Chapter 3

Luiz Max Fagundes de Carvalho

October 11, 2017

THE UNIVERSITY
of EDINBURGH

School of BiologicalSciences
Institute of Evolutionary Biology
West Mains Road
Edinburgh, EH9 3FL
Scotland

Contents

Chapter 1

Convergence diagnostics for Markov Chain Monte Carlo in Bayesian phylogenetics: the case of time-trees

1.1 Motivation

Markov chain Monte Carlo (MCMC) methods have become a standard tool for approximating complex posterior distributions encountered in Bayesian inference (**?**). In phylogenetics, most if not all Bayesian approaches rely on MCMC for approximating the posterior distribution of trees (**???**). These methods rely on constructing a Markov chain whose stationary distribution is the (target) distribution one wishes to sample from. A fundamental issue is to determine when the chain has reached stationarity and samples are being drawn from the target distribution. Whilst much attention has been given to this issue in the statistical literature, most diagnostic methods assume univariate, continuous parameter spaces. Discrete, high-dimensional parameter spaces such as those encountered in phylogenetics pose additional challenges to development of effective convergence diagnostic tools.

Available methods for diagnosing convergence of MCMC for Bayesian phylogenetics include tracking clade (split) frequencies both within and between chains (**?**), multi-dimensional scaling of tree distance matrices **??** and network-based clustering (**?**). These methods are mostly graphical in nature, and only recently have more formal convergence metrics been proposed (**??**).

An important thing to notice is that it is not possible to say with complete certainty when a Markov chain has converged to its target distribution. Rather, convergence tools are designed to identify failure to converge. As argued by **?**, when the data do not conform with the model (e.g., come from a mixture of trees rather than a single tree) apparent convergence can be misleading. **?** and **?** further reinforce the point that multiple convergence diagnostics need to be employed in order to mitigate the risk of determining convergence when in fact chains have not reached the desired target. Thus, no single method or

tool is likely to supersede all the others, as there are cases where one method fails to detect problems but others identify failure to converge. Successful application of convergence detection tools fundamentally depends on combining several metrics/tools in one coherent framework (e.g. the approaches of **?** and **?**).

An additional issue with currently available methods is that most assume either unrooted trees and/or contemporaneous sequences, limiting their applicability in cases where one deals with time-calibrated phylogenies, or time-trees (see below). **?** attempt to integrate most of the popular visualisation methods, along with some quantitative indicators, into one framework. My goal here is to expand upon their approach and make the necessary adaptations to accommodate time-calibrated trees with hundreds of taxa. In what follows I review some key concepts in Bayesian phylogenetics as well as the state-of-the-art for convergence diagnostics in phylogenetic MCMC. I then proceed on to discussing the limitations of available methods when dealing with time-calibrated trees and suggest adaptations.

1.1.1 Time-tree space

To understand the challenges of assessing convergence of phylogenetic MCMC, it is desirable to describe the parameter space which it attempts to explore. First, it is convenient to define some notation. Let $t \in \mathbb{T}$ be a binary rooted tree on n taxa (leaves), with $2n-3$ edges. Each of the edges in t can associated with a unique node numbering. We can supplement t with a set of branch lengths $\boldsymbol{b} = \{b_1, b_2, \ldots, b_{2n-3}\}$, $\boldsymbol{b} \in \boldsymbol{B} \subseteq \mathbb{R}_+^{2n-3}$. Denote the object $(t, \boldsymbol{b}) = \tau \in \boldsymbol{\Psi}$. For convenience, we will henceforth call t a **topology** and τ a **tree**. We will also use the terms tree and phylogeny interchangeably.

The first important thing to notice is that while $\mathbb{T}$ is discrete and finite (despite

Throughout this paper we will use $\boldsymbol{\Psi}$ to denote the parameter space encompassing topologies and branch lengths, henceforth called "tree space", and $\tau \in \boldsymbol{\Psi}$ to denote a bifurcating, rooted tree with branch lengths on n taxa.

An striking feature of the space of phylogenetic trees is its sheer size. A well-known counting argument shows that there are $|\mathbb{T}| = 1 \times 3 \times \ldots \times (2n-5) \times (2n-3) = (2n-3)!!$ binary rooted phylogenies on $n \geq 3$ taxa. In her review of the geometry of tree space, **?** argues that the power of the tree model "comes from the property that adds the complexity: the vast number of trees to explain different possible evolutionary scenarios" (pg e83).

As argued by **?**, the complexity of tree space can be seen as a major reason for the development of specialised software for Bayesian phylogenetics as opposed to the use of common MCMC packages such as Stan and JAGS.

An useful way of representing discrete tree space is equipping it with a metric and construct a graph $G_\delta(V, E)$ where each tree corresponds to a topology and there is an edge between two edges (trees) if they are a distance $d \leq \delta$ apart under the chosen metric.

1.1.2 Tree metrics

? proposed a new metric based on ...

1.1.3 MDS

?

? developed an R package to aid MDS visualisation under diffent metrics, with special focus on the KC metric (**?**).

1.2 Convergence of continuous parameters

REVIEW basics (PSRF, Gweke, etc) + ESS

1.3 Convergence in tree space

1.3.1 Clade frequencies

The approach of **?** is to analyse clade/split frequencies to assess convergence of MCMC in a phylogenetic space. The program AWTY (short for "are we there yet?") provides graphical facilities for assessing convergence by analysing various aspects of the distribution of sampled clades. Most of the diagnostic plots rely on two independent chains.

By plotting clade frequencies estimated in two independent chains against each other (scatterplot), one can assess whether both chains have converged to similar distributions. Lack of convergence can be detected when points fall away from the identity ($x = y$) line. For a single chain, one useful diagnostic is plotting cumulative clade frequencies along the chain. If these trajectories present long-term trends, it means clade frequencies have not stabilised, indicating lack of convergence.

Absence-presence plots show whether a particular clade was absent or present in the tree sampled at each iteration of the chain. If there are long periods where the clade is either absent or present, this indicates the chain has not mixed well and might not have converged. On the other hand, a traceplot of this kind where the indicator variable frequently switches between 0 and 1 indicates good mixing. This notion of "clade-switching" can be made more precise (see below).

Finally, one can also plot the distance

A modern incarnation of AWTY, RWTY (**?**) seeks ...

How do clade frequencies relate to distance to true tree (for simulated data)? Explore Figure 4 in **?**: can we link ESS, clade frequencies and MDS?

Let $\boldsymbol{X}_i = \{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\} \in [0,1]^n$ be a collection of samples from a Markov chain such that $X_i^{(j)} = 1$ if clade i was sampled in the j-th iteration and 0 otherwise. Also, for $s_i = \sum_k X_i^{(k)}$ we call $f_i = s_i/n$ the *frequency* of clade i.

1.3.2 Clade switching

Let $m_i = \min(n - s_i, s_i)$, it can be shown that the maximum number of transitions that can be observed from $\boldsymbol{X}_i$ is either $J_i = 2m_i$[1].

[1]Technically, J_i depends on the first state $X_i^{(1)}$. Suppose w.l.o.g. that $m_i = s_i$. Then $J_i = 2m_i - 1$ if $X_i^{(1)} = 1$ and $J_i = 2m_i$ otherwise.

Let $\delta_i = \Delta(\boldsymbol{X}_i)$, where $\Delta(\cdot)$ a function that counts the number of state transitions in $\boldsymbol{X}_i$. Then $\sigma_i = \delta_i / J_i \in [0, 1]$ is a score that measures the relative efficiency of sampling by comparing how how many transitions happened compared to the theoretical maximum.

Choice of which clades to track.

1.3.3 Multi-dimensional scaling

Multi-dimensional scaling (MDS) (**?**)

explain + cite relevant papers

- Stress function - illustrative plot two panels (a) graph (b) MDS

In other words, can we tell apart the MDS of two runs, one converged and one not converged (as assessed with other criteria)?

1.3.4 Graph (network) analysis of tree space

Whidden & Matsen How to fix the approach of **?** ?

1.4 Accommodating time-calibrated phylogenies

1.5 Combining diagnostic measures

1.6 Final Remarks

Bibliography

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Drummond, A. J. and Bouckaert, R. R. (2015). *Bayesian evolutionary analysis with BEAST*. Cambridge University Press.

Hillis, D. M., Heath, T. A., and St John, K. (2005). Analysis and visualization of tree space. *Syst. Biol.*, 54(3):471–482.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *science*, 294(5550):2310–2314.

Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. (2017). treespace: statistical exploration of landscapes of phylogenetic trees. *Molecular Ecology Resources*.

Kendall, M. and Colijn, C. (2016). Mapping phylogenetic trees to reveal distinct patterns of evolution. *Molecular biology and evolution*, 33(10):2735–2743.

Lakner, C., Van Der Mark, P., Huelsenbeck, J. P., Larget, B., and Ronquist, F. (2008). Efficiency of markov chain monte carlo tree proposals in bayesian phylogenetics. *Systematic biology*, 57(1):86–103.

Lanfear, R., Hua, X., and Warren, D. L. (2016). Estimating the effective sample size of tree topologies from bayesian phylogenetic analyses. *Genome Biology and Evolution*, 8(8):2319–2332.

Li, S., Pearl, D. K., and Doss, H. (2000). Phylogenetic tree construction using markov chain monte carlo. *Journal of the American Statistical Association*, 95(450):493–508.

Matsen, F. A. (2006). A geometric approach to tree shape statistics. *Systematic biology*, 55(4):652–661.

Mossel, E. and Vigoda, E. (2005). Phylogenetic mcmc algorithms are misleading on mixtures of trees. *Science*, 309(5744):2207–2209.

Nylander, J. A., Wilgenbusch, J. C., Warren, D. L., and Swofford, D. L. (2008). AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics*, 24(4):581–583.

Robert, C. and Casella, G. (2011). A short history of markov chain monte carlo: subjective recollections from incomplete data. *Statistical Science*, pages 102–115.

St. John, K. (2017). Review paper: The shape of phylogenetic treespace. *Systematic Biology*, 66(1):e83.

Suchard, M. A., Weiss, R. E., and Sinsheimer, J. S. (2001). Bayesian selection of continuous-time markov chain evolutionary models. *Molecular biology and evolution*, 18(6):1001–1013.

Warren, D. L., Geneva, A. J., and Lanfear, R. (2017). Rwty (r we there yet): An r package for examining convergence of bayesian phylogenetic analyses. *Molecular biology and evolution*, 34(4):1016–1020.

Whidden, C. and Matsen, F. A. (2015). Quantifying mcmc exploration of phylogenetic tree space. *Systematic biology*, page syv006.