Al buio non si trova

Biostatistics in the 21st century

Luiz Max Carvalho
lmax.fgv@gmail.com

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

FGV EMAp

Le dirò con due parole, chi son

Personal

- Born and raised in Petrópolis-RJ;
- Eldest of three kids;
- Married and father of two sassy girls;
- Mais Querido supporter.

Academic

- BSc in Microbiology & Immunology (UFRJ, 2012);
- PhD Evolutionary Biology (Edinburgh, 2018);
- Post doctoral researcher at ENSP/Fiocruz (2019);
- Lecturer (Assistant Professor) at EMAp since Jan/2020.

E che faccio

Applications of Statistics/Mathematics

Applications in Epidemiology, (Molecular) Biology, Ecology, Psychology, Linguistics, etc.

Applied Statistics

Markov Chain Monte Carlo, Model combination and selection, Statistical Phylogenetics.

Junior colaborators (Or, you know, students) I

Mentoring/Honours

- Yure Oliveira: "Bayesian consistency under the normalised power prior";
- Rodrigo Kalil: "Extending joint models".

Junior colaborators (Or, you know, students) II

MSc

- Ezequiel Braga (MSc) "Principled Bayesian analysis under the normalised power prior";
- Eduardo Adame (MSc) "Exact MCMC methods for the normalised power prior";
- Iara Castro (MSc) "Using survival analysis to understand cancer treatment outcomes";
- Wellington Silva (MSc) "Adaptive truncation of infinite series: applications to Statistics";
- Igor Michels (MSc) "Calibration of Bayesian player-level models for Brazilian football".

Junior colaborators (Or, you know, students) III

PhD

- Felipe Schardong (PhD) “Mathematical modelling of antimicrobial resistance”;
- Atilio Pellegrino (PhD) “Model combination for epidemiological forecasting”.

Postdocs

- Fernanda Valente: "Spatio-temporal modelling of dengue and its vectors";
- Rodrigo Alves: "Tree-valued stochastic processes".

Problem I: efficiently utilising available information

Loads of historical data: how to build informative priors?

Let $\boldsymbol{y_0} = (y_{01}, \ldots, y_{0n_0})$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ be **historical** and **current** data, respectively.

Question: how do I build a prior that

- Uses information in $\boldsymbol{y_0}$ efficiently but also
- Does not lead to borrowing too much information when the data sets are not compatible?

Applications: clinical trials, quality control, policy-making.

I got the power

Normalised power prior[1]

$$\tilde{p}(\theta, \eta \mid \boldsymbol{y}_0) := \frac{L_0(\boldsymbol{y}_0 \mid \theta)^\eta \pi_0(\theta \mid \psi) \pi_A(\eta \mid \phi)}{c(\eta; \psi, \phi)}$$

- How pick π_A such that prediction error (say) is minimised?
- How to **efficiently** compute

$$c(\eta; \psi, \phi) = \int_\Theta L(\boldsymbol{y}_0 \mid t)^\eta \pi(t \mid \psi) \, d\mu(t)$$

by leveraging its special properties as function of η?

[1] https://doi.org/10.1002/sim.9124

Open problems

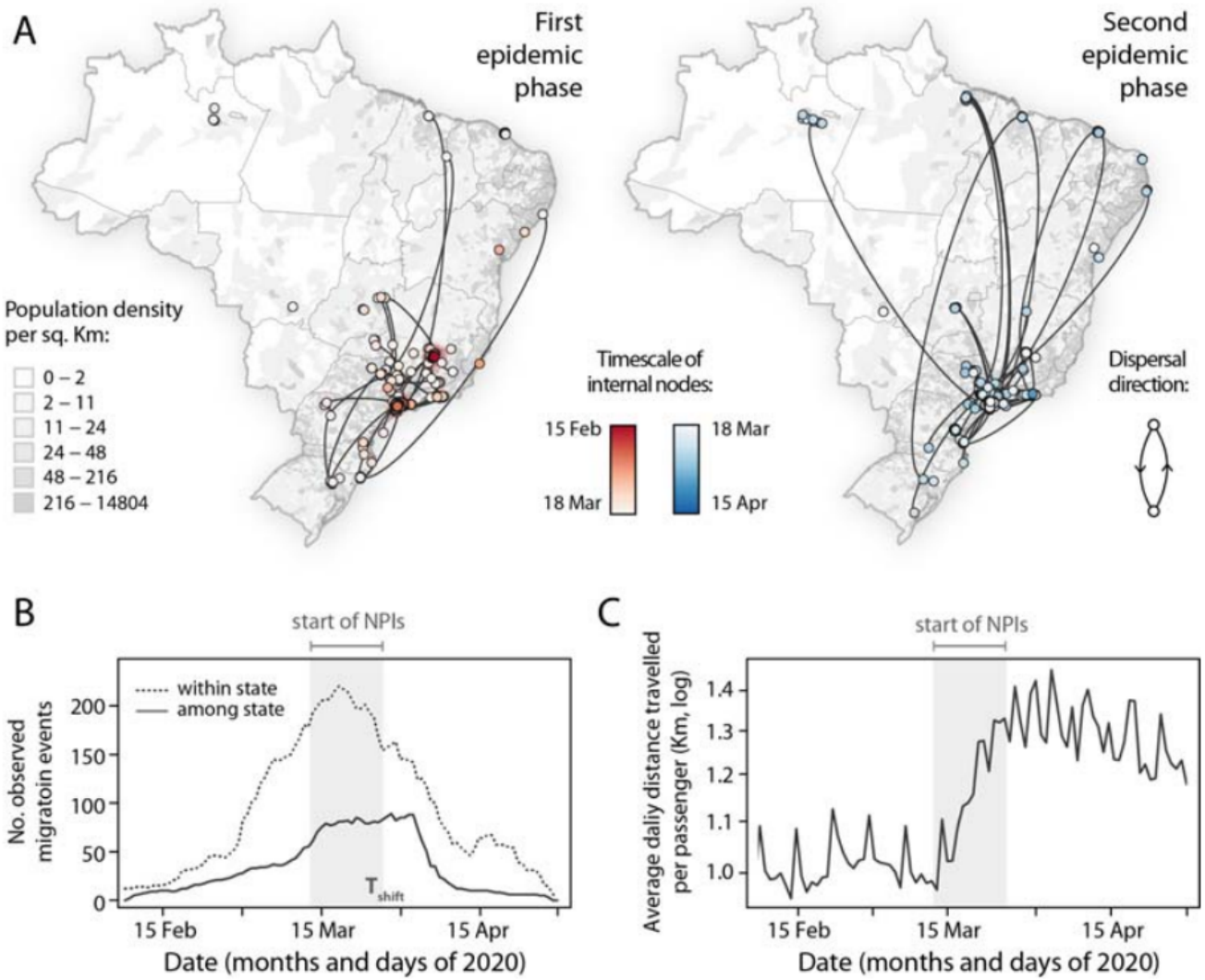
What happens to the posterior

$$p(\theta, \eta \mid \boldsymbol{y}_0, \boldsymbol{y}) \propto L(\boldsymbol{y} \mid \theta)\tilde{p}(\theta, \eta \mid \boldsymbol{y}_0),$$

- in various asymptotic regimes (e.g. $n \to \infty$ with $n/n_0 = r$)?
- For finite (n, n_0) when $\text{dist}(\boldsymbol{y}_0, \boldsymbol{y}) > \delta$?
- Can we sample exactly from this doubly-intractable distribution?
- How to pick $\pi_A(\cdot \mid \phi)$?

Problem II: dealing with huge complex data

Where did this virus come from?[2]



[2] https://doi.org/10.1126/science.abd2161

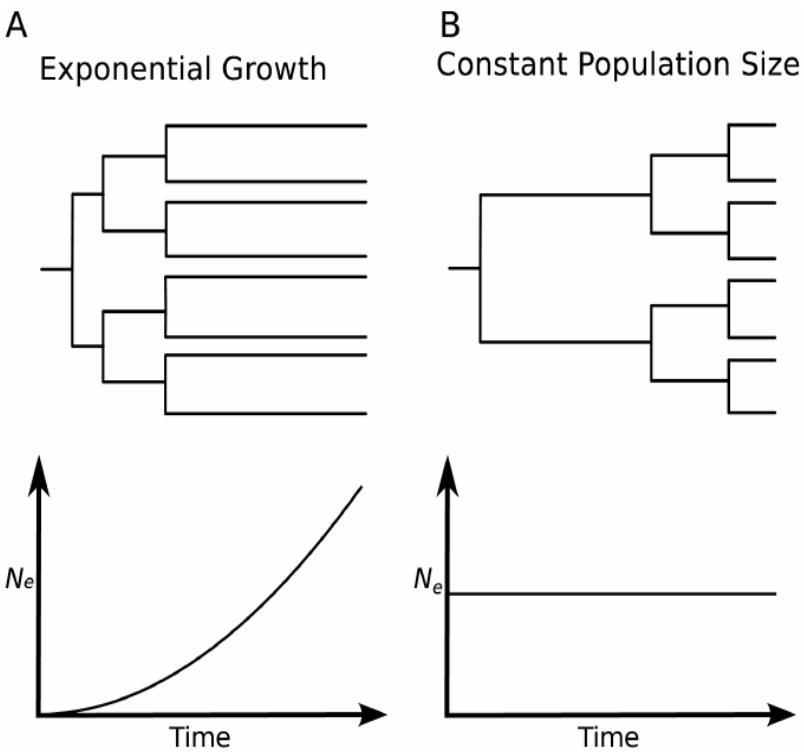
Motivation

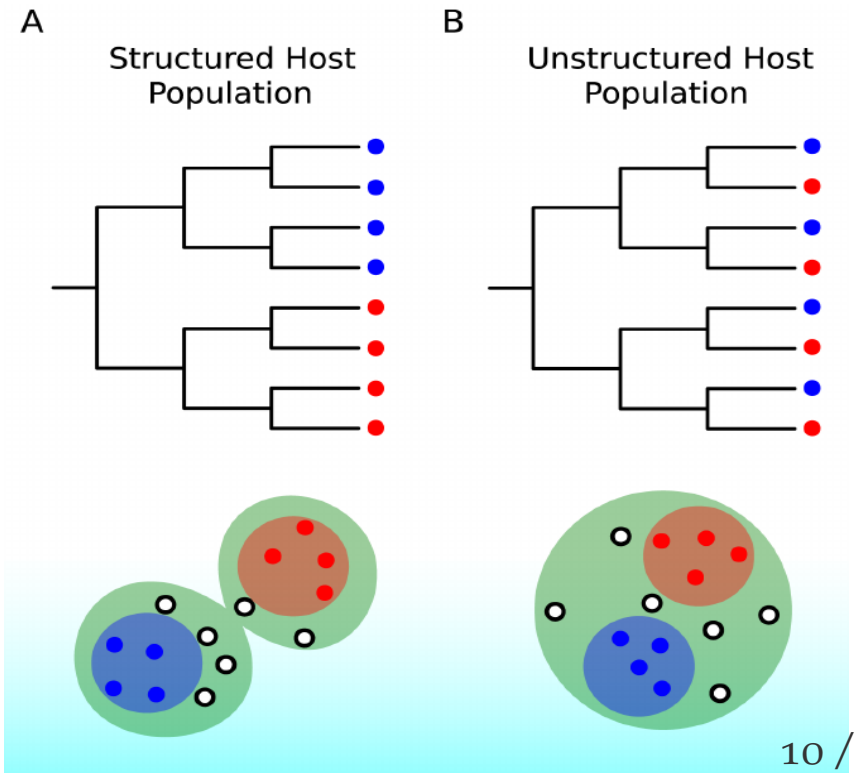
Phylodynamics of fast-evolving viruses

Inferring spatial and temporal dynamics from genomic data:

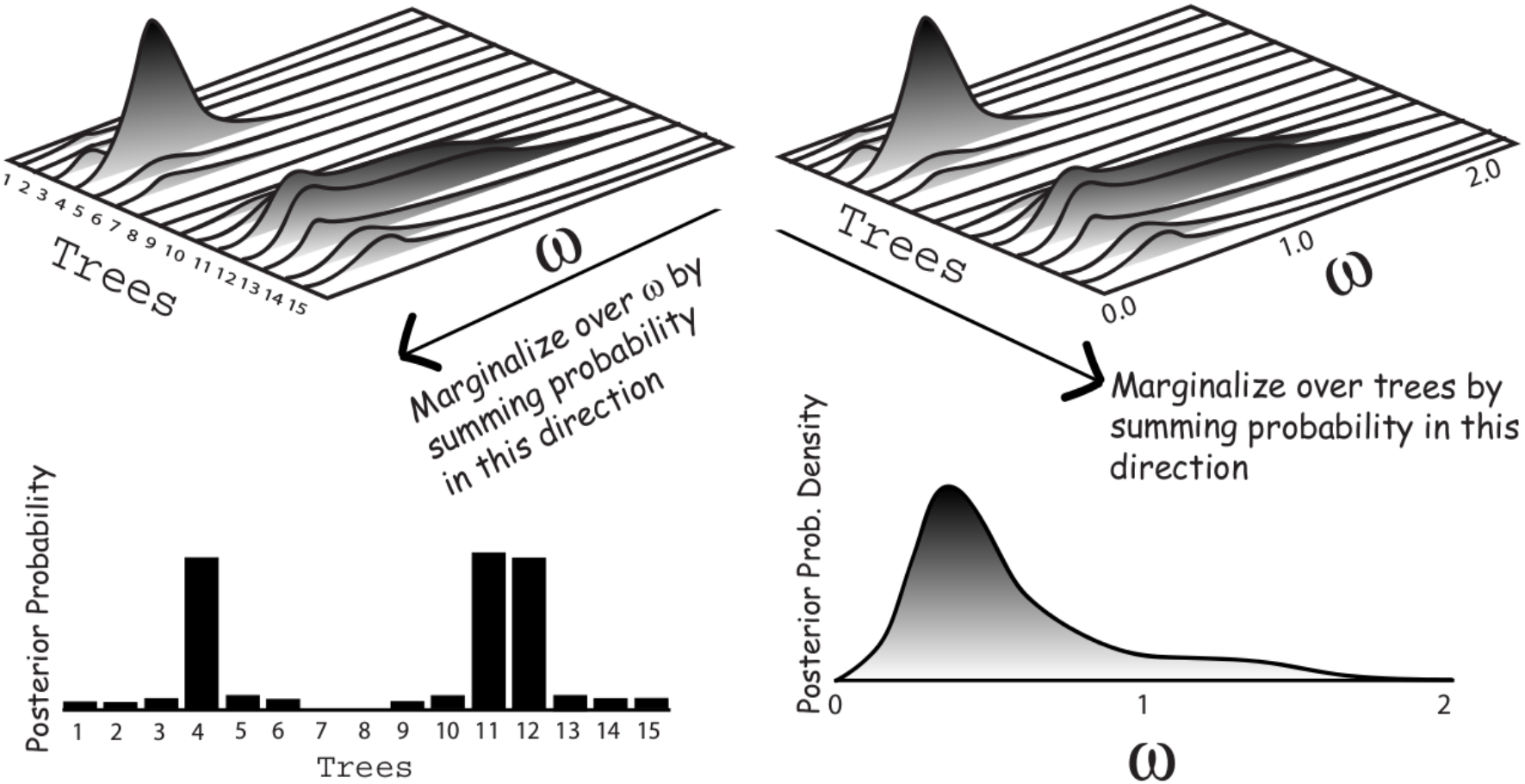
Phylogenies*!

* plus complicated models



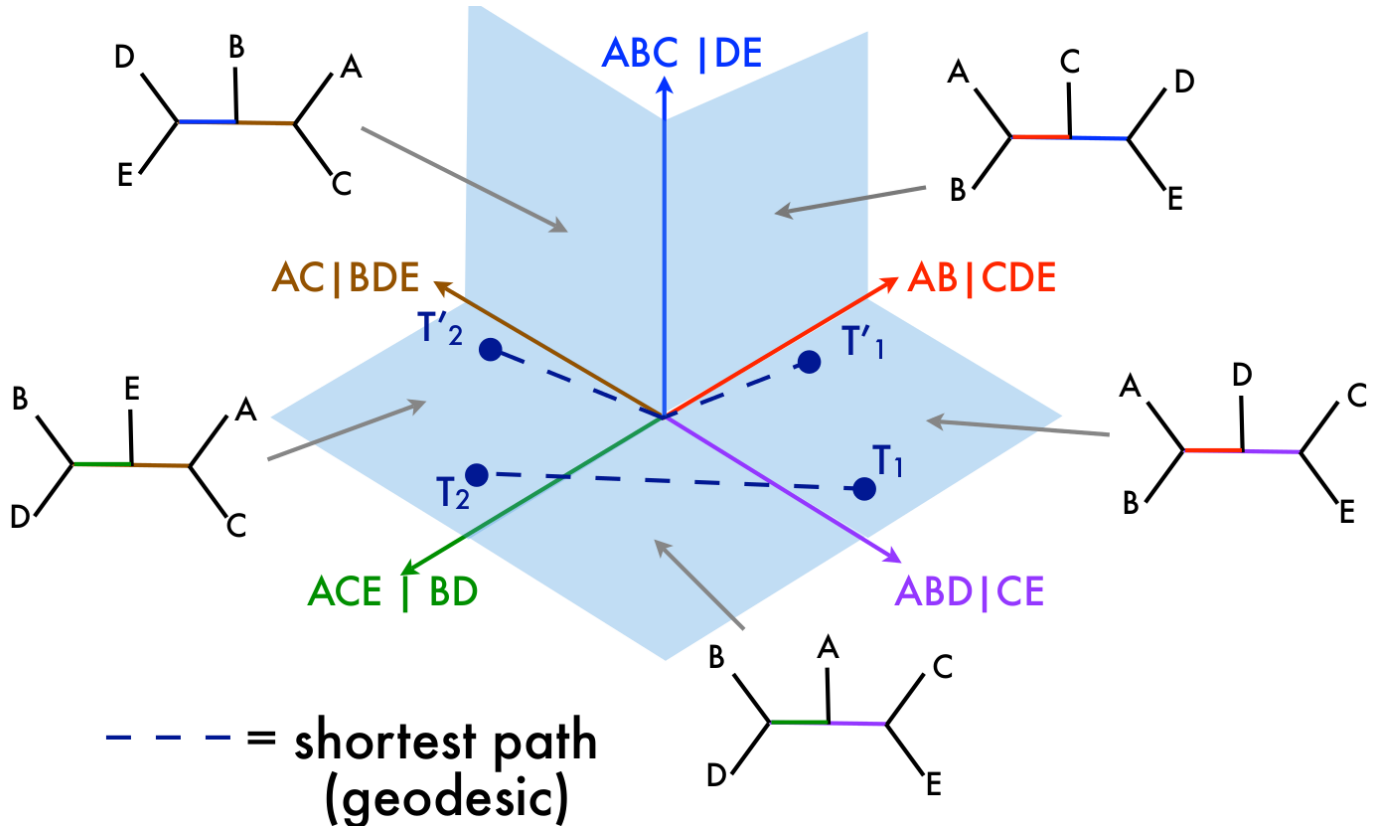


The end product



This place is weird..

Traversing cubic complexes efficiently[3]



Applications: Molecular Epidemiology, Evolutionary Biology.

[3] https://youtu.be/h9bWRQ6aeKA

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? (Custom) **Tools!**
- Optimal scaling: what's the **optimal** acceptance probability?

Take home

A light in the dark

Maths gives us methods with provable guarantees

Computational methods are key

Learn to program and learn Computational Statistics[4]

Maths at the service of Science

My research employs: combinatorics, probability theory, basic calculus, optimisation and classical **and** Bayesian Statistics.

We've got loads to do!

Biomedical statistics is where most of the cool data and problems **with actual impact** are.

---

[4] Here's a place to start:
https://github.com/maxbiostat/Computational_Statistics

THE END