Evaluating Markov chain Monte Carlo for phylogenetics

The case of exchangeable distributions

Luiz Max Carvalho

School of Applied Mathematics, Getulio Vargas Foundation

Available from: https://github.com/maxbiostat/presentations/

Plan for today

Phylogenetics

Concepts and problems

Exchangeable phylogenetic distributions

Proportional to distinguishable arrangements (PDA):
Coalescent, Yule.

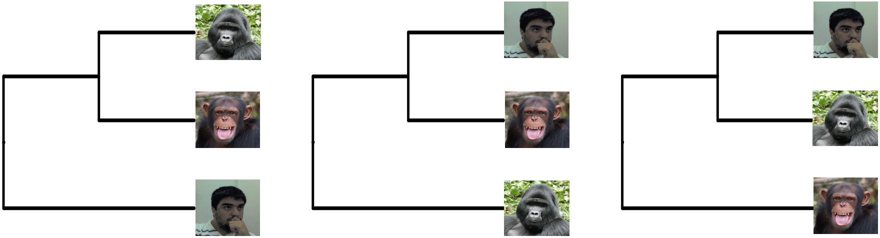
Multivariate ESS

Accounting for dependence between clades

Illustration

Simple illustration using lazy Metropolis-Hastings

Trees are hypotheses



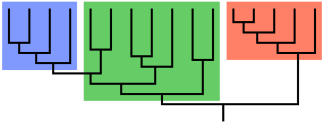
Target

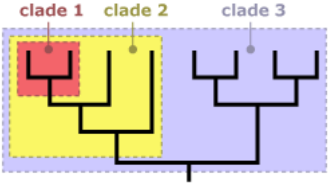
$$p(t, \boldsymbol{b}, \boldsymbol{\omega}|D) = \frac{f(D|t, \boldsymbol{b}, \boldsymbol{\omega})\pi(t, \boldsymbol{b}, \boldsymbol{\omega})}{\sum_{t_i \in \boldsymbol{T}_n} \int_{\boldsymbol{B}} \int_{\boldsymbol{\Omega}} f(D|t_i, \boldsymbol{b}_i, \boldsymbol{\omega})\pi(t_i, \boldsymbol{b}_i, \boldsymbol{\omega}) d\boldsymbol{\omega} d\boldsymbol{b}_i}. \quad (1)$$

- D: observed sequence (DNA) data;
- $\boldsymbol{T}_n$: set of all binary ranked trees ($\mathbb{G}^{(2n-3)!!}$);
- $\boldsymbol{b}_k$: set of branch lengths of $t_k \in \boldsymbol{T}_n$ ($\mathbb{R}_+^{2n-2}$, kind of) ;
- $\boldsymbol{\omega}$: set of parameters of interest such as substitution model parameters, migration rates, heritability coefficients, etc.

A lower-dimensional projection

A clade is a partition of the set of leaves and two clades $A = A_1|A_2$ and $B = B_1|B_2$ are said to be compatible if at least one of $A_i \cap B_j$, $i, j = 1, 2$ is empty. Here's a picture[1]:





[1] Pictures taken from Wikipedia and from https://evolution.berkeley.edu/evolibrary/news/080301_elephantshrew

Why clades?

- **Dimension!** $|\mathbb{T}_n| = (2n-3)!!$ *vs* $|\mathbb{C}_n| = 2^{n-1} - 1$
- Interpretability;
- Under simplifying assumptions, clades are independent (Larget, 2013[2]);
- Clade distribution is known under popular prior distributions.

[2] but see Whidden & Matsen, 2015 and Zang & Matsen, 2018.

Exchangeable distributions on $\mathbb{T}_n$

Aldous, 1996 proposes the following desiderata for a probability model on $\mathbb{T}_n$:

1) **Exchangeability:** For each n the random cladogram $t \in \mathbb{T}_n$ is exchangeable in the species' labels, i.e., invariant under permutations;
2) **Group elimination:** For all $k \in [1, n-1]$, if we condition on a clade with members $\{k+1, k+2, \ldots, n\}$, the remaining cladogram on $\{1, 2, \ldots, k\}$ has the same distribution indexed on $\mathbb{T}_k$.

This gives rise to the β-splitting family of distributions.

The Beta-splitting model

For $n \geq 2$, we have a symmetric distribution $\boldsymbol{q}_n = (q_n(i); i = 1, 2, \ldots, n-1)$, with $q_n(i) = q_n(n-i)$.

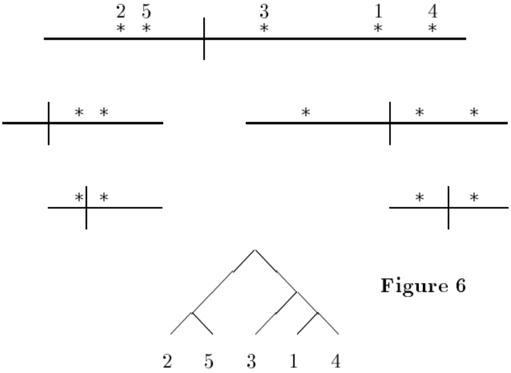


Figure: Figure 6 in Aldous (1996): the beta-splitting model.

The Beta-splitting model (cont.)

Take

$$q_n(i) = \frac{\binom{n}{i} \int_0^1 x^i (1-x)^{n-i} f(x)\, dx}{1 - 2\int_0^1 x^n f(x)\, dx},$$

with

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} x^\beta (1-x)^\beta, x \in (0,1).$$

For $\beta \in (-2, -1]$, this encompasses classical models:

- $\beta = 0$: Yule model;
- $\beta = -3/2$: uniform on $\mathbb{T}_n$.

Properties of PDA models

Zhu, Degnan & Steel (2011) show that:

Theorem (Joint distribution of clades)

Let A and B be two clades with $|A| = a$ and $|B| = b$. Under a PDA model, the joint probability of A and B is

$$p_n(A,B) = \begin{cases} p_n(a), \text{ if } A \equiv B; \\ R_n(a,b), \text{ if } A \subsetneq B; \\ R_n(b,a), \text{ if } B \subsetneq A; \\ \bar{p}(a, n-a), \text{ if } A \cap B = \emptyset \text{ and } A \cap B = \mathcal{X}; \\ r_n(a,b), \text{ if } A \cap B = \emptyset \text{ and } A \cap B \subsetneq \mathcal{X}; \\ 0, \text{ otherwise}, \end{cases} \quad (2)$$

Properties of PDA models (cont.)

where

$$p_n(a) := \begin{cases} \frac{2n}{a(a+1)}\binom{n}{a}^{-1}, \text{ if } & 1 \leq a \leq n-1; \\ 0, \text{ otherwise}, \end{cases} \qquad ,$$

$$\bar{p}_n(a,b) := \frac{4a!b!(n-a-b))!}{(n-1)!(a+b)([a+b]^2-1)!},$$

$$R_n(a,b) := \frac{4n}{a(a+1)(b+1)}\binom{n}{b}^{-1}\binom{b}{a}^{-1},$$

$$r_n(a,b) := \frac{4a!b!(n-a-b))!}{(n-1)!}G_n(a,b), \text{ with}$$

$$G_n(a,b) := \frac{n}{ab(a+1)(b+1)} - \frac{a(a+1)+b(b+1)+ab}{ab(a+1)(b+1)(a+b+1)} + \frac{1}{(a+b)[(a+b)^2-1]}.$$

Clade correlations

$$\rho_n(A,B) = \frac{p_n(A,B) - p_n(A)p_n(B)}{\sqrt{p_n(A)[1-p_n(A)]p_n(B)[1-p_n(B)]}}.$$

Theorem (**Minimum and maximum correlation**)

For $n \geq 4$, the minimum and maximum values for $\rho_n(A,B)$ are, respectively

$$\rho_{\min}(n) = -\frac{2}{3n-5},$$

$$\rho_{\max}(n) = \frac{2u(n)k(n) - 4n^2(n-1)}{2n(n-1)\sqrt{\left[\lfloor\frac{n}{2}\rfloor\left(\lfloor\frac{n}{2}\rfloor+1\right)k(n)-2n\right]\left[\lceil\frac{n}{2}\rceil\left(\lceil\frac{n}{2}\rceil+1\right)k(n)-2n\right]}},$$

Further observations on the clade correlation under PDA

Let $c(n)$ be the proportion of entries in the clade correlation matrix that are **positive**.

Theorem (Sparsity of exchangeable priors)

The following facts imply that the exchangeable PDA prior induces a "flat" correlation matrix as the number of taxa n grows:

i) $\lim_{n\to\infty} \rho_{\min}(n) = 0$;

ii) $\lim_{n\to\infty} c(n) = 0$.

Additionally, $\lim_{n\to\infty} \rho_{\max}(n) = 1/4$.

How can we put these things to good use?

For correcntess, we can check

a) Clade frequencies;

b) Clade correlations;

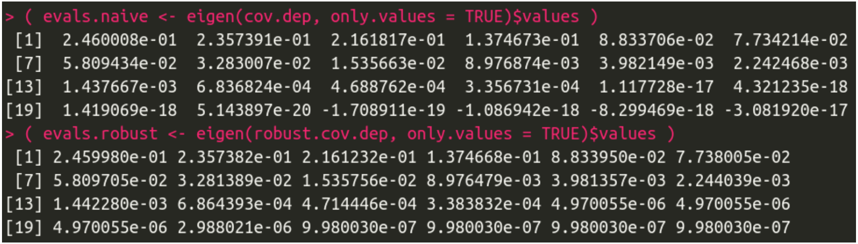
c) Minimum and maximum correlation;

As we shall see, we can use this approach to assess correctness and efficiency **simultaneously**!

Measuring efficiency

Thus, we can employ the idea from Vats, Flegal & Jones (2019): Magee et al, 2021 point out that trees are fundamentally multivariate objects.

$$\text{mESS} = M \left(\frac{\det(\mathbf{\Lambda})}{\det(\mathbf{\Sigma})} \right)^{1/p}. \tag{3}$$



```
> ( evals.naive <- eigen(cov.dep, only.values = TRUE)$values )
 [1]  2.460008e-01  2.357391e-01  2.161817e-01  1.374673e-01  8.833706e-02  7.734214e-02
 [7]  5.809434e-02  3.283007e-02  1.535663e-02  8.976874e-03  3.982149e-03  2.242468e-03
[13]  1.437667e-03  6.836824e-04  4.688762e-04  3.356731e-04  1.117728e-17  4.321235e-18
[19]  1.419069e-18  5.143897e-20 -1.708911e-19 -1.086942e-18 -8.299469e-18 -3.081920e-17
> ( evals.robust <- eigen(robust.cov.dep, only.values = TRUE)$values )
 [1] 2.459980e-01 2.357382e-01 2.161232e-01 1.374668e-01 8.833950e-02 7.738005e-02
 [7] 5.809705e-02 3.281389e-02 1.535756e-02 8.976479e-03 3.981357e-03 2.244039e-03
[13] 1.442280e-03 6.864393e-04 4.714446e-04 3.383832e-04 4.970055e-06 4.970055e-06
[19] 4.970055e-06 2.988021e-06 9.980030e-07 9.980030e-07 9.980030e-07 9.980030e-07
```

Figure: Eigenvalues can be numerically unstable.

Simple Metropolis-Hastings on the SPR graph

For $T \in \mathbb{T}\mathbb{m}$ let $N(T)$ be the set of all trees $u \in \mathbb{T}_n$ which are on subtree prune-and-regraft operation away from T.

Define $a(x) := 1 - \sum_{z \in N(x)} \frac{1}{|N(x)|} \min\left\{1, \frac{|N(x)|}{|N(z)|}\right\}$.

$$p_{\mathrm{MH}}(x, y) = \begin{cases} \frac{1}{|N(x)|} \min\left\{1, \frac{|N(x)|}{|N(y)|}\right\}, y \in N(x), \\ a(x), y = x \\ 0, y \notin N(x). \end{cases}$$

Lazy Metropolis-Hastings

We can (artificially) change the performance of the original MH by adding a probability $\rho \in (0,1)$ of staying in the same place. Then

$$p_{\mathrm{LMH}}(x,y) = \begin{cases} p_{\mathrm{MH}}(x,y), y \in N(x) \,\&\, a(x) = 0, \\ 0, y = x \,\&\, a(x) = 0, \\ \frac{1-\rho}{1-a(x)} p_{\mathrm{MH}}(x,y), y \in N(x) \,\&\, a(x) > 0, \\ \rho, y = x \,\&\, a(x) > 0, \\ 0, y \notin N(x). \end{cases}$$

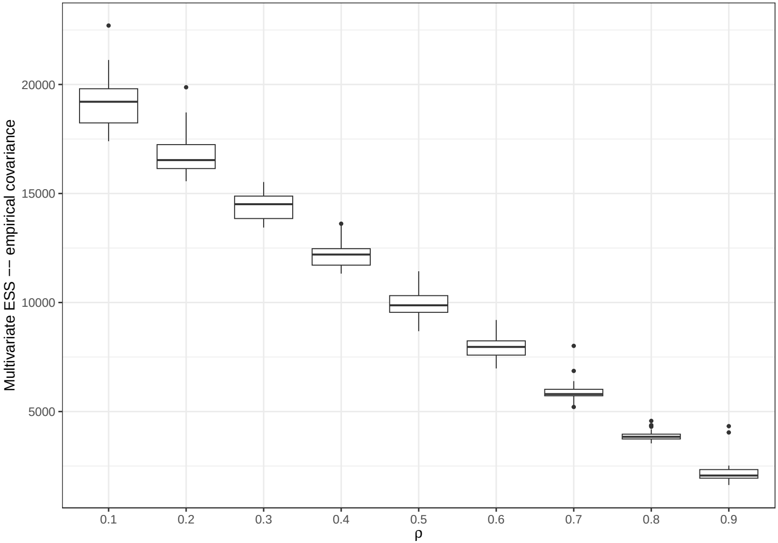
A small illustration

For $n = 5$ and $\rho \in \{0.1, 0.2, \ldots, 0.9\}$, run $K = 50$ replicates of $M = 10,000$ iterations each. Then project onto clade space and compute

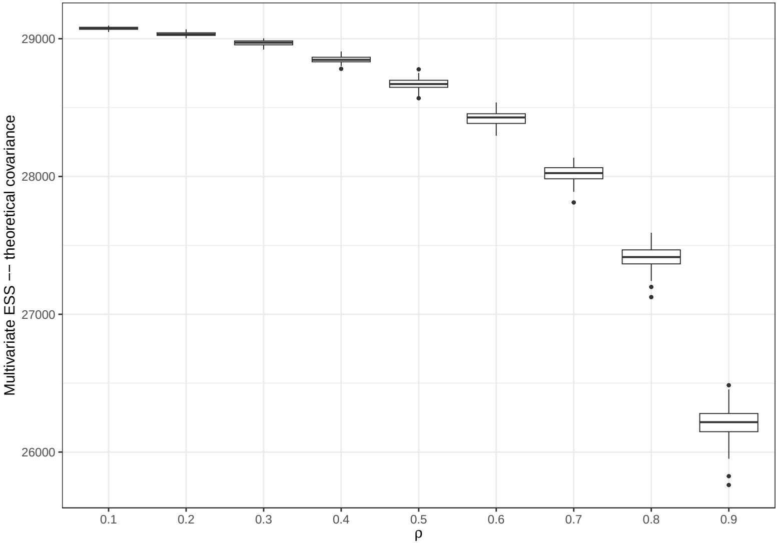
i) **empirical**: the multivariate ESS with both $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}$ estimated from the data;

ii) **theoretical**: the multivariate ESS with $\boldsymbol{\Sigma}$ set to its theoretical value.

Results I


Multivariate ESS -- empirical covariance
20000
15000
10000
5000
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
ρ

Results II



Multivariate ESS –– theoretical covariance
29000
28000
27000
26000
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
ρ

Take home

Trees are weird

We need more (better!) theory for the space of phylogenies

Lower dimensional projections can help

Projecting onto clades

A framework for jointly assessing correctness and efficienty of MCMC

These ideas can be employed for real-world MCMC samplers such as the ones in BEAST and Mr. Bayes

THE
END