\, \boldsymbol{Z} + 4\text{diag}\left\{ \frac{1}{\eta_1}, \ldots, \frac{1}{\eta_r} \right\} \right) \tag{6}$$

Prior modification

$$(\boldsymbol{B}, \sigma_\epsilon^2) = \pi(\boldsymbol{B}) \times \frac{1}{\sigma_\epsilon^2}$$

where $\pi(\boldsymbol{B})$ is such that all the coefficients are independent across species:

$$\forall j = 1, \ldots, S \qquad \boldsymbol{\beta_j} \overset{iid}{\sim} N(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

where $\boldsymbol{\beta_j}$ is the j-th row of $\boldsymbol{B}$ (if $\boldsymbol{B}$ is $S \times k$) and σ is an hyperparameter that gives the degree of uninformativeness of the prior.

$\boldsymbol{D} = diag(\boldsymbol{\Sigma}^\star)$, with $\boldsymbol{\Sigma}^\star = \boldsymbol{A}\,\boldsymbol{A}' + \sigma_\epsilon^2\,\boldsymbol{I}$.

Without loss of generality, we perform these changes of variables:

$$\begin{cases} \boldsymbol{V}^\star = \boldsymbol{V}\,\boldsymbol{D}^{1/2} \\ \boldsymbol{B}^\star = \boldsymbol{D}^{1/2}\,\boldsymbol{B} \end{cases}$$

$\boldsymbol{V} = \boldsymbol{D}^{-\frac{1}{2}}\,\boldsymbol{V}^\star \rightarrow \boldsymbol{V} = \boldsymbol{V}^\star\,\boldsymbol{D}^{-\frac{1}{2}}$

Full Conditionals for steps 7 and 8

- $[\boldsymbol{V}_i^\star \mid \boldsymbol{B}^\star, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Y}]$: for each site $i = 1, \ldots, n$ and each species $j = 1, \ldots, S$ we sample $\boldsymbol{V}_{i,j}^\star$ from a univariate truncated normal:

$$\boldsymbol{V}_{i,j}^\star \mid \boldsymbol{B}^\star, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{W}, y_{i,j} \sim trunc.N(\boldsymbol{B_{j.}}^\star \boldsymbol{x_i} + \boldsymbol{A}\, \boldsymbol{w_i}, \sigma_\epsilon^2) \tag{7}$$

  Where the normal is truncated to the positive axis if $y_{i,j} = 1$ and to the negative axis if $y_{i,j} = 0$.

- $[\boldsymbol{B}^\star \mid \boldsymbol{V}^\star, \boldsymbol{X}, \boldsymbol{A}]$: for each species $j = 1, \ldots, S$ we sample $\beta_j^\star$ from a multivariate normal

$$\boldsymbol{\beta_j}^\star \sim N((\frac{1}{\sigma^2}\, \boldsymbol{I} + \frac{1}{\sigma_\epsilon^2}\, \boldsymbol{X}'\, \boldsymbol{X})^{-1} \frac{1}{\sigma_\epsilon^2}\, \boldsymbol{X}'(\boldsymbol{V_{.j}}^\star - \boldsymbol{W}\, \boldsymbol{A_{j.}}), (\frac{1}{\sigma^2}\, \boldsymbol{I} + \frac{1}{\sigma_\epsilon^2}\, \boldsymbol{X}'\, \boldsymbol{X})^{-1}) \tag{8}$$

  Where $\boldsymbol{W}$ is the $n \times r$ matrix whose lines are $\boldsymbol{w_i}$.

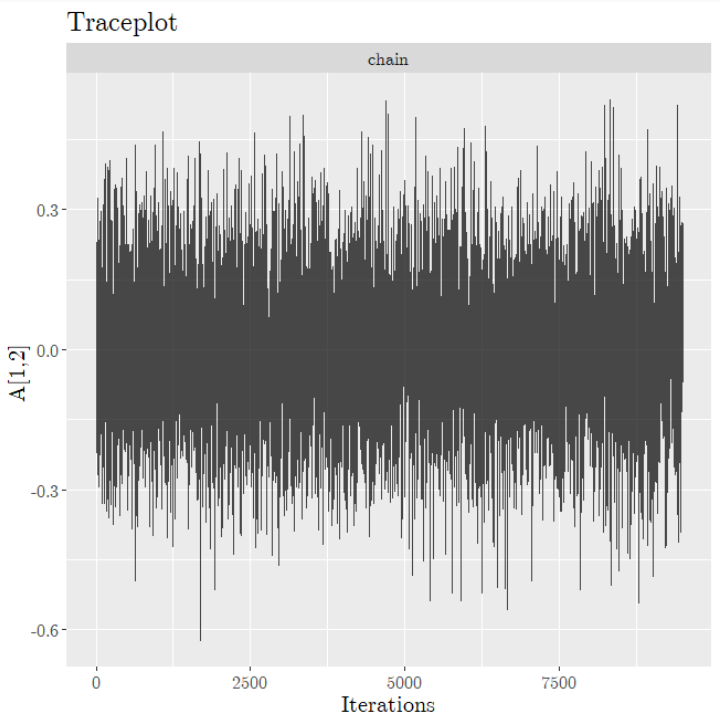
Pseudo-Code

1: **function** GJAM_GIBBS_SAMPLER( )
2: **for** $j = 1, \ldots, N$ **do**
resample Z_j according to the following
3: **if** $j \notin \boldsymbol{k}$ **then** sample from 1.1
4: **else** sample Z_j from 1.2
5: **end if**
6: **end for**
7: **for** $i = 1, \ldots, N$ **do** resample $\boldsymbol{w_i}$ from 2
8: **end for**
Resample the vector of labels $\boldsymbol{k}$ from 3
Resample the vector of labels $\boldsymbol{p}$ from 4, thus
9: **for** $j = 1, \ldots, N-1$ **do** sample ξ_j and thus p_j
10: **end for**
$p_N = 1 - \sum_{j=1}^{N-1} p_j$
sample σ_ϵ^2 from 5
sample $\boldsymbol{D_z}$ from 6
sample $\boldsymbol{V}$ from 7
sample $\boldsymbol{B}$ from 8
11: **end function**

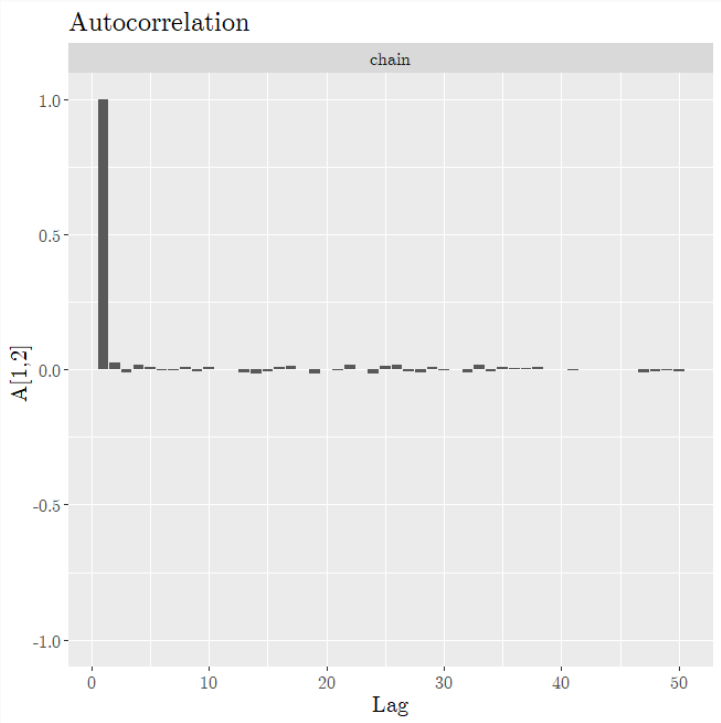
Implementation

- the procedure is computationally expensive
- `R` code vs `C++` code
- efficient linear algebra with the `Eigen library` and the package `RcppEigen`

Study of the chain convergence

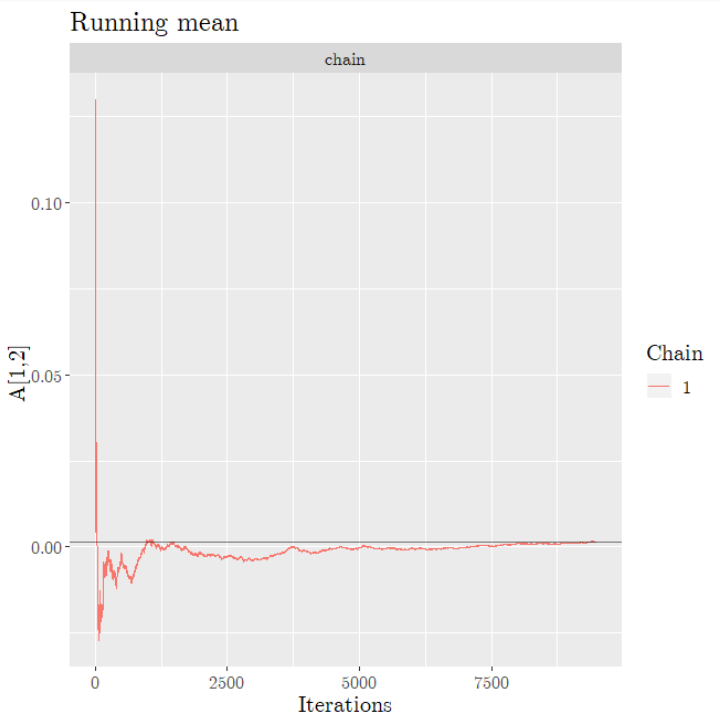


(a) Traceplot.

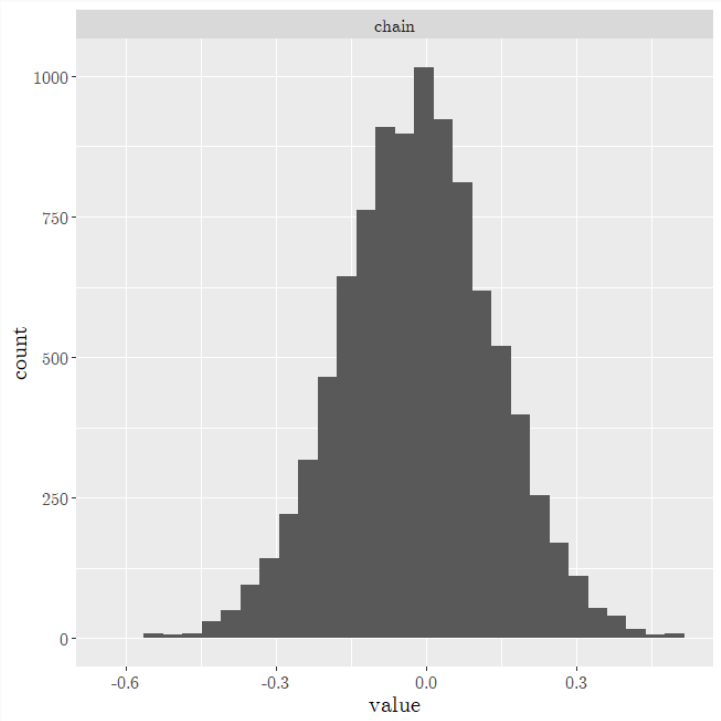


(b) Autocorrelation.

Study of the chain convergence



(c) Cumulative mean.



(d) Frequencies.

Speedup of our code

S	n	r	$N_{stick} = \min(150, S)$	**time [s]**	
				Taylor&al	Rcpp
150	100	4	150	644.344	274.189
50	100	4	50	203.803	84.918
30	100	4	30	100.205	40.082
20	100	4	20	87.177	39.626

When S grows, we need increase number of Gibbs iterations for having convergence.

Speedup of our code

X2.31

Further Developments

- Different results might be obtained with different priors for the parameters
- Parallelization of some steps
- Instead of treating the number of latent factor `r` as fixed, we can use a shrinkage approach to let the model choose the best number of latent factor, by implementing in our model the work of Bhattacharya and Dunson (2011)

References

A. Bhattacharya and D. B. Dunson. "Sparse Bayesian infinite factor models". In: *Biometrika* 98, (June 2011), pp. 291–306. DOI: 10.1093/biomet/asr013.

Jorge Soberón and Andrew Peterson. "Interpretation of Models of Fundamental Ecological Niches and Species' Distributional Areas". In: *Biodiversity Informatics* 2 (Jan. 2005). DOI: 10.17161/bi.v2i0.4.

Daniel Taylor-Rodríguez et al. "Joint Species Distribution Modeling: Dimension Reduction Using Dirichlet Processes". In: *Bayesian Anal.* 12.4 (Dec. 2017), pp. 939–967. DOI: 10.1214/16-BA1031. URL: https://doi.org/10.1214/16-BA1031.

Thanks for your attention!