Politecnico di Milano

MSc in Mathematical Engineering
Applied Statistics

Final Project of:
Bayesian Statistics

Joint Species Distribution Models

Gibbs sampling implementation in `RCpp`

Students:
Contini Matteo
Fiorello Lorenzo
Pasquale Angelo

Advisers:
Poggiato Giovanni[a]
Dr. Corradin Riccardo

Course Lecturer:
Prof. Guglielmi Alessandra

[a]Laboratoire d'Ecologie Alpine - Inria Grenoble

Milan - December 8, 2019

Abstract

abstract-text

Contents

1 Review and comparison of Joint Species Distribution Models (JSDM)

1.1 Models description

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

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

- Latent variable $\boldsymbol{z} \sim \mathcal{N}_{N,J}(\mathbf{mean}, \mathbf{var})$ (associated to $\boldsymbol{y}$).
- $\boldsymbol{X} \in \mathbb{R}^{N \times K}$ matrix of measured covariates.

1.2 The core model (CM)

One of the first JSDMs was proposed by Pollock et al. (2014), and we will refer to it as the core model (CM). This model is built on the multivariate probit regression model (Chib and Greenberg (1998)), by using a latent variable parametrisation of a probit model rather than the probit link directly.

The community (i.e. the set of species) present at site i is thus characterized by the multidimensional latent variable $z_{i,.}$:

$$\begin{aligned} y_{ij} &= 1(z_{ij} > 0) \\ z_{ij} &= B_{.j} X_{i.} + e_{ij} \\ B_{kj} &\overset{\text{ind}}{\sim} \mathcal{N}(\omega_j, \sigma_j) \\ e_{i.} &\overset{\text{iid}}{\sim} \mathcal{MVN}(\mathbf{0}, \boldsymbol{R}) \end{aligned} \tag{CM}$$

where

$$\begin{cases} \omega_j \overset{\text{iid}}{\sim} \mathcal{N}(0, 100) \\ \sigma_j \overset{\text{iid}}{\sim} \mathcal{U}(0, 100) \\ \boldsymbol{R} \sim \mathcal{IW}(J+1, \boldsymbol{I}) \quad \text{(correlation coefficients prior)} \end{cases}$$

1.3 Hierarchical Model of Species Communities (HMSC)

Hierarchical Model of Species Communities (HMSC) is a model appeared in a sequence of papers (Ovaskainen et al. (2017), Tikhonov and Ovaskainen (2017), Tikhonov, Abrego, et al. (2017)) that aim to give a very complete framework that takes into account all possible information about species in one single hierarchical model.

HMSC is a very similar to (CM), but allows the regression coefficients to be correlated:

$$\begin{aligned} y_{ij} &= 1(z_{ij} > 0) \\ z_{ij} &= B_{.j} X_{i.} + e_{ij} \\ B_{.j} &\overset{\text{ind}}{\sim} \mathcal{MVN}(\boldsymbol{\omega}, \boldsymbol{\Lambda}) \\ e_{ij} &= \nu_{ij} + \varepsilon_{ij} \\ \nu_{ij} &= \eta_{i.} \boldsymbol{A_{j.}} \\ \varepsilon_{ij} &\overset{\text{iid}}{\sim} \mathcal{MVN}(0, 1) \\ \eta_{i.} &\overset{\text{iid}}{\sim} \mathcal{MVN}(0, \boldsymbol{I}_{n_f}) \end{aligned} \tag{HMSC}$$

where $\boldsymbol{\omega} = (\omega_k)_{k=1,\dots,K}$ and

$$\begin{cases} \omega_k \overset{\text{iid}}{\sim} \mathcal{N}(0, 100) & \\ \boldsymbol{\Lambda} \sim \mathcal{IW}(5, \boldsymbol{I}) & \text{(variance-covariance matrix of the regression coefficients)} \\ \boldsymbol{R} \sim \mathcal{IW}(J+1, \boldsymbol{I}) & \text{(correlation coefficients prior)} \end{cases}$$

(HMSC) biggest improvement concerns the different representation of the error e_{ij}. In the (CM), the full rank matrix $\boldsymbol{R}$ represents the variation in species occurrences and co-occurrences that cannot be attributed to the responses of the species to the measured covariates.

With S species, each covariance matrix $\boldsymbol{R}$ is bijective mapping to a space of $\frac{J(J+1)}{2}$ unrestricted parameters (Lewandowski, Kurowicka, and Joe (2009)), making their estimation numerically challenging. In order to reduce the parameter space, (HMSC) uses latent factors and latent loadings. Under the classic assumption made in factor iid models that the latent factors marginally follow multivariate normal distribution $\eta_{i.} \overset{\text{iid}}{\sim} \mathcal{MVN}(0, \boldsymbol{I}_{n_f})$, the latent loadings provide then a parametrisation of $\boldsymbol{R}$ as $\boldsymbol{R} = \boldsymbol{A}^T \boldsymbol{A} + \boldsymbol{I}_S$, where $\boldsymbol{A} = \{A_{ij}\}$ is the $S \times n_f$ matrix containing all latent loadings. The utility of the latent factor approach comes from the dimension-reduced parametrization of $\boldsymbol{R}$ in case where $n_f \ll S$. Instead of fixing the number of latent factors n_f, (HMSC) treats n_f as an unknown parameter through the shrinkage approach of Bhattacharya, Pati, and Dunson (2014). This variance decomposition could be considered similar to a linear regression where the latent loadings $A_{j,q}$ are the parameters of the regression, and the latent factors are interpreted to model some missing covariates, which have an impact on the species occurrences and are not represented in the matrix.

1.4 Generalized Joint Additive Model (GJAM)

GJAM is a Joint Species Distribution Model that aims to fit all type of response data, using a latent variable. This is an important feature: since in ecology the collection of data can be very heterogeneous, it is suitable to have a single model to deal with multifarious data. For presence-absence data it is a multivariate probit regression model that takes on two different forms depending on S , the number of species to be modeled, for the same reasons we discussed above: when the number of species S is big, the model suffers from the "curse of dimensionality". The small dataset form (i.e. when S is small) is equivalent to the core model (CM), but the regression coefficients B_{jk} are indipendent and vague:

$$\begin{aligned} y_{ij} &= 1(z_{ij} > 0) \\ z_{ij} &= B_{.j} X_{i.} + e_{ij} \\ B_{.j} &\overset{\text{ind}}{\sim} \mathcal{MVN}(\boldsymbol{0}, 100\, \boldsymbol{I}) \\ e_{i.} &\overset{\text{iid}}{\sim} \mathcal{MVN}(\boldsymbol{0}, \boldsymbol{R}) \\ \boldsymbol{R} &\sim \mathcal{IW}(J+1, \boldsymbol{I}) \end{aligned} \tag{GJAM1}$$

The big dataset form (i.e. when S is big and dimension reduction is needed) proposes a latent factor approach similar to (HMSC):

$$\begin{aligned} y_{ij} &= 1(z_{ij} > 0) \\ z_{ij} &= B_{.j} X_{i.} + \boldsymbol{A} \end{aligned} \tag{GJAM2}$$

ask GioP to review notation from CM to GJAM

2 GJAM using Bayesian nonparametric priors

2.1 The Dirichlet Process

2.1.1 Theoretical background

The polya urn representation of the DP can also be reinterpreted as a Chinese restaurant process (CRP). The CRP(α) is a single parameter distribution over partitions of integers.
The idea of this representation is the following: suppose a restaurant have infinite number of tables and sequence

of customers labeled by $1, ...n$.
The first customer sits at the first table and then each new customer joins a table populated by n_j customers with probability $\frac{n_j}{\alpha+n}$, where n is the overall number of customers that already entered the restaurant, or can sit at a new table with probability $\frac{\alpha}{\alpha+n}$. We can truncate this process after N customers and then, each time we want to change our partition, we take the last customer from his table and we assign him to a new one according to the same rule as before. In this way we'll deal with a finite process and we'll have the possibility of updating our posterior by reassign customers at each iteration.

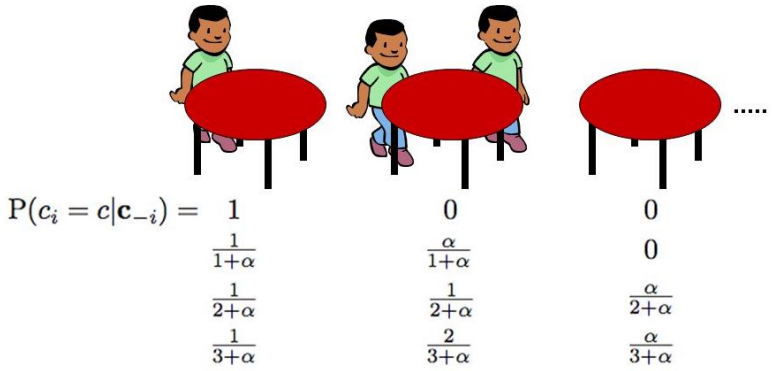

Figure 1: Representation of the Chinese process

The CRP is a random partition induced by DP.

2.1.2 Approximation

2.2 Application to GJAM

3 Sampling strategy

To construct a sampler for Bayesian statistical models it is necessary to specify priors over all of the parameters for which we want to make inference. These express our prior belief (before observing any data) about the probability distribution of the parameters.

3.1 Gibbs sampling

- $[\boldsymbol{Z} \mid \boldsymbol{D_z}, \boldsymbol{B}, \boldsymbol{W}, \sigma^2_\varepsilon, \boldsymbol{V}]$: 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 $j = 1, \dots N$

  - If $j \notin k$, sample
  $$\boldsymbol{Z}_j \sim \mathcal{N}_r(0, \boldsymbol{D_z}) \tag{1}$$
  - If $j \in k$, let $S_j = \{l = 1, \dots S s.t. k_l = j\}$ and let
  $$\boldsymbol{Z} \mid \boldsymbol{D_z}, \boldsymbol{B}, \boldsymbol{W}, \sigma^2_\varepsilon, \boldsymbol{V} \overset{\text{ind}}{\sim} \mathcal{N}_r(\mu_{z_l}, \Sigma_{\boldsymbol{Z}_j}) \tag{2}$$
  where $\Sigma_{\boldsymbol{Z}_j} = \left(\frac{|S_j|}{\sigma^2_\varepsilon} \boldsymbol{W}^T \boldsymbol{W} + \boldsymbol{D_z}^{-1}\right)^{-1}$, $\mu_{\boldsymbol{Z}_j} = \Sigma_{\boldsymbol{Z}_j} \boldsymbol{W}^T \frac{1}{\sigma^2_\varepsilon} \sum_{l \in S_j} (\boldsymbol{V}^{(l)} - \boldsymbol{X}\, \beta_l)$, 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.

- $[\boldsymbol{W} \mid \boldsymbol{B}, \boldsymbol{A}, \sigma^2_\varepsilon, \boldsymbol{V}]$
$$\boldsymbol{w}_i \mid \boldsymbol{B}, \boldsymbol{A}, \sigma^2_\epsilon, \boldsymbol{V}_i \sim \mathcal{N}_r \left(\Sigma_{\boldsymbol{W}}\, \boldsymbol{A}\, \frac{1}{\sigma^2_\varepsilon} (\boldsymbol{V}_i - \boldsymbol{B}\, x_i)\right) \tag{3}$$
where $\Sigma_{\boldsymbol{W}} = (\frac{1}{\sigma^2_\epsilon}\, \boldsymbol{A}^T\, \boldsymbol{A} + I_r)^{-1}$

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

$$[\boldsymbol{k}\,|p,\boldsymbol{B},\boldsymbol{Z},\sigma_{\epsilon}^2,\boldsymbol{V}] = \prod_{l=1}^{q}\left\{\sum_{j=1}^{N} p_{lj}\delta_j(k_l)\right\} \tag{4}$$

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

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

$$\begin{aligned} p_1 &= \xi_1 \\ p_j &= (1-\xi_1)\ldots(1-\xi_{j-1})\,\xi_j \;\text{ for }\; j = 2,3,\ldots,N-1 \\ p_N &= 1 - \sum_{j=1}^{N-1} p_j \end{aligned} \tag{5}$$

with

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

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

$$\sigma_{\epsilon}^2|\,\boldsymbol{A},\boldsymbol{W},\boldsymbol{B},\boldsymbol{V} \sim \mathcal{IG}\left(\frac{nS+\nu}{2}+1, \frac{\sum_{i=1}^{n}||\,\boldsymbol{V}_i - \boldsymbol{B}\,x_i - \boldsymbol{A}\,w_i||^2}{2} + \frac{\nu}{G^2}\right) \tag{6}$$

- $[\boldsymbol{D_z}\,|\,\boldsymbol{Z}]$

$$\boldsymbol{D_z}\,|\,\boldsymbol{Z} \sim \mathcal{IW}\left(\boldsymbol{D_z}\,|2+r+N-1, \boldsymbol{Z}'\,\boldsymbol{Z} + 4\text{diag}\left\{\frac{1}{\eta_1},\ldots,\frac{1}{\eta_r}\right\}\right) \tag{7}$$

- $[\boldsymbol{V}_i\,|\,\boldsymbol{Y}_i,\beta,\boldsymbol{R}]$

- $[\boldsymbol{B}\,|\textit{to be computed...}]$

  FIRST METHOD : Sample the vector of regression coefficients $\boldsymbol{B}_j$ for each species j from a multivariate normal distribution:

$$\boldsymbol{B}_j \sim \mathcal{N}((\sigma I + X_j' X_j)^{-1} X_j' z_j, (\sigma I + X_j' X_j)^{-1}) \tag{8}$$

  SECOND METHOD : For each element of each vector of regression coefficients $\boldsymbol{B}_j$ we use a diffuse normal prior with mean 0 and variance 100 by setting σ to 10. This is a widely used prior which exhibits little influence on the posterior:

$$B_{.j} \overset{\text{ind}}{\sim} \mathcal{MVN}(\boldsymbol{0}, 100\,\boldsymbol{I}) \tag{9}$$

3.2 Pseudocode for the GJAM

4 Implementation in Rcpp

5 Simulations

Precondition:

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

References

[BPD14] Anirban Bhattacharya, Debdeep Pati, and David Dunson. "Anisotropic function estimation using multi-bandwidth Gaussian processes". In: *Ann. Statist.* 42.1 (Feb. 2014), pp. 352–381. DOI: 10.1214/13-AOS1192. URL: https://doi.org/10.1214/13-AOS1192.

[CG98] S. Chib and E. Greenberg. "Analysis of Multivariate Probit Models". In: *Biometrika* 85 (June 1998), pp. 347–361. DOI: 10.1093/biomet/85.2.347.

[LKJ09] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. "Generating random correlation matrices based on vines and extended onion method". In: *Journal of Multivariate Analysis* 100 (Oct. 2009), pp. 1989–2001. DOI: 10.1016/j.jmva.2009.04.008.

[Ova+17] Otso Ovaskainen et al. "How are species interactions structured in species-rich communities? A new method for analysing time-series data". In: *Proceedings of the Royal Society B: Biological Sciences* 284 (May 2017), p. 20170768. DOI: 10.1098/rspb.2017.0768.

[Pol+14] Laura J. Pollock et al. "Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM)". In: *Methods in Ecology and Evolution* 5.5 (2014), pp. 397–406. DOI: 10.1111/2041-210X.12180. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12180. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12180.

[Tay+17] 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.

[Tik+17] Gleb Tikhonov, Nerea Abrego, et al. "Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context". In: *Methods in Ecology and Evolution* 8 (Apr. 2017), pp. 443–452. DOI: 10.1111/2041-210X.12723.

[TO17] Gleb Tikhonov and Otso Ovaskainen. "Making more out of ecological community data: conceptual framework and its implementation as models and software". In: (Dec. 2017).

Appendices

A Appendix 1.

B Appendix 2.