useR ! 2019 – Thursday, July 11th 2019, – Toulouse

Using Rcpp packages for easy and fast Gibbs sampling MCMC from within R

Ghislain Vieilledent[1] and Jeanne Clément[1]

[1] **Cirad** UMR AMAP

cirad

AMAP botAnique et Modélisation de l'Architecture des Plantes et des végétations

Plan

Plan

botAnique et Modélisation
de l'Architecture des Plantes et des végétations

Seamless R and C++ Integration

C++ code

```cpp
#include <Rcpp.h>
using namespace Rcpp;
// [[Rcpp::export]]
int addition(int a, int b) {
  return  a+b;
}
```

compile and execute

R code

```r
Rcpp::sourceCpp("Code/addition.cpp")
addition(2,2)
```

```
## [1] 4
```

RcppGSL for random draws and RcppArmadillo for matrix calculation

- **GNU Scientific Library** written in C : fiable random number generators, fast random number ditributions, vectors and matrices.

- **Armadillo** C++ linear algebra : cube, matrices and vectors easy generation and manipulation, decompositions, factorisations, inverses and equation solvers.

- **RcppGSL** and **RcppArmadillo** R packages provides an easy-to-use interface between GSL as well as Armadillo data structures or functions and R using.

Benefits of using GSL's randow draws instead of R's

- More available distributions :
  - R : gaussian, uniform, binomial, beta, chi-squared, exponential, fisher, gamma, geometric, hypergeometric, negative binomial, poisson, student, weibull.
  - RcppGSL : gaussian, uniform, binomial, beta, chi-squared, exponential, fisher, gamma, geometric, hypergeometric, negative binomial, poisson, student, weibull, multivariate gaussian, multinomial, whishart, dirichlet, cauchy, lognormal, logistic, logarithmic, pareto, laplace, landau...
- GSL's random draws are implemented in C and optimised.
- Easy implementation of wanted distributions using RcppGSL's elementary distributions.

Rcpp function using RcppGSL and RcppArmadillo

```
1  #include <RcppArmadillo.h>
2  #include <gsl/gsl_rng.h>
3  #include <gsl/gsl_randist.h>
4
5  // [[Rcpp::depends(RcppArmadillo)]]
6  // [[Rcpp::depends(RcppGSL)]]
7  // [[Rcpp::export]]
8
9  Rcpp::List alpha_sample(arma::mat X, arma::mat beta,
10                          arma::mat Y, int Valpha, const int↩
                                seed) {
11   // Defining constants
12   const int NSITE = Y.n_rows;
13   const int NSPECIES = Y.n_cols;
14
15   // Initialize random number generator
16   gsl_rng *s = gsl_rng_alloc(gsl_rng_mt19937);
17   gsl_rng_set(s, seed);
```

Rcpp function using RcppGSL and RcppArmadillo

```
19    // Declaring new objects to store results
20    arma::vec alpha; alpha.zeros(NSITE);
21
22    // Draw in the posterior distribution
23    for (int i=0; i < NSITE ; i++) {
24      double small_v = arma::sum(Y.row(i)-X.row(i)*beta);
25      double big_V = 1/(1/Valpha + NSPECIES);
26      alpha(i) = big_V*small_v + gsl_ran_gaussian_ziggurat(s↩
            , std::sqrt(big_V));
27    }
28
29    Rcpp::List results = Rcpp::List::create(Rcpp::Named("↩
        alpha") = alpha);
30    return results;
31  }
```

Plan

2. Rcpp functions to fit hierarchical Species Distribution Models
   - Models definition
   - Inference method
   - Evaluation of their efficiency on simulated dataset
   - Compilation time of Gibbs sampler in R or in C++ with GSL and R draws

botAnique et Modélisation
de l'Architecture des Plantes et des végétations

Models definition

Data and parameters :

- Response variable : $Y = (y_i)_{i=1,\dots,nsite}$ such as :

$$y_i = \begin{cases} 0 & \text{if the species is absent on the site } i \\ 1 & \text{if the species is present on the site } i. \end{cases}$$

- Explanatory variables : $X = (X_i)_{i=1,\dots,nsite}$ avec $X_i = (x_i^1, \dots, x_i^p)' \in \mathbb{R}^p$ where p is the number of bioclimatic variables considered for each site i.
- Species fixed effect : β_0 and $\beta = (\beta_1, \dots, \beta_p)'$, the intercept and regression coefficients which are assumed to have prior distribution $\mathcal{N}(0, 10^6)$.

Models definition

Model 1 : hSDM with logit link function

$$\text{logit}(\theta_i) = \beta_0 + X_i'\beta$$

- Link function logit : $\text{logit} : p \to = \log(p) - \log(1-p)$.
- $y_i \sim \text{Bin}(n_i, \theta_i)$ with n_i the number of visits for site i.

Model 2 : hSDM with probit link function

$$\text{probit}(\theta_i) = \beta_0 + X_i'\beta$$

- Link function probit : $\text{probit} : q \to \Phi^{-1}(q)$ where Φ correspond to the $\mathcal{N}(0,1)$ cumulative distribution function.
- Latent variable $z_i = \beta_0 + X_i'\beta + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0,1)$ and such as :

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

- We easily shown that $y_i|z_i \sim \mathcal{B}ernouilli(\theta_i)$.

Model 1 : Gibbs sampler with Metropolis algorithm

Gibbs sampler :

Use to get sample of random variable, $Z = (Z_1, Z_2, Z_3)$ for example, with known probability distributions $\pi_i(z_i)$ for $i = 1, 2, 3$.

- **Initialisation** : $z^{(0)} = 0_{\mathbb{R}^3}$.
- **Iteration t** : Generate $z^{(t)}$ as follows :
- $z_1^{(t)} \sim \pi_1 \left(z_1 \, | z_2^{(t-1)}, z_3^{(t-1)} \right)$
- $z_2^{(t)} \sim \pi_2 \left(z_2 \mid z_1^{(t)}, z_3^{(t-1)} \right)$
- $z_3^{(t)} \sim \pi_3 \left(z_3 \mid z_1^{(t)}, z_2^{(t)} \right)$

So we need conditional posterior distributions of betas which are not explicitly calculable in this case.

Model 1 : Gibbs sampler with Metropolis algorithm

Metropolis Hastings algorithm :

Use to generate betas according to an estimation of their conditional posterior distributions.

- **Initialisation** : $z^{(0)} = 0_{\mathbb{R}^3}$.
- **Iteration t** :
- Generate $z^* \sim q(z^{(t-1)}, .)$, as symetric proposal distribution $q(z^{(t-1)}, .)$ we have chosen $\mathcal{N}(z^{(t-1)}, 1)$.
- Calculate the acceptance probability :

$$\alpha = min\left(1, \frac{\pi(z^*)}{\pi(z^{(t-1)})}\right)$$

.

- Retain

$$z^{(t)} = \begin{cases} z^* & \text{with probability } \alpha \\ z^{(t-1)} & \text{with probability } 1-\alpha. \end{cases}$$

Model 2 : Gibbs sampler with conjugate priors

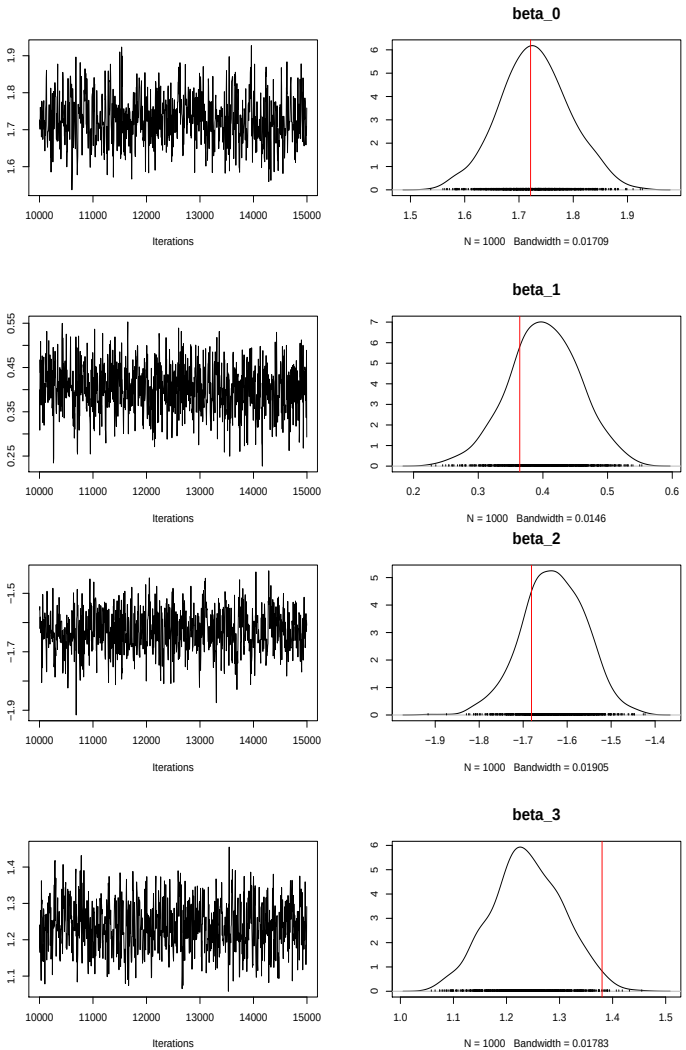
We use conjugate priors in a Gibbs sampler to estimate posterior distribution of parameters with the following proposition :

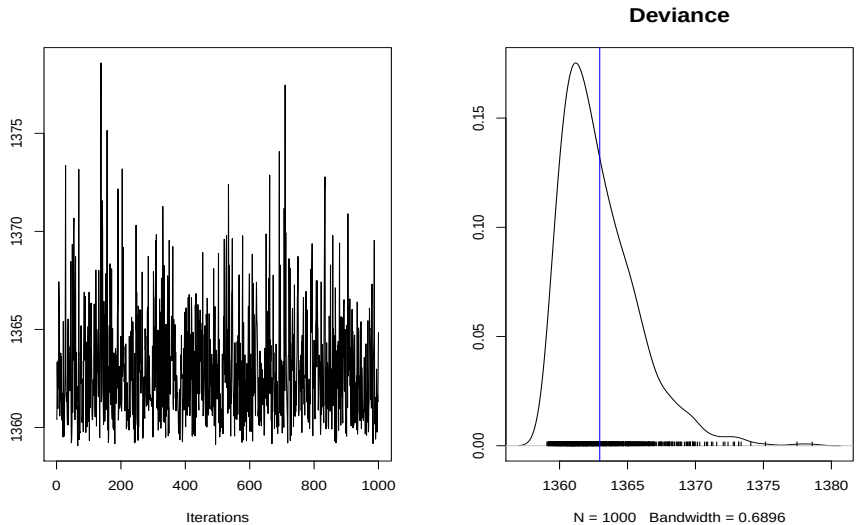
Proposition 1

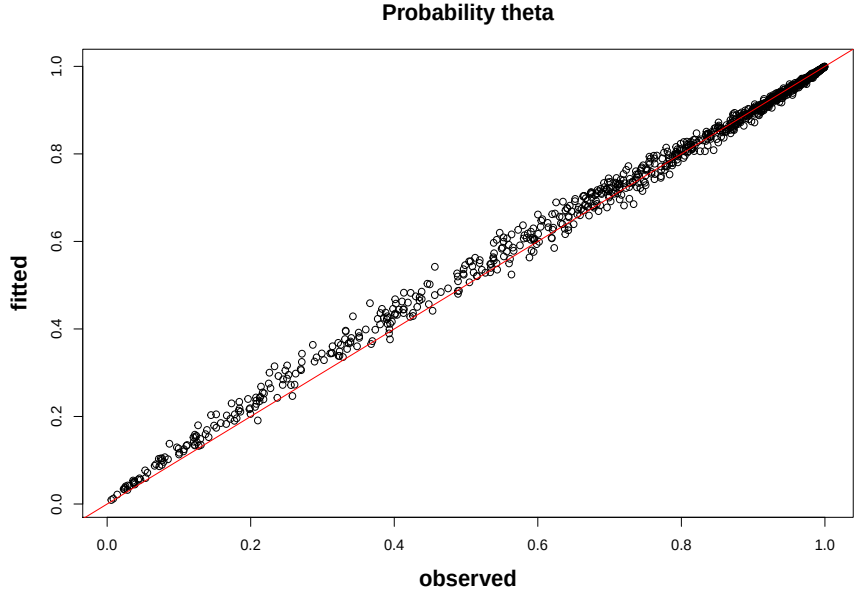
$$\begin{cases} Y \mid \beta & \sim \mathcal{N}_n(X\beta, I_n) \\ \beta & \sim \mathcal{N}_p(m, V) \end{cases} \Rightarrow \begin{cases} \beta|Y & \sim \mathcal{N}_p(m^*, V^*) \text{ with} \\ m^* & = (V^{-1} + X'X)^{-1}(V^{-1}m + X'Y) \\ V^* & = (V^{-1} + X'X)^{-1} \end{cases}$$

.

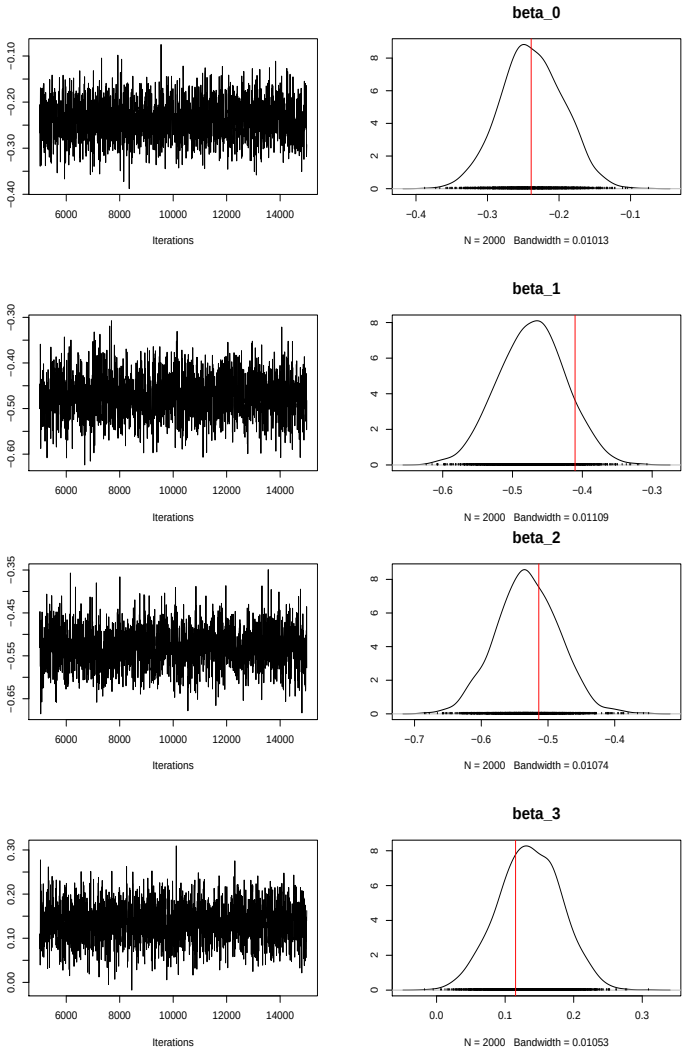
Model 1 : hSDM with logit link function

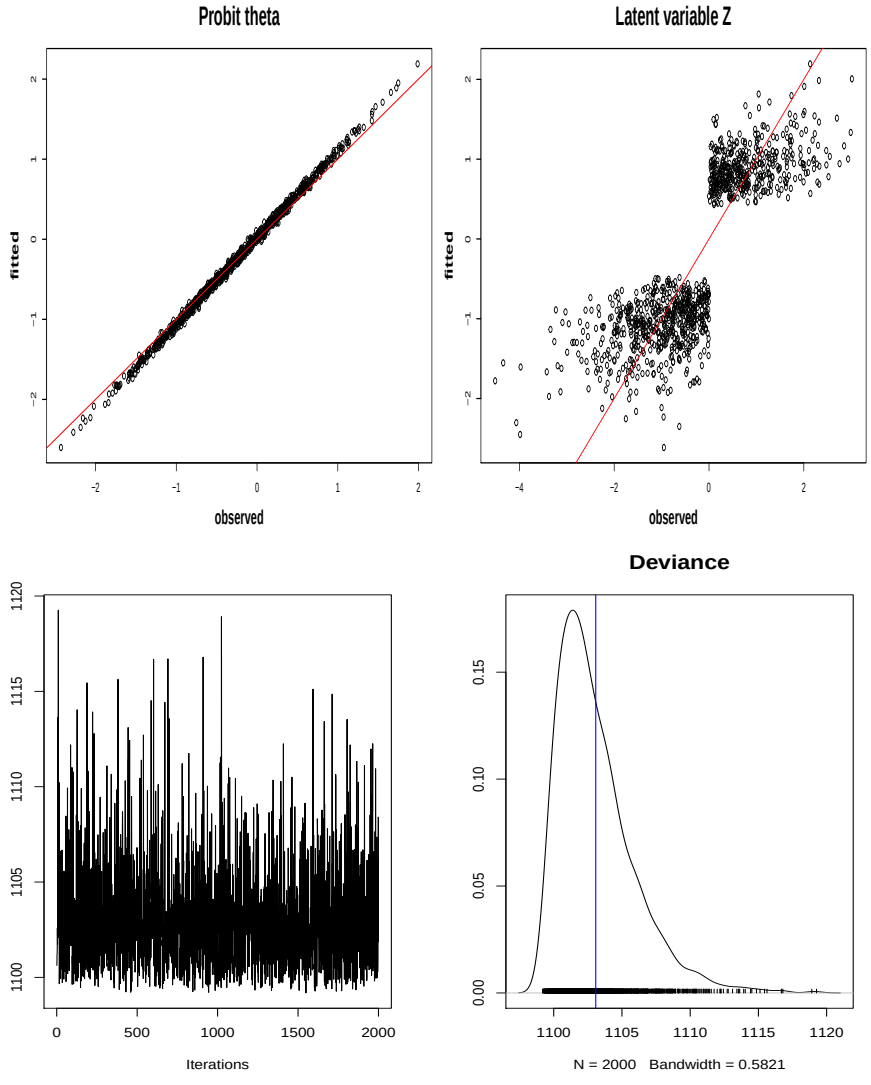






Model 2 : hSDM with probit link function





Compilation time of Gibbs sampler in R or in C++ with GSL and R random draws to fit model 1

Compilation times in secondes

	GSL draws	R draws	R function
Compilation time in secondes	57.4	56.7	4372.4

GSL radom draws in Rcpp

```
1 double x_prop = x_now + gsl_ran_gaussian_ziggurat(r, ↩
      sigmap_beta(p));
2 double u = gsl_rng_uniform(r);
```

R random draws in Rcpp

```
1 double x_prop = x_now + R::rnorm(0, sigmap_beta(p));
2 double u = R::runif(0,1);
```

Plan

cirad

botAnique et Modélisation de l'Architecture des Plantes et des végétations

Definition of the joint Species Distribition Model used :

$$\text{probit}(\theta_{ij}) = \alpha_i + \beta_{0j} + X_i\beta_j + W_i'\lambda_j$$

- Link function probit : $\text{probit} : q \to \Phi^{-1}(q)$ where Φ correspond to the $\mathcal{N}(0,1)$ cumulative distribution function.
- Response variable : $Y = (y_{ij})_{j=1,\dots,nspecies}^{i=1,\dots,nsite}$ with :

$$y_{ij} = \begin{cases} 0 & \text{if species } j \text{ is absent on the site } i \\ 1 & \text{if species } j \text{ is present on the site } i. \end{cases}$$

- Latent variable $z_{ij} = \alpha_i + \beta_{0j} + X_i'\beta_j + W_i'\lambda_j + \epsilon_{ij}$, with $\epsilon_{ij} \sim \mathcal{N}(0,1)$ and such that :

$$y_{ij} = \begin{cases} 1 & \text{if } z_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

It can be easily shown that : $y_{ij} \sim \mathcal{B}ernouilli(\theta_{ij})$.

Data and parameters

- Explanatory variables : $X = (X_i)_{i=1,\dots,nsite}$ avec $X_i = (x_i^1, \dots, x_i^p) \in \mathbb{R}^p$ where p is the number of bioclimatic variables considered for each site i.
- Species fixed effect : β_{0j} and $\beta_j = (\beta_j^1, \dots, \beta_j^p)'$, the intercept and regression coefficients for each species j. We use a prior distribution $\mathcal{N}(0, 10^6)$ for all betas.
- Latent variables : $W_i = (W_i^1, \dots, W_i^q)$ where q is the number of latent variables considered, we assume that $W_i \sim \mathcal{N}(0, I_q)$.
- Species fixed effect on latent variables : $\lambda_j = (\lambda_j^1, \dots, \lambda_j^q)'$. We use a prior distribution $\mathcal{N}(0, 10)$ for all lambdas not concerned by constraints to 0 on upper diagonal and to strictly positive values on diagonal.
- Random site effect : $\alpha_i \sim \mathcal{N}(0, V_\alpha)$.
- Variance of random site effect : $V_\alpha \sim \mathcal{IG}(\text{shape} = 0.5, \text{rate} = 0.005)$.

Inference method

Gibbs sampler with conjugate prior to estimate conditional posterior distributions of parameters :

- jSDM's function jointly estimate betas and lambdas of each species by drawing them under a multivariate gaussian distribution configured according to proposition 1.
- boral's function separately estimate betas and lambdas by drawing each parameter under gaussian distribution.
- Estimation of random site effect variance using the following proposal :

$$\begin{cases} x \mid \sigma^2 & \sim \mathcal{N}_n(\theta,\ \sigma^2 I_n) \\ \sigma^2 & \sim \mathcal{IG}(a, b) \\ \theta & \text{connu} \end{cases} \Rightarrow \begin{cases} \sigma^2 | x \sim \mathcal{IG}(a', b') \text{ with} \\ a' = a + \frac{n}{2} \\ b' = \frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)^2 + b. \end{cases}$$

Compilation times and accuracy of estimated parameters

Compilation times in minuts on four real datasets and one simulated

	Simulated	Mosquito	Eucalypts	Frogs	Fungi
boral	96.9	5.8	17.2	1.2	38.0
jSDM	7.0	1.3	1.8	0.3	4.1

Root-Mean-Square Error (RMSE) for simulated data

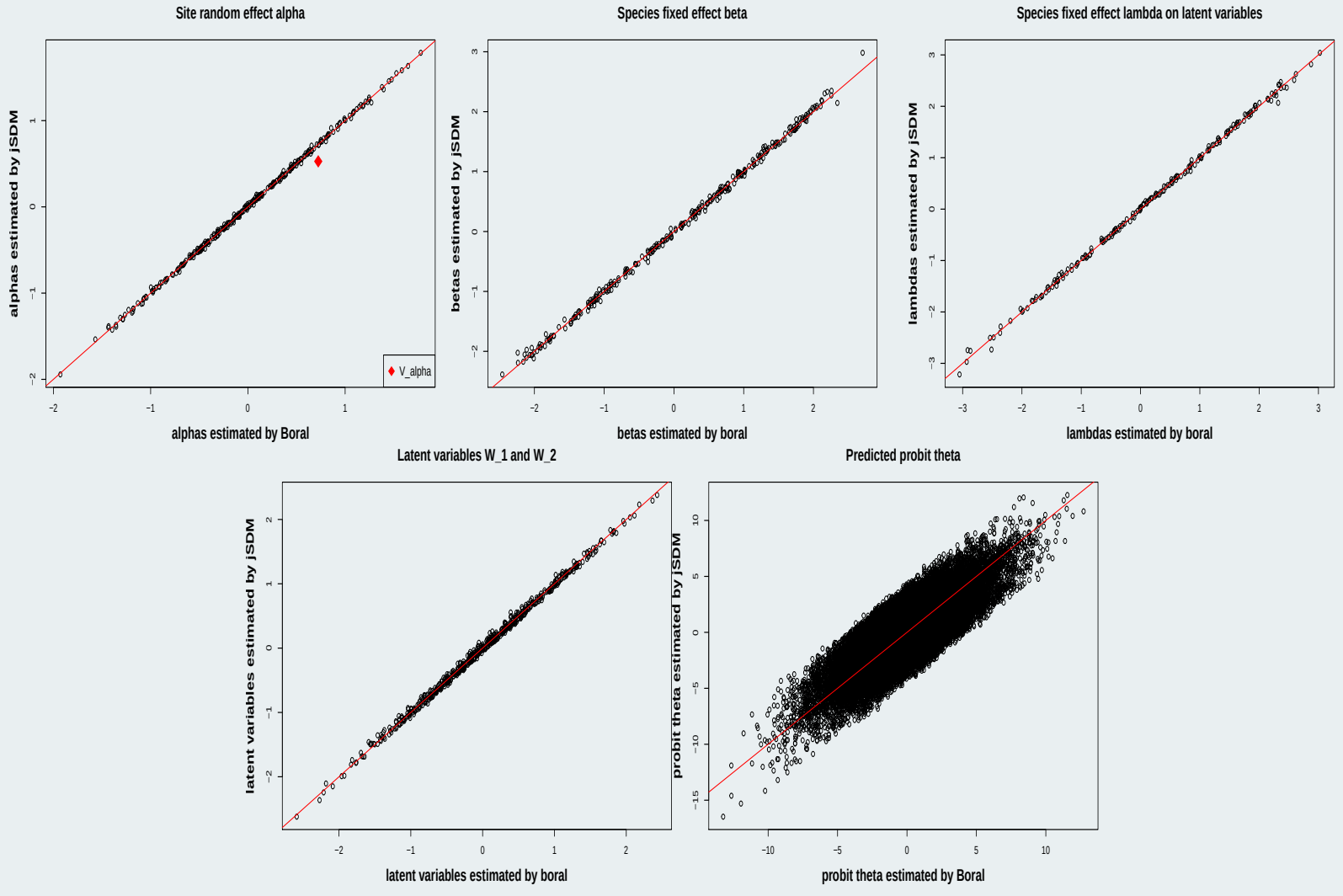
	boral	jSDM
RMSE	1.8	0.6

Obtained deviances with estimated parameters

	Simulated	Mosquito	Eucalypts	Frogs	Fungi
boral	40485.6	6935.8	8778.9	883.9	12870.6
jSDM	15651.0	1230.7	1921.9	150.2	1981.6

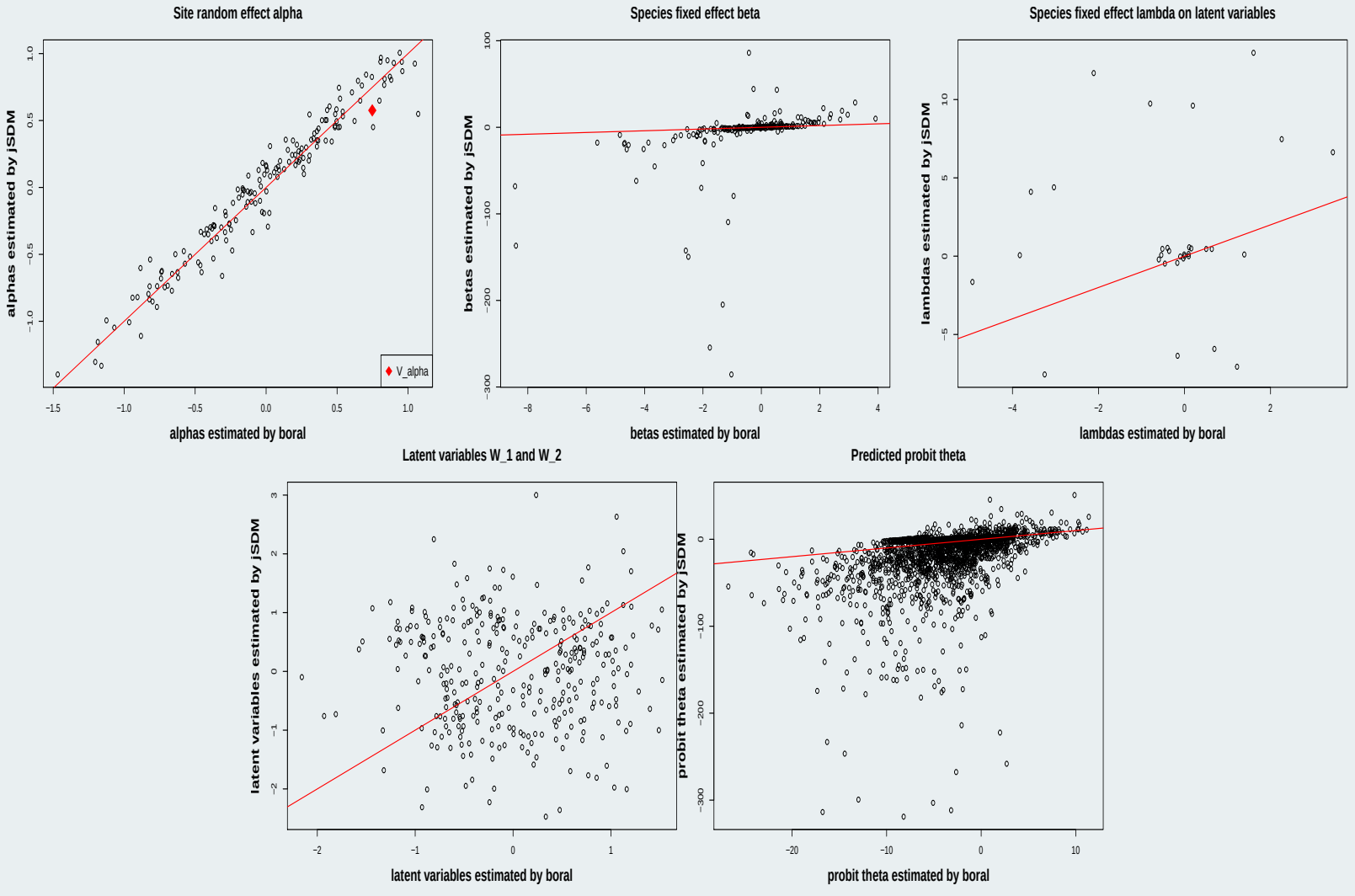
Comparison of estimated parameters by both packages

Simulated dataset :300 sites and 100 species



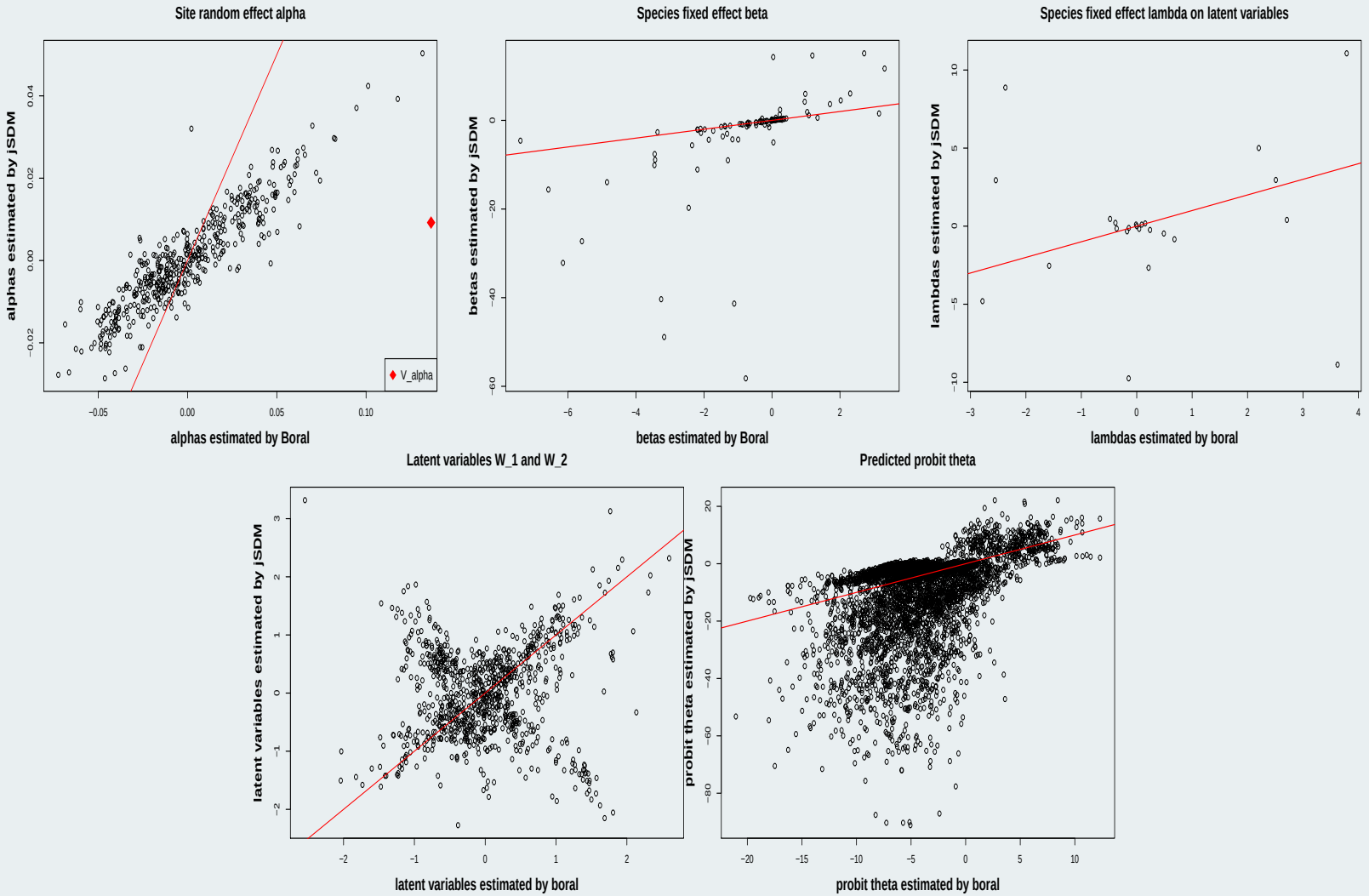
Comparison of estimated parameters by both packages

Mosquito dataset : 167 sites and 16 species



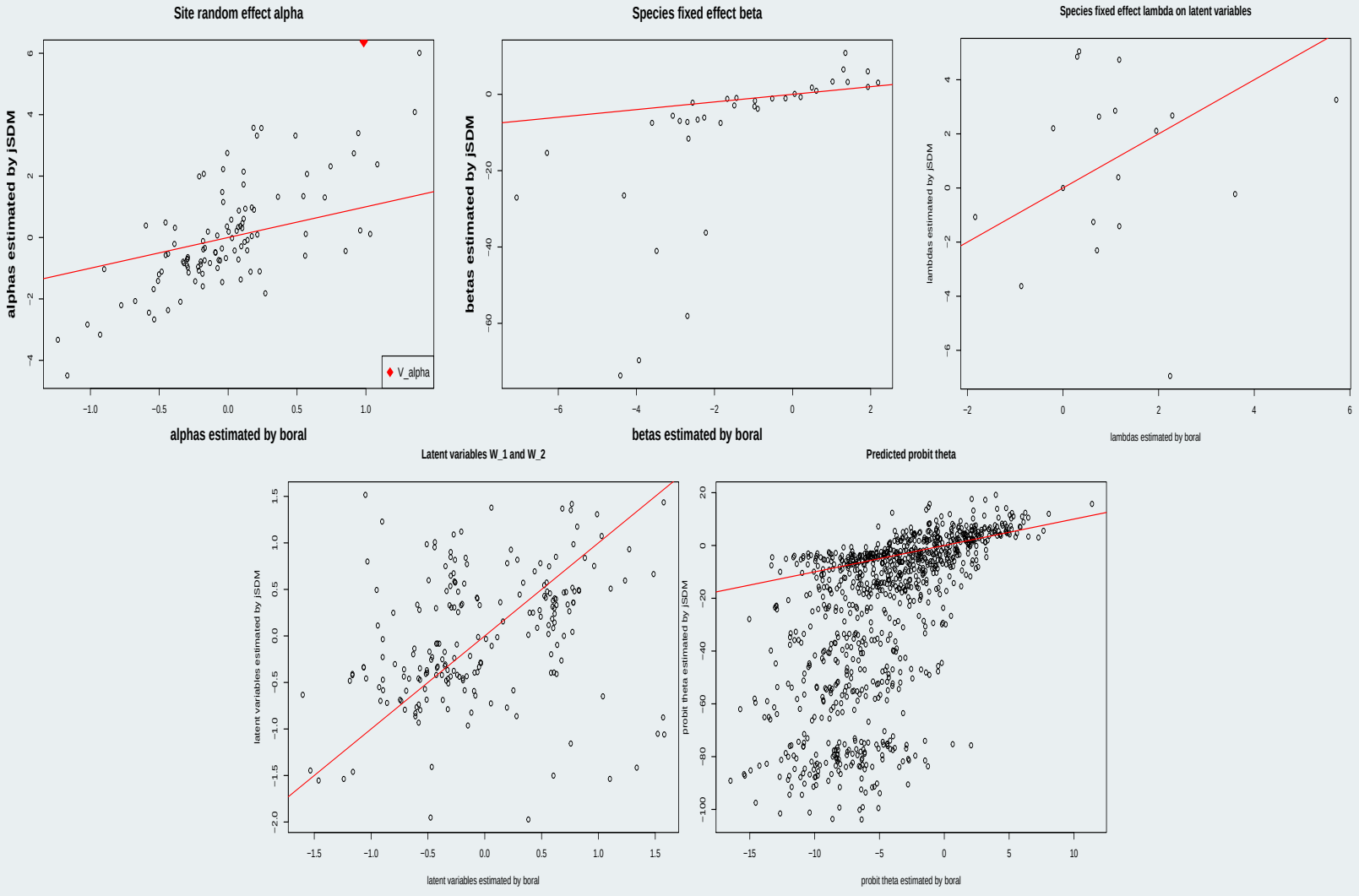
Comparison of estimated parameters by both packages

Eucalypts dataset : 458 sites and 12 species



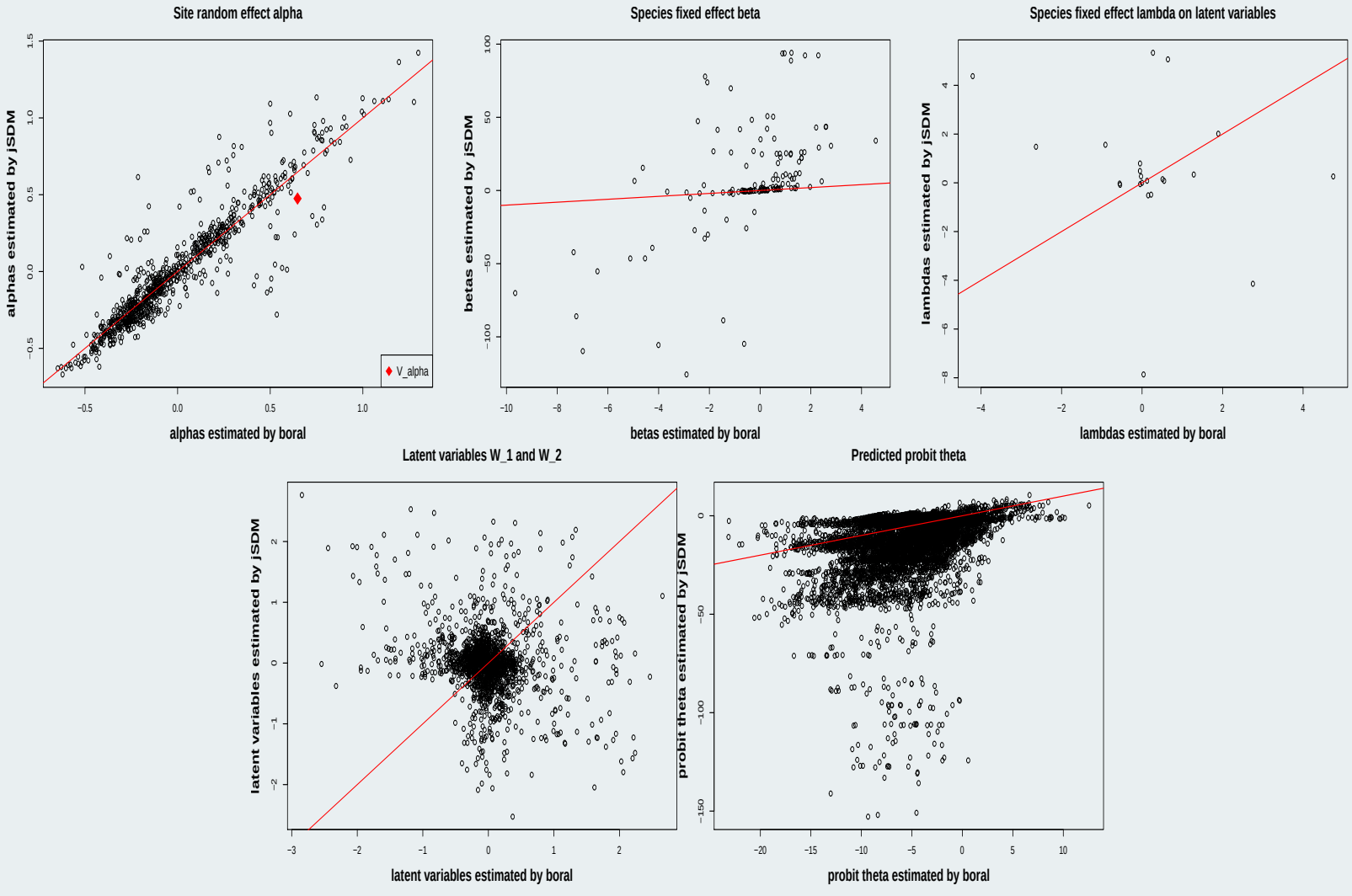
Comparison of estimated parameters by both packages

Frogs dataset : 104 sites and 9 species



Comparison of estimated parameters by both packages

Fungi dataset : 800 sites and 11 species



. . . Merci pour votre attention . . .

https://ecology.ghislainv.fr/jSDM

cirad

AMAP

botAnique et Modélisation de l'Architecture des Plantes et des végétations