Markov Chain Monte Carlo for phylogenetics

a helicopter ride

Luiz Max Carvalho [lmax.fgv@gmail.com]

School of Applied Mathematics
Getúlio Vargas Foundation, Rio de Janeiro.
COLMEA November 2022

Acknowledgments

Andrew Rambaut
UoE

Marc Suchard
UCLA

Guy Baele
KU Leuven

Rodrigo B. Alves
FGV EMAp

Remco Bouckaert
Auckland

Plan for today

Problem

What are trees and why are interested in them?

MCMC in tree space

A journey through a strange land

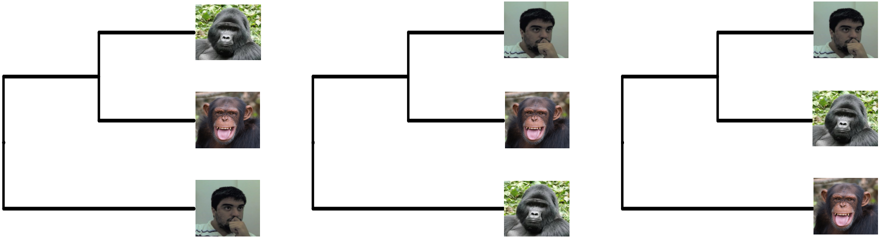
Validation

Checking against exchangeable phylogenetic distributions and simulation-based calibration (SBC).

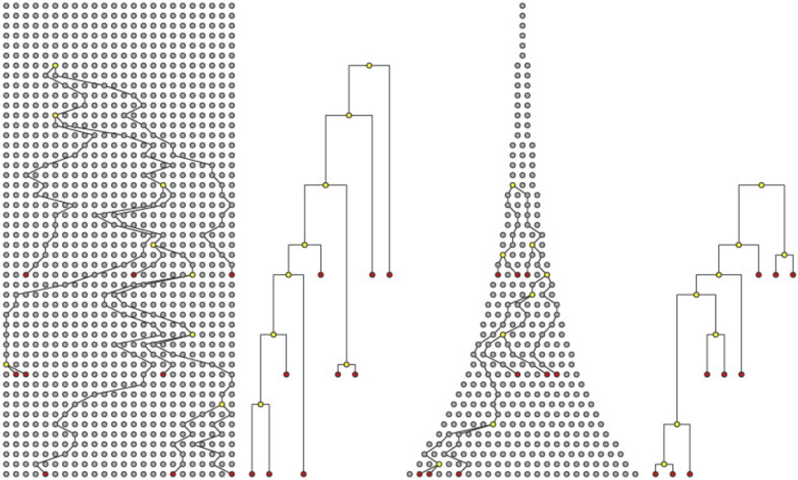
Perspectives

Open problems!

Trees are hypotheses



Trees and the coalescent



Central object: time-calibrated trees

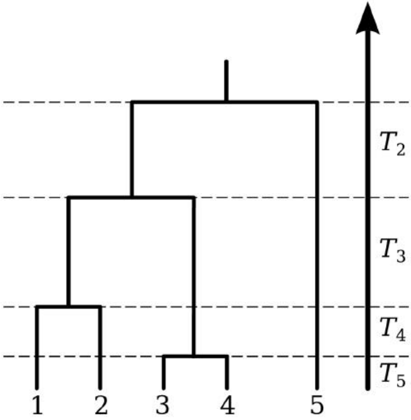


Figure: Figure 4 from Volz et al. (2013).

Let T_n denote the time for n lineages to *coalesce*, i.e., merge into one ancestral lineage, in a population of size N_e. Then:

$$Pr(T_n = t) = \lambda_n e^{-\lambda_n t}$$

$$\lambda_n = \binom{n}{2}\frac{1}{N_e} = \binom{n}{2}\frac{1}{N_e \tau}$$

where N_e is the effective population size and τ is the generation time. Let T_{mrca} denote the age of the most recent common ancestor:

$$\begin{aligned}\mathbb{E}[T_{\text{mrca}}] &= \mathbb{E}[T_n] + \mathbb{E}[T_{n-1}] + \ldots + \mathbb{E}[T_2] \\ &= 1/\lambda_n + 1/\lambda_{n-1} + \ldots + 1/\lambda_2 \\ &= 2N_e\left(1 - \frac{1}{n}\right)\end{aligned}$$

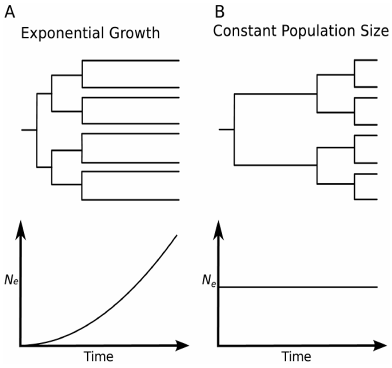
Motivation

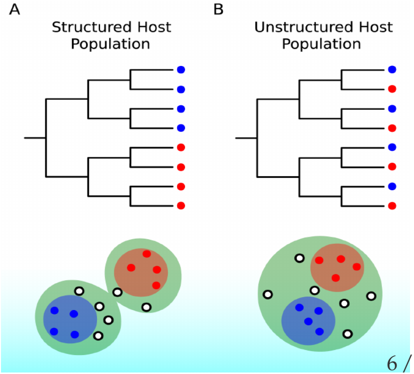
Phylodynamics of fast-evolving viruses

Inferring spatial and temporal dynamics from genomic data:

Phylogenies*!

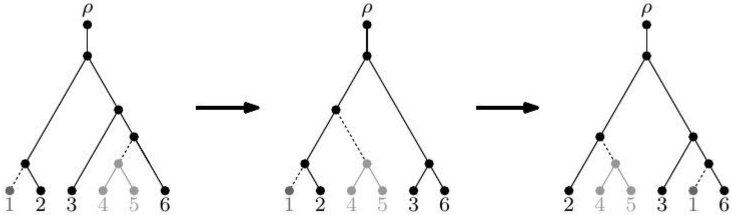
* plus complicated models





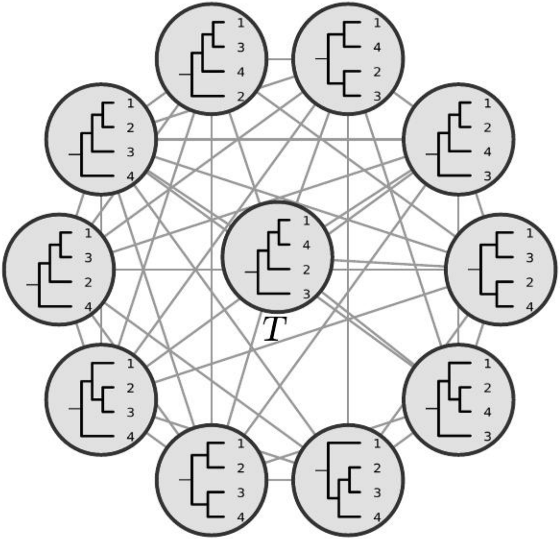
Discrete tree space: tree surgery

Subtree prune-and-regraft (SPR):



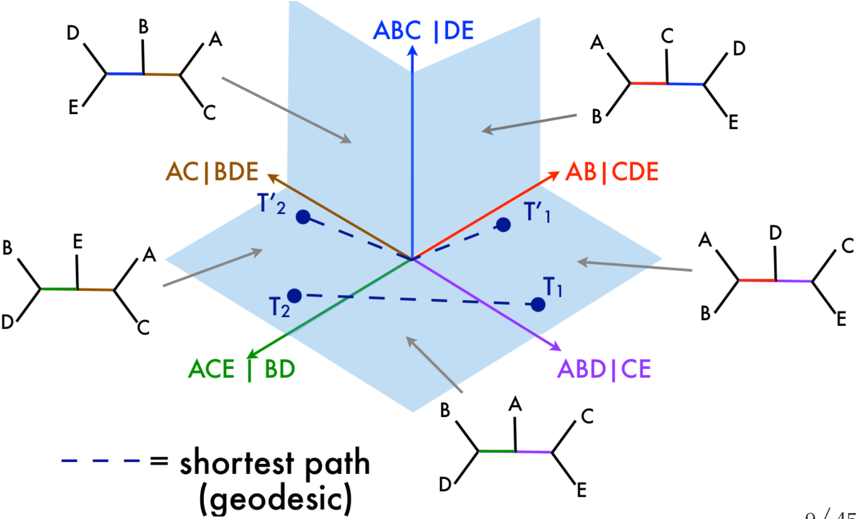
Discrete tree space: SPR graph

For curvature results, see Whidden & Matsen(2017).

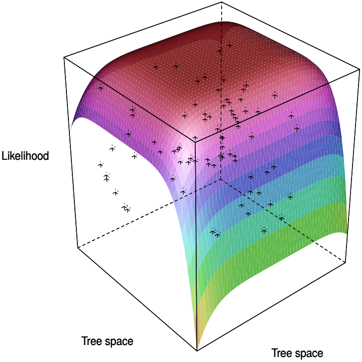


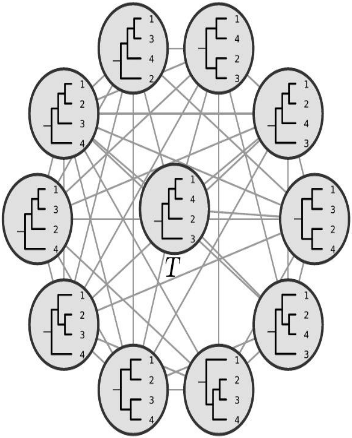
Continuous tree space: BHV

Billera, Holmes & Vogtmann (2001).



Tree space: a strange land





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.

(Adaptive) Metropolis-Hastings for trees

General MH setup.

Let $\tau = (t, \boldsymbol{b})$ denote a tree with topology t and branch lengths $\boldsymbol{b}$. For two trees τ and τ', denote the transition kernel by $q_\gamma(\tau|\tau') := \Pr(\tau' \to \tau|\gamma)$.

Accepting with probability

$$A_\gamma(\tau|\tau') = \min\left(1, \frac{p(\tau', \boldsymbol{\omega}|D)q_\gamma(\tau|\tau')}{p(\tau, \boldsymbol{\omega}|D)q_\gamma(\tau'|\tau)}\right)$$

leads to the desired target.

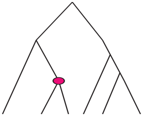
Note: Here $\gamma > 0$ is a so-called tuning parameter.

Height-constrained kernels: SubTreeLeap (STL)

1. Excluding the root, pick a node i in τ uniformly at random, i.e., with probability $1/(2n-3)$;
2. Draw a patristic distance δ from the distance kernel $k(\delta|\sigma)$;
3. Find the set of destination nodes ${\mathbf{D_i}}^{\delta}$ that are within distance δ **and** whose heights are not less than $h(i)-\delta$;
   If ${\mathbf{D_i}}^{\delta} = :$
   - prune p_i and regraft it at height $h_b = h(p_i) - \delta$ or $h_a = h(p_i) + \delta$ with probability $1/2$, creating a new tree τ', else
   - pick a node $j \in {\mathbf{D_i}}^{\delta}$ with probability $Pr(i \rightarrow j) = 1/|{\mathbf{D_i}}^{\delta}|$, prune the tree at p_i and regraft it at p_j, creating a new tree τ';

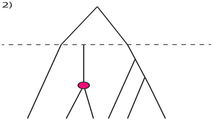
STL – illustration

1)



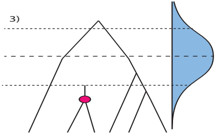
Pick a node

2)



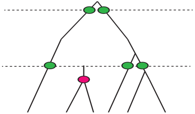
Disconnect its parent

3)



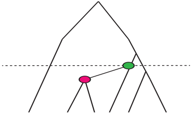
Draw a new height from a
normal centred on old height
of parent. Also consider the
symmetrical height above or
below the old height.

4)



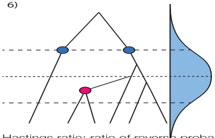
Pick uniformly from branches
subtending that height and the
symmetrical height above or
below (in this case 5).

5)



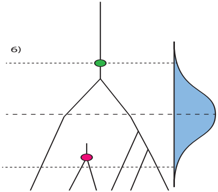
Attach parent to the chosen
location.

6)



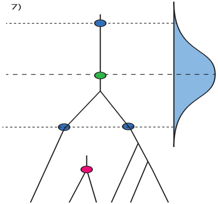
Hastings ratio: ratio of reverse probability
(1 / number of reverse locations, i.e., 1/2) to
forwards probability (i.e., 1/5).
Hastings ratio = 5 / 2

6)



There is always at least 1 target
location (above the root).

7)



In this case the HR would be 1/3

STL – properties

- Adaptive → more efficient (?);
- Height-constrained → time-precedence constraints are respected;
- Changes topology and branch lengths **simultaneously** → presumably more efficient;
- Inherits cool properties from SPR.
  - We know a bunch of things about the SPR graph;
  - SPR graph admits a Hamiltonian (Gordon et al., 2013).

STL – ergodicity

Carvalho (2019), Chapter 2.

Remark

Assume strictly positive branch lengths. Then SubTreeLeap induces an irreducible Markov chain on $\mathbb{G}$.

Sketch: Starting at $x \in \mathbb{G}$, notice there exists $\delta_y^\star > 0$ such that $P\left(x \to y \mid \delta_y^\star\right) > 0$ for any tree $y \in \mathbb{G}$ in the SPR neighbourhood of x.

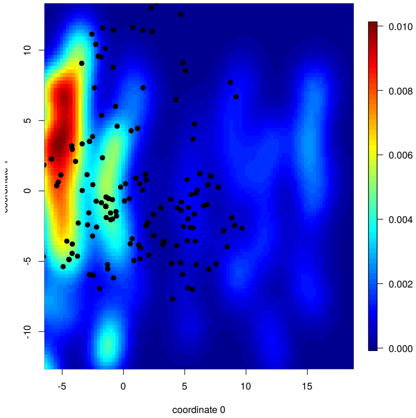
Theorem

Assume the target satisfies $p(A) > 0$ *for all* $A \subset \mathbf{\Psi}$. *Then, SubTreeLeap induces an ergodic Markov chain on* $\mathbf{\Psi}$.

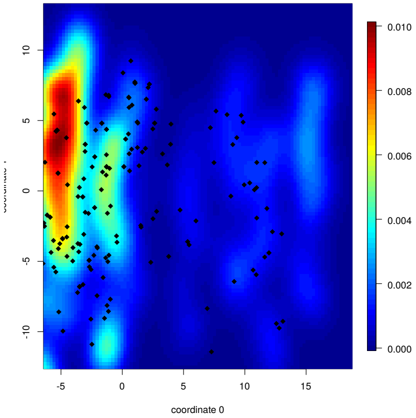
Sketch: Employ the remark to get to the case where $d_{\text{SPR}}(x, y) = 0$ and then establish Harris recurrence.

Traversing tree space – Topology

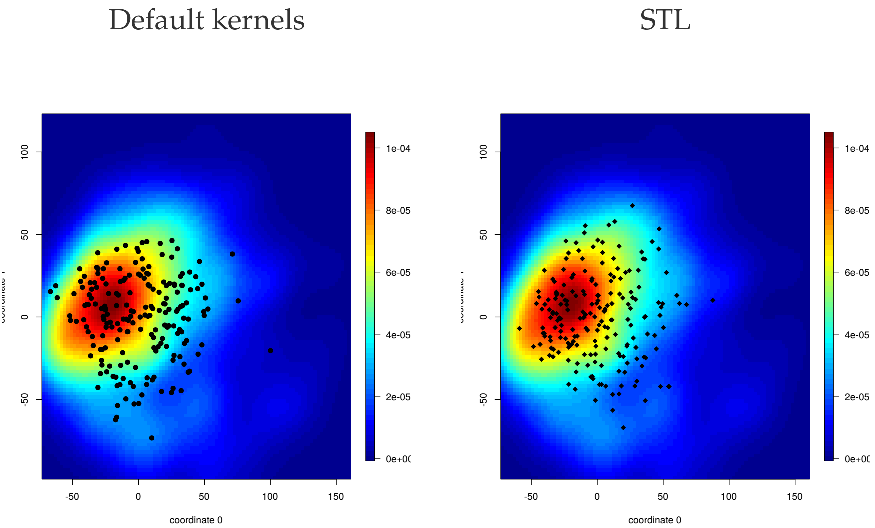
Default kernels



STL



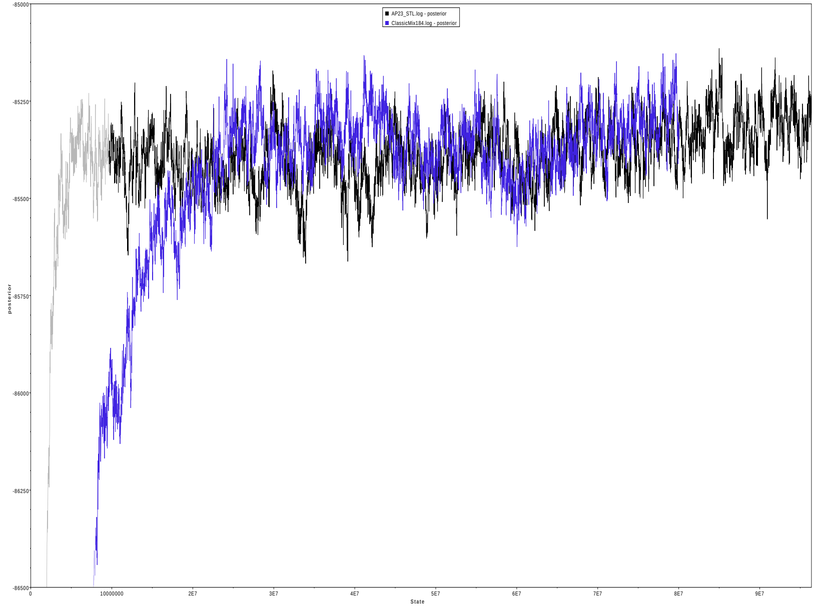
Traversing tree space – Topology + branch lengths



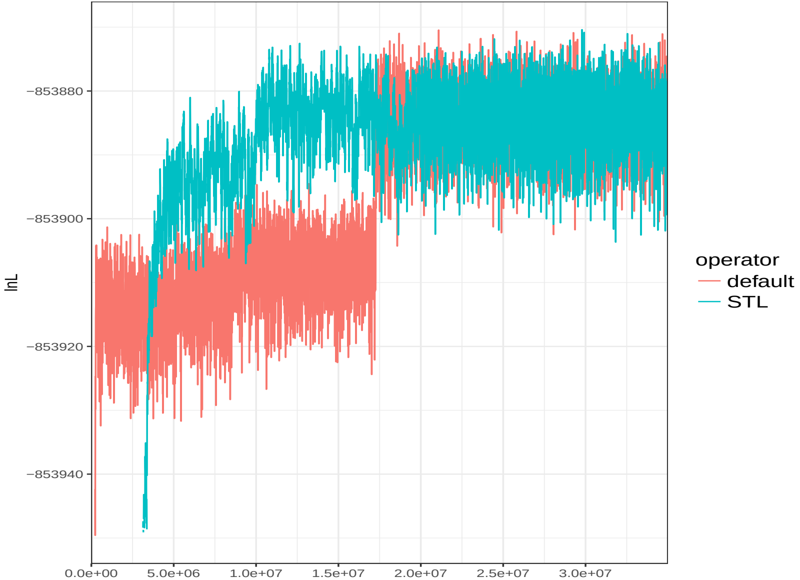
Default kernels

STL

Ebola virus full genome (1610 taxa (!), 18990 NT sites)

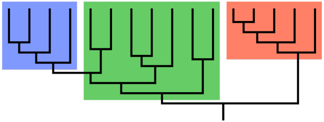


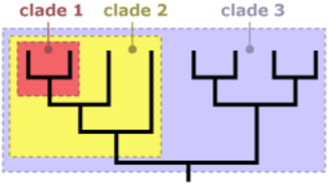
Metazoans (contemporaneous, 55 taxa, 30257 AA sites)



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?

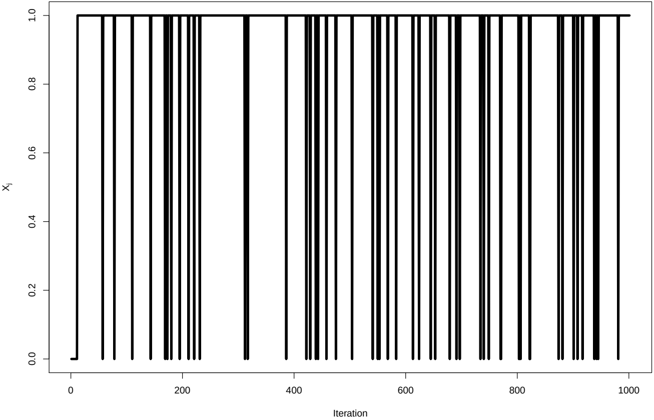
- **Dimension**: $|\mathbb{T}_n| = (2n-3)!!$ *vs* $|\mathbb{C}_n| = 2^{n-1} - 1$
- Interpretability;
- Under simplifying assumptions, clades are independent (Larget, 2013[2]);
- Clade distribution is known under popular prior distributions.

---

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

Setup

Let $X_j^{(i)} \in \{0, 1\}$ be the indicator of whether clade j in the tree sampled at the i-th iteration and $\hat{p}_j = M^{-1} \sum_{i=1}^{M} X_j^{(i)}$ be a simple MCMC estimator of its marginal success probability.



Playing pretend

Theorem

The Metropolis-Hastings process (with uniform invariant) on the SPR graph is ϵ-lumpable w.r.t. clades.

Pretend for a second $\left(X_j^{(i)}\right)_{i\geq 0}$ is Markov on $\mathcal{X} = \{0, 1\}$ and reparametrise the usual two-state model as

$$\tilde{\boldsymbol{P}}_x := \begin{bmatrix} 1-\alpha & \alpha \\ \alpha\frac{1-p}{p} & \frac{p-\alpha(1-p)}{p} \end{bmatrix}, \tag{2}$$

What an explicit model buys you

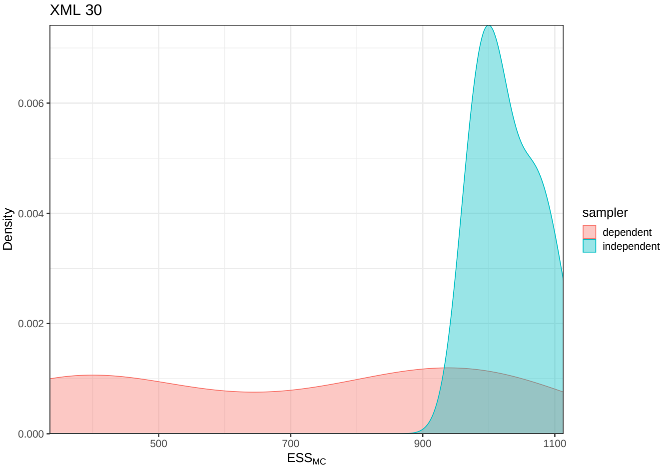
Under this model we can derive

- Distribution of occupation times;
- Distribution of state-transitions (0 → 1 or 1 → 0);
- Effective sample size:

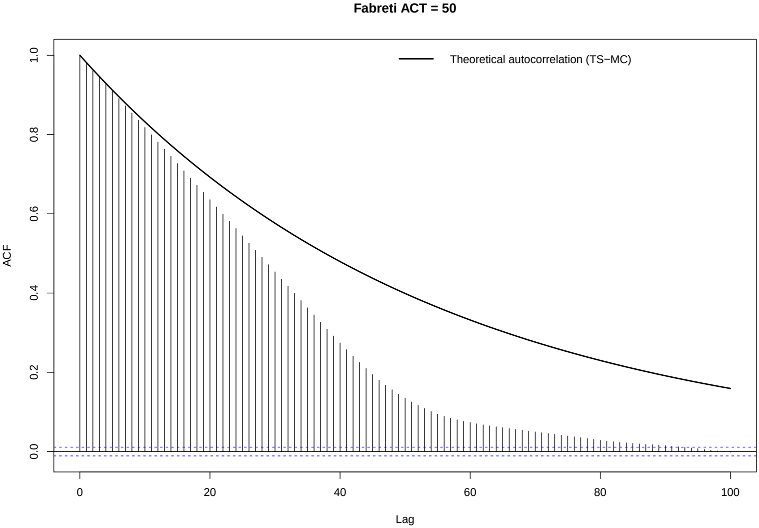
$$\begin{aligned}
\text{ESS} &= \frac{M}{1 + 2\sum_{t=1}^{\infty} \rho_t}, \\
&= \frac{M}{1 + 2\frac{p-\alpha}{\alpha}}, \\
&= \frac{\alpha}{2p - \alpha} M.
\end{aligned} \tag{3}$$

Looking cool!

We can fake phylogenetic MCMC quite well. In particular we can sample from the posterior “exactly".



Autocorrelation spectra in practice



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

$$p_n(A, B) = \begin{cases} p_n(a), \text{ if } A \equiv B; \\ R_n(a, b), \text{ if } A \subsetneq B; \\ R_n(b, a), \text{ if } B \subsetneq A; \\ \bar{p}(a, n-a), \text{ if } A \cap B = \emptyset \text{ and } A \cap B = \mathfrak{X}; \\ r_n(a, b), \text{ if } A \cap B = \emptyset \text{ and } A \cap B \subsetneq \mathfrak{X}; \\ 0, \text{ otherwise}, \end{cases} \tag{4}$$

Properties of PDA models (cont.)

where

$$
\begin{aligned}
p_n(a) &:= \begin{cases} \frac{2n}{a(a+1)}\binom{n}{a}^{-1}, \text{ if } \quad 1 \le a \le n-1; \\ 0, \text{ otherwise}, \end{cases} \qquad\qquad , \\
\bar{p}_n(a,b) &:= \frac{4a!b!(n-a-b))!}{(n-1)!(a+b)([a+b]^2-1)!}, \\
R_n(a,b) &:= \frac{4n}{a(a+1)(b+1)}\binom{n}{b}^{-1}\binom{b}{a}^{-1}, \\
r_n(a,b) &:= \frac{4a!b!(n-a-b))!}{(n-1)!}G_n(a,b), \text{ with} \\
G_n(a,b) &:= \frac{n}{ab(a+1)(b+1)} \\
&\quad - \frac{a(a+1)+b(b+1)+ab}{ab(a+1)(b+1)(a+b+1)} \\
&\quad + \frac{1}{(a+b)[(a+b)^2-1]}.
\end{aligned}
$$

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

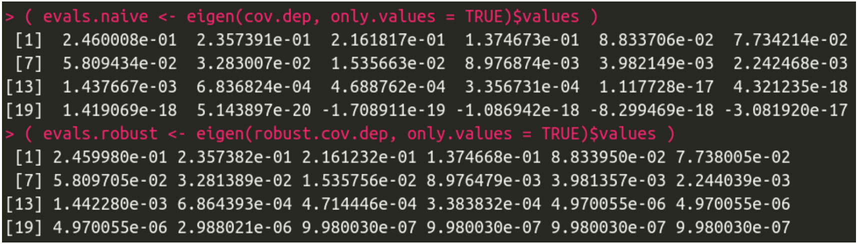
a) Clade frequencies;
b) Clade correlations;
c) Minimum and maximum correlation;

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.

$$\text{mESS} = M\left(\frac{\det(\boldsymbol{\Lambda})}{\det(\boldsymbol{\Sigma})}\right)^{1/p}. \tag{5}$$



```
> ( 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 $T \in \mathbb{T}\mathbb{n}$ 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

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

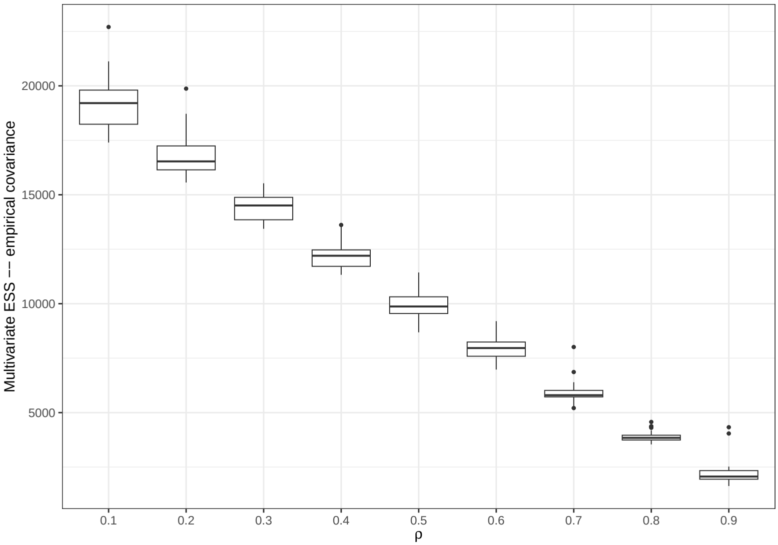
A small illustration

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

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

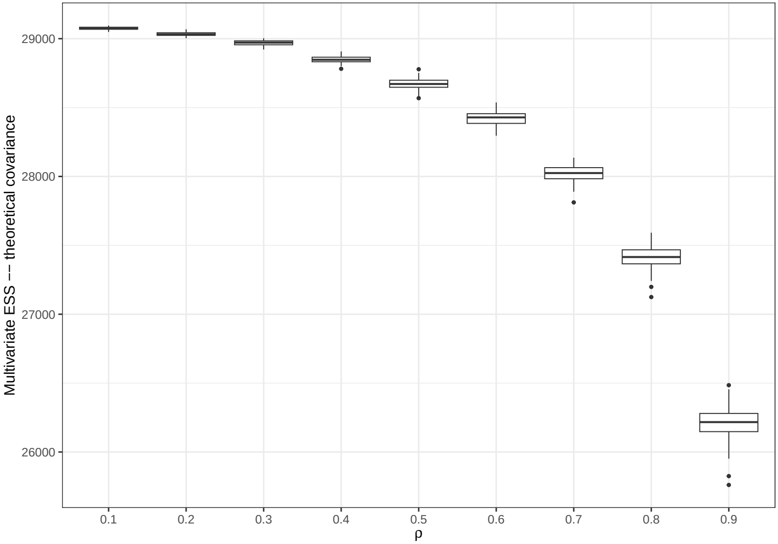
B) **theoretical**: the multivariate ESS with $\mathbf{\Sigma}$ set to its theoretical value.

Results A



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

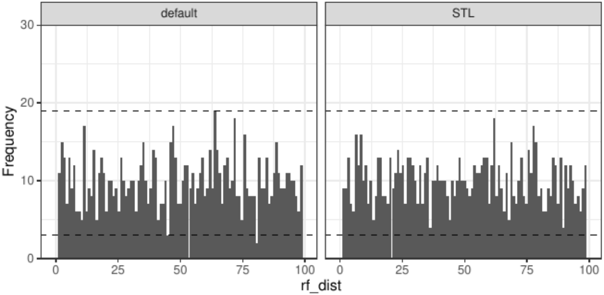
Results B


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

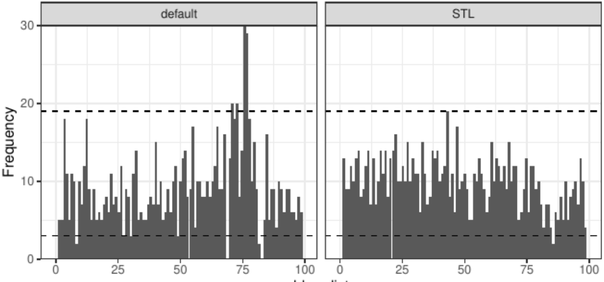
SBC for trees

0. Generate a reference tree from the prior $\bar{\tau}_0 \sim \pi_T(\tau|\boldsymbol{\gamma})$;
   for each iteration in 1:N, **do**:
1. Generate $\bar{\tau} \sim \pi_T(\tau|\boldsymbol{\gamma})$;
2. Compute the distance $\bar{\delta} = d_\sigma(\bar{\tau}, \bar{\tau}_0)$ according to the metric of choice;
3. Generate some (alignment) data $\tilde{y} \sim p(y|\bar{\tau}, \boldsymbol{\alpha})$;
4. Draw (approximately) $\boldsymbol{\tau}_s = \{\tau_s^{(1)}, \tau_s^{(2)}, \ldots, \tau_s^{(L)}\}$ from the posterior $\pi(\tau|\tilde{y})$;
5. Compute distances $\boldsymbol{\delta}_s = \{\delta_1, \delta_2, \ldots, \delta_L\}$ with $\delta_i = d_\sigma(\tau_s^{(i)}, \bar{\tau}_0)$;
6. Compute the rank $r(\boldsymbol{\delta}_s, \bar{\delta}) = \sum_{i=1}^{L} \mathbb{I}(\delta_i < \bar{\delta})$.

Some results: tree distances



(a) Robinson-Foulds, $\mathrm{RF}_0(\tau)$



Some results: continuous parameters

Simulation Based Calibration

prior sample: ../../truth.log
posterior samples: combined.log
Use ranking for bins

Tree.height

Missed: 0

Tree.treeLength

Missed: 0

kappa

Missed: 0

gammaShape

Missed: 0

popSize

Missed: 0

CoalescentConstant

Missed: 0

freqParameter.1

Missed: 1

freqParameter.2

Missed: 0

freqParameter.3

Missed: 0

freqParameter.4

Missed: 0

Statistics in the space of phylogenetic trees

- Central Limit Theorem(s) in BHV space: Barden, Le & Owen (2013);
- "Statistics in the Billera-Holmes-Vogtmann space": Weyenberg (2015);
- Consistency of the MLE: RoyChoudhury, Willis & Bunge (2015);
- How to turn tree space into an Euclidean space: Barden & Le (2017);
- Quantifying uncertainty about phylogenies: Willis & Bell (2018);
- Confidence sets for phylogenies: Willis (2018);
- Probabilistic path Hamiltonian Monte Carlo for phylogenies: Dinh et al. (2017).

Open problems in MCMC for phylogenies

Open problems:

- How can we construct more efficient proposals? How to exploit structure?
  - **Geometry!**
- How to quantify exploration of the target?
  - Exploit subtrees;
  - Exploit quasi-lumpability (?);
  - Multi-dimensional scaling (?).
- **Optimal scaling: what's the optimal acceptance probability?**

Take home[3]

Searching trees is **hard**

Complicated and **HUGE** parameter space

---

[3] this talk is available online

Take home[3]

Searching trees is **hard**

Complicated and **HUGE** parameter space

Height-preserving tree rearrangements are **good**

Use the extra information provided by the tip dates

---

[3] this talk is available online

Take home[3]

Searching trees is hard

Complicated and **HUGE** parameter space

Height-preserving tree rearrangements are good

Use the extra information provided by the tip dates

Validation is hard but feasible

Using the coalescent and SBC (with clever metrics) gives us a bit of hope.

---

[3] this talk is available online

Take home[3]

Searching trees is **hard**

Complicated and **HUGE** parameter space

Height-preserving tree rearrangements are **good**

Use the extra information provided by the tip dates

Validation is hard but feasible

Using the coalescent and SBC (with clever metrics) gives us a bit of hope.

Much more work is needed

We should prepare for an era of plenty

---

[3] this talk is available online

THE
END