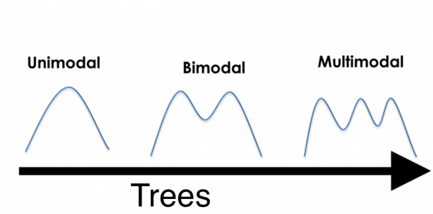
Statistics for phylogenetic time trees

Lars Berling

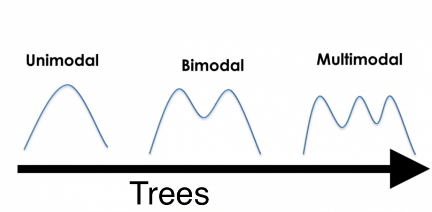
11/02/2025

UC UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

The goal



The goal



Bayesian Phylogenetic Inference via MCMC

- Key object of interest is often the rooted tree topology.
- **MCMC Output**: Sample of trees (typically thousands)
- Challenges in estimating mean and variance in treespace due to its high-dimensional, non-Euclidean nature.

Treespaces

τ-space

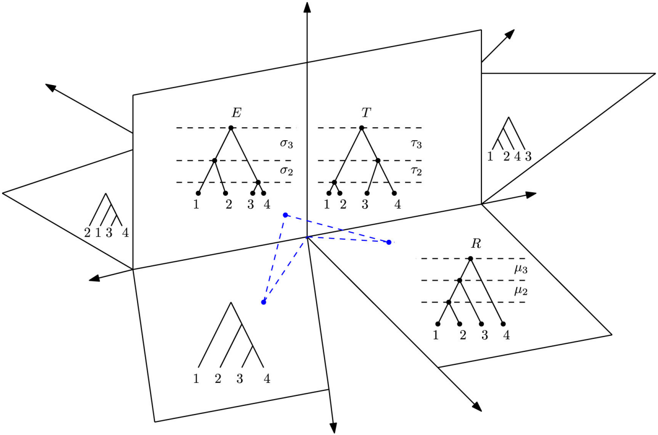


Figure: Three-dimensional projection of a part of τ space with 4 taxa.

The downfall of stratified spaces

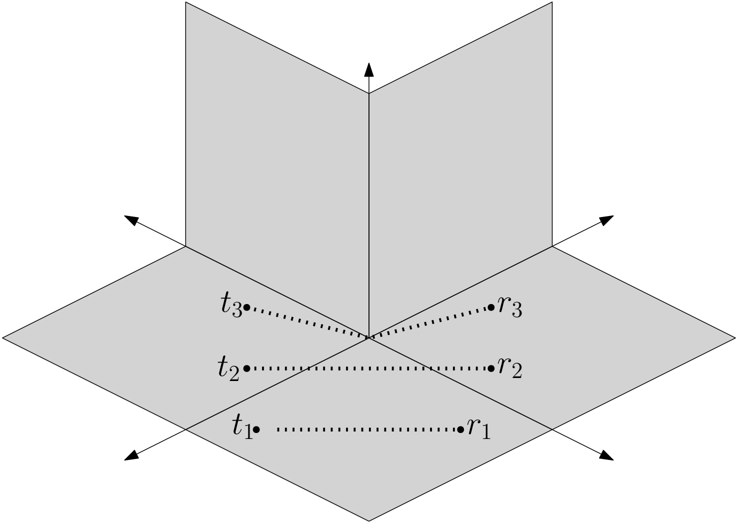
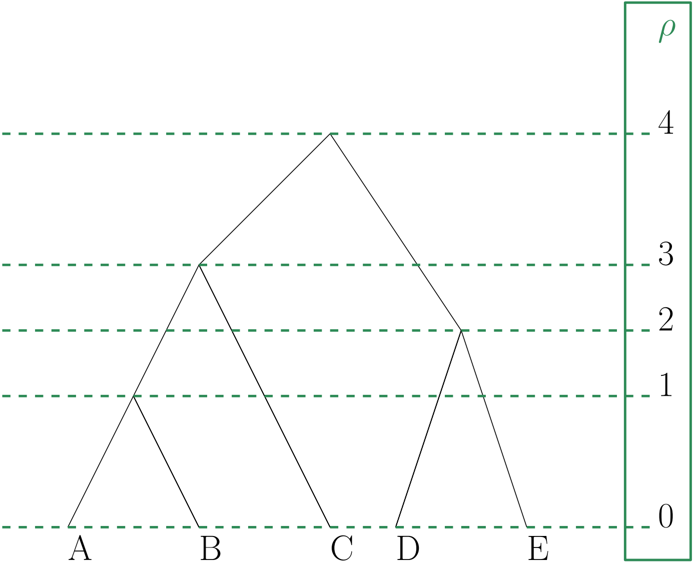
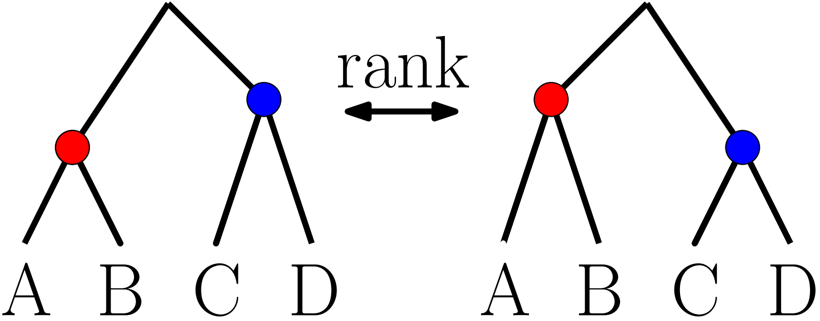


Figure: Geodesics and cone paths

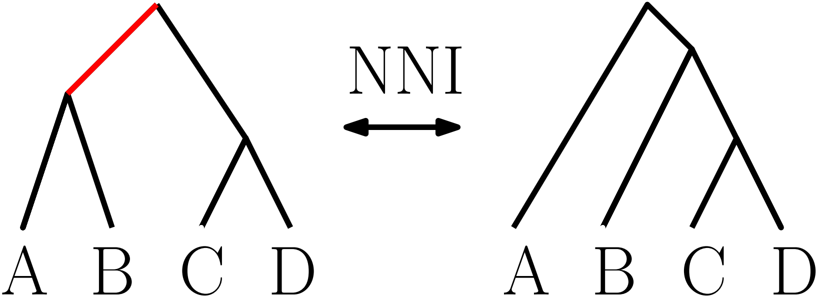
The space of ranked trees



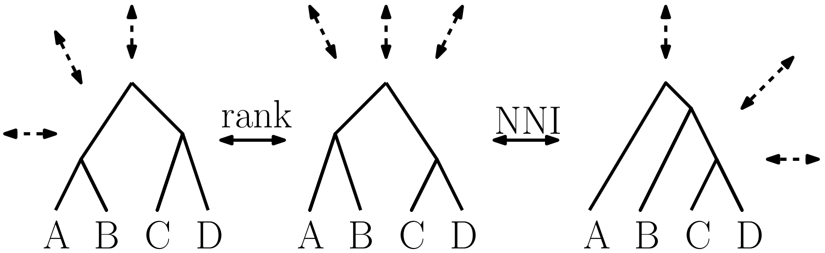
The space of ranked trees



The space of ranked trees



The space of ranked trees



The RNNI graph

This graph is the **treespace** of ranked trees, **R**anked **N**earest **N**eighbour **I**nterchange space.

The shortest path problem

Induced distance

The minimal **number of rearrangement operations** to transform one tree into another

Equivalent: find a **path of minimal length** in the RNNI graph

The shortest path problem

Induced distance

The minimal **number of rearrangement operations** to transform one tree into another

Equivalent: find a **path of minimal length** in the RNNI graph

Theorem

The **FindPath** algorithm computes a shortest path in RNNI, with time complexity $\mathbf{O}(\mathbf{n^2})$. [a]

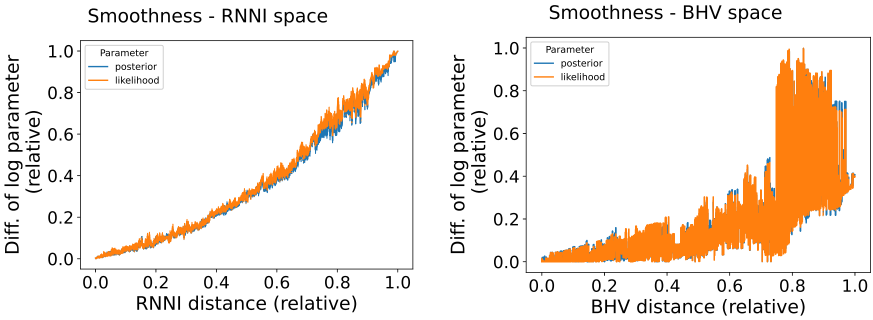
[a] Collienne, Lena, and Alex Gavryushkin. "Computing nearest neighbour interchange distances between ranked phylogenetic trees." Journal of Mathematical Biology 82.1-2 (2021): 8.

Probability distributions are 'continuous'

Comparing **probability distributions** in BHV and RNNI tree space on **one data set**. **x**-Axis displays **tree metric distance** (relative), **y**-axis displays **difference in probability** (relative)

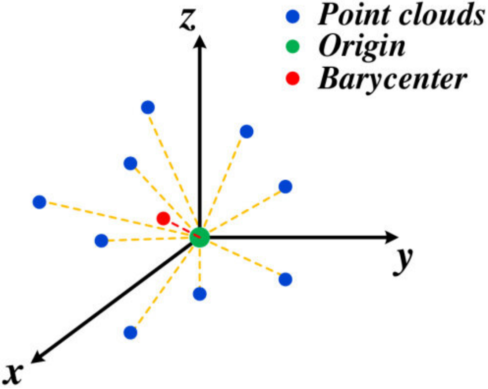
Probability distributions are 'continuous'

Comparing **probability distributions** in BHV and RNNI tree space on **one data set**. **x**-Axis displays **tree metric distance** (relative), **y**-axis displays **difference in probability** (relative)



Mean tree

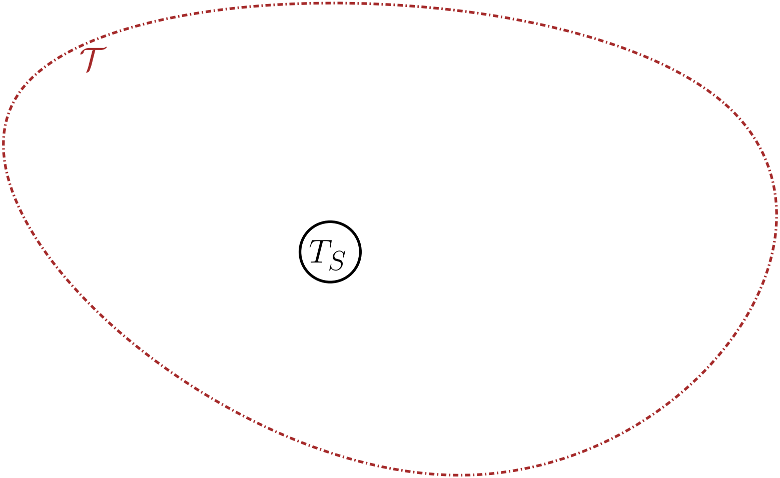
Geometric means



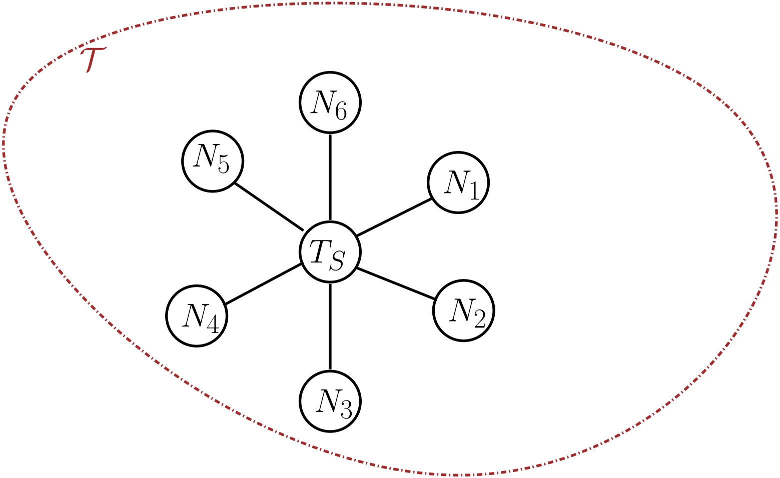
Fréchet variance

$Var(t)_{\mathcal{T}} = \sum_{t_i \in \mathcal{T}} d(t_i, t)^2$

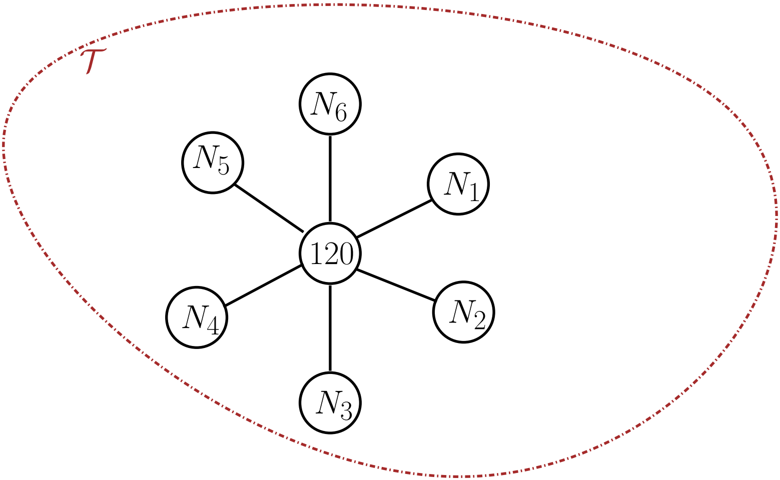
The algorithm



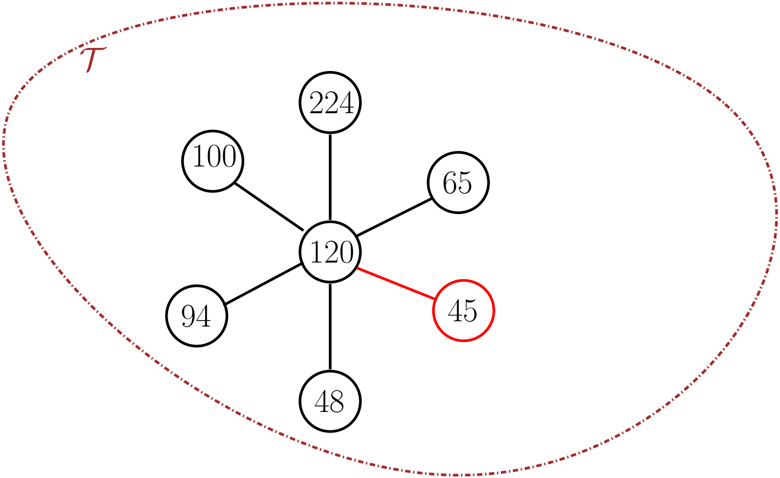
The algorithm



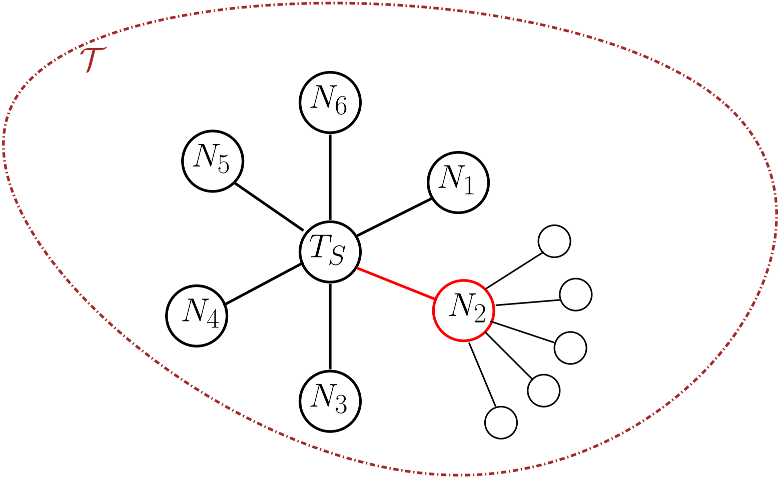
The algorithm



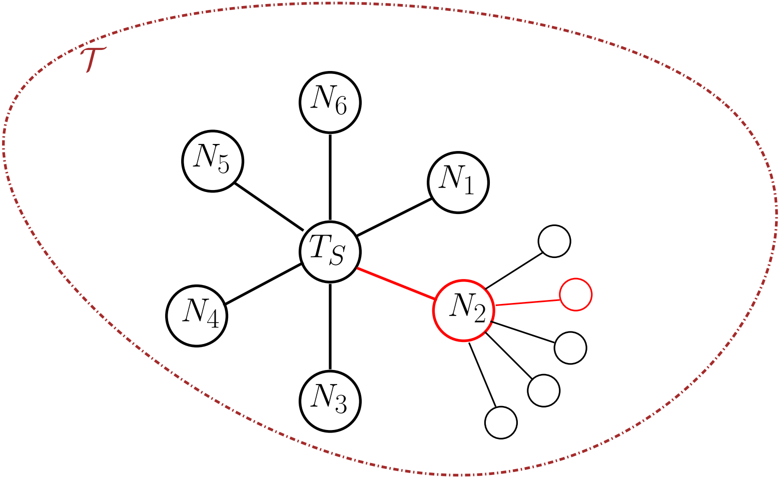
The algorithm



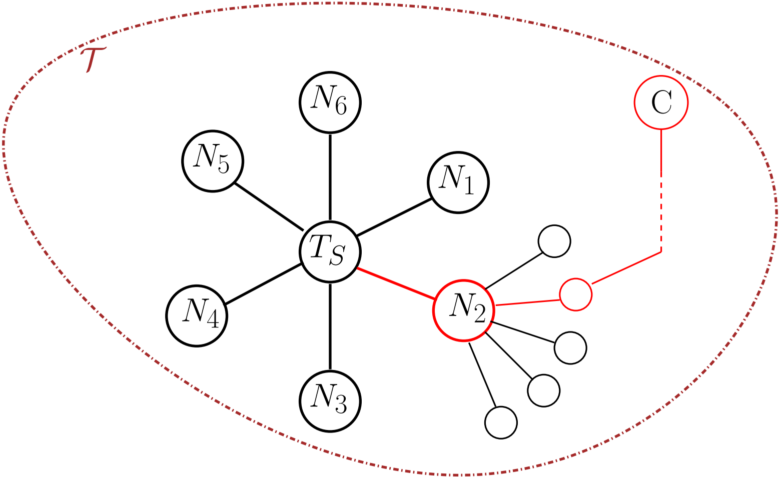
The algorithm



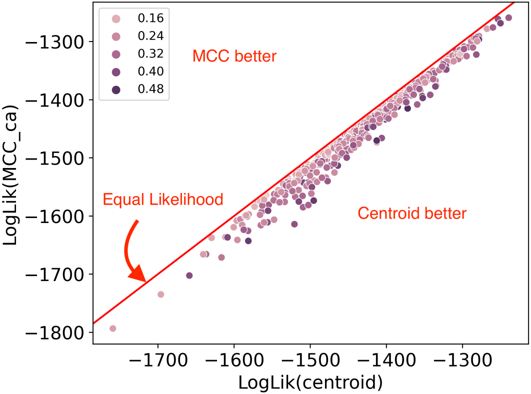
The algorithm



The algorithm



Comparing Likelihood to MCC



MCC

Maximum Clade Credibility Tree from treeannotator (BEAST).

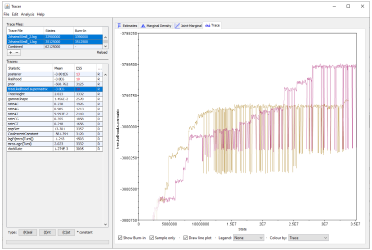
Convergence Assessment

Convergence

- Sampling from the stationary distribution

→ Parameter trace no trend

- **E**ffective **S**ample **S**ize at least 200 (rule of thumb)



Trace File	States	Burn-In
2chains50mil_2.log	33900000	3390000
2chains50mil_1.log	35125000	3512500
Combined	62125000	-

Statistic	Mean	ESS	...
posterior	-3.801E6	13	R
likelihood	-3.8E6	10	R
prior	-568.762	3125	R
treeLikelihood.supermatrix	-3.8E6	10	R
TreeHeight	2.023	3332	R
gammaShape	1.456E-2	2570	R
rateAC	0.238	1926	R
rateAG	0.985	1213	R
rateAT	9.993E-2	2110	R
rateCG	0.355	1858	R
rateGT	0.248	1656	R
popSize	13.301	3357	R
CoalescentConstant	-561.394	3120	R
logP(mrca(Tursi))	-1.243	4503	R
mrca.age(Tursi)	2.023	3332	R
clockRate	1.274E-3	3095	R

☑ Show Burn-in ☑ Sample only ☑ Draw line plot Legend: None Colour by: Trace

Figure: Not converged

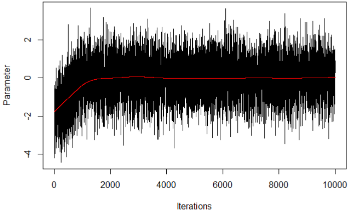


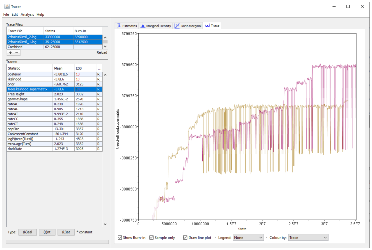
Figure: Converged

Convergence

- Sampling from the stationary distribution

→ Parameter trace no trend

- **E**ffective **S**ample **S**ize at least 200 (rule of thumb)



Trace Files:

Trace File	States	Burn-In
2chains50mill_2.log	33900000	3390000
2chains50mill_1.log	35125000	3512500
Combined	62125000	-

Traces:

Statistic	Mean	ESS	...
posterior	-3.801E6	13	R
likelihood	-3.8E6	10	R
prior	-568.762	3125	R
treeLikelihood.supermatrix	-3.8E6	10	R
TreeHeight	2.023	3332	R
gammaShape	1.456E-2	2570	R
rateAC	0.238	1926	R
rateAG	0.985	1213	R
rateAT	9.993E-2	2110	R
rateCG	0.355	1858	R
rateGT	0.248	1656	R
popSize	13.301	3357	R
CoalescentConstant	-561.394	3120	R
logP(mrca(Tursi))	-1.243	4503	R
mrca.age(Tursi)	2.023	3332	R
clockRate	1.274E-3	3095	R

Type: (R)eal (I)nt (C)at * constant

☑ Show Burn-in ☑ Sample only ☑ Draw line plot Legend: None Colour by: Trace

Figure: Not converged

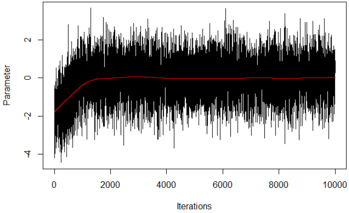
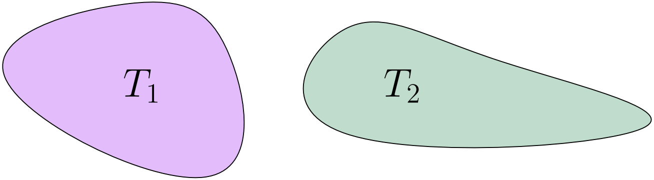


Figure: Converged

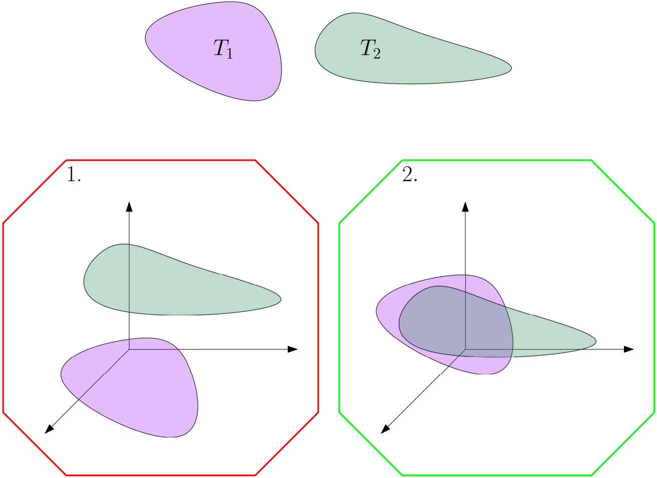
Do two sets have the same underlying distribution?

independently sampled sets of trees

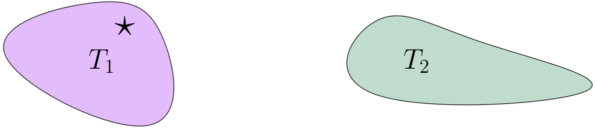


Do two sets have the same underlying distribution?

independently sampled sets of trees

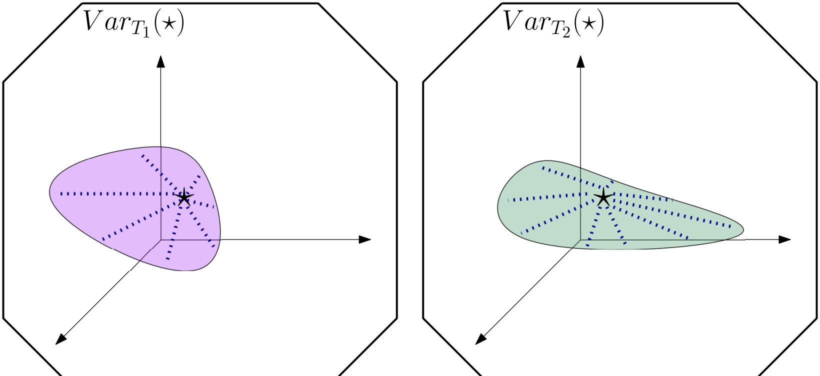


Gelman Rubin diagnostic for trees



$\star$ = new sample in T_1

⋯ = RNNI distance



Gelman Rubin diagnostic for trees

Potential scale reduction factor

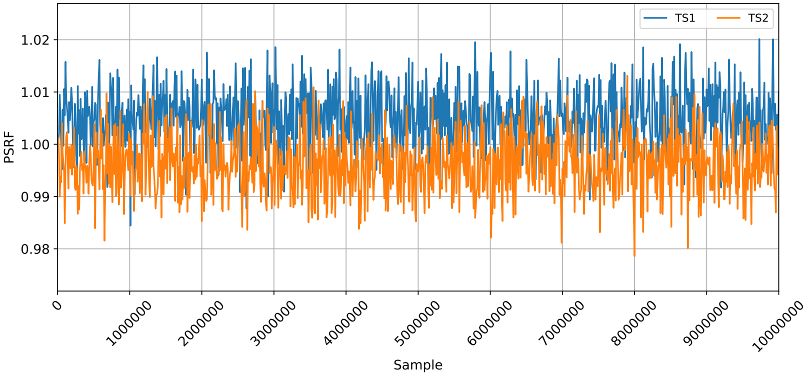
$$PSRF(t|\mathcal{T}_1, \mathcal{T}_2) = \sqrt{\frac{Var(t)_{\mathcal{T}_2}}{Var(t)_{\mathcal{T}_1}}}, t \in \mathcal{T}_1$$ [a]

[a] Inference from Iterative Simulation Using Multiple Sequences, A. Gelman and D. Rubin

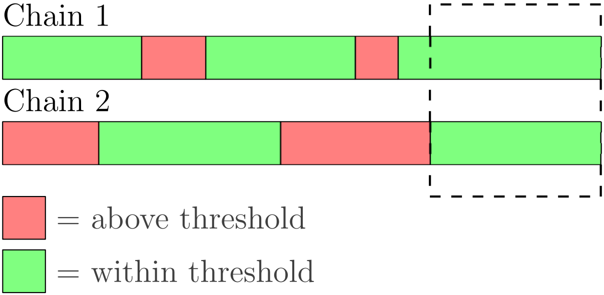
Fréchet variance (normalized)

$$Var(t)_{\mathcal{T}} = \frac{\Sigma_{t_i \in \mathcal{T}} d(t_i, t)^2}{|\mathcal{T}|}$$

Gelman Rubin diagnostic for trees



Gelman Rubin diagnostic for trees



Further assessment of overlap

- **E**ffective **S**ample **S**ize at least 200 (rule of thumb)
- Further downstream analysis: Summarizing/ computing a mean tree

Results on DS1-DS11

		mean - 0.05						mean - 0.02						mean - 0.01					
	ESS-threshold	C1	C2	C3	C4	C5	C6	C1	C2	C3	C4	C5	C6	C1	C2	C3	C4	C5	C6
DS1	200	0	0	1	1	1	1	0	0	0	1	1	1	0	0	0	0	1	1
	500	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
DS2	200	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
	500	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	0	1
DS3	200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	500	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DS4	200	0	1	1	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0
	500	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0
DS5	200	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
	500	1	0	1	1	1	1	0	0	1	1	1	1	0	0	1	1	0	1
DS6	200	1	1	0	1	1	1	0	0	0	0	1	1	0	0	0	0	0	1
	500	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
DS7	200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	500	1	1	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	1
DS8	200	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
	500	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
DS9	200	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	500	0	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1
DS10	200	1	1	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1
	500	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DS11	200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	500	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Orange cell = convergence, Blue cell = non-convergence

Conclusion

Summary and Outlook

- Developed a mean tree within the RNNI treespace

→ Theoretical foundation is still an open problem

- Convergence assessment of samples of trees

→ Multimodal distributions?

→ Advanced characterization of tree distributions

Conclusion

Software

- BEAST2 package: ASM
  `https://github.com/rbouckaert/asm`
- Python package: tetres
  `https://github.com/bioDS/tetres`

References

- Lars Berling; Lena Collienne; Alex Gavryushkin, Estimating the mean in the space of ranked phylogenetic trees, Bioinformatics (2024)
- Lars Berling; Remco Bouckaert; Alex Gavryushkin, An Automated Convergence Diagnostic for Phylogenetic MCMC Analyses, IEEE/ACM TCBB (2024)