Journal of Theoretical Biology 336 (2013) 246–249

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Letter to Editor

Consistency of Bayesian inference of resolved phylogenetic trees

ARTICLE INFO

Keywords:
Bayesian phylogenetics
Statistical consistency
Gene tree
Species tree

ABSTRACT

Bayesian inference is now a leading technique for reconstructing phylogenetic trees from aligned sequence data. In this short note, we formally show that the maximum posterior tree topology provides a statistically consistent estimate of a fully resolved evolutionary tree under a wide variety of conditions. This includes the inference of gene trees from aligned sequence data across the entire parameter range of branch lengths, and under general conditions on priors in models where the usual 'identifiability' conditions hold. We extend this to the inference of species trees from sequence data, where the gene trees constitute 'nuisance parameters', as in the program *BEAST. This note also addresses earlier concerns raised in the literature questioning the extent to which statistical consistency for Bayesian methods might hold in general.

1. Introduction

Bayesian inference has become a mainstream approach for inferring phylogenetic tree topology from aligned DNA sequence data (Lemey et al., 2009). The approach has a number of desirable features, and there exist powerful software packages for analysing genetic sequence data in this way. At the same time, some potential theoretical limitations of Bayesian phylogenetics have been identified and studied. These include potential problems with the convergence of MCMC-based Bayesian methods (Mossel and Vigoda, 2005), and properties that appear to be surprising at first, such as the Bayesian star 'paradox' (Steel and Matsen, 2007; Susko, 2008; Yang, 2007).

A further property of Bayesian phylogentic inference was raised in a simulation study of Kolackzkowski and Thornton (2009), suggesting that Bayesian methods applied to unresolved four-leaf trees (with a zero-length interior edge) with certain combinations of long/short pendant branches tended to show increasing bias towards one of the three particular resolved trees as the sequence length increased. By contrast, maximum likelihood was found to favour each of the three resolutions equally. Kolackzkowski and Thornton (2009) initially suggested the possibility that for data generated by a resolved four-leaf tree with a certain combination of short and long edges, Bayesian inference might even be statistically inconsistent (i.e. the tree with the highest posterior probability for the data being different from the tree that generated the data, with a probability that does not tend to zero as the sequence length grows) even for models for which maximum likelihood is known to be statistically consistent (Chang, 1996). While Kolackzkowski and Thornton (2009) stepped back from this suggestion in a subsequent correction to their original paper, the issue drew attention to a lack of a formal proof of the statistical consistency of Bayesian inference for in molecular phylogenetics. We provide this here by establishing a more general result that includes the phylogenetic setting as a particular case.

This enhanced generality serves a further purpose, as it allows us to establish formally the statistical consistency of Bayesian species tree estimation directly from sequence data where the gene trees (and their branch lengths) are treated as further 'nuisance parameters' (as in the program *BEAST, Heled and Drummond, 2010).

While it might be possible that these results could be derived from other theoretical results in Bayesian statistics, we provide here a self-contained and essentially elementary proof that is tailored towards easy application in the phylogenetic setting. This follows the spirit of Joseph Chang's tailored version of Wald's theorem that provided a convenient tool to check and establish the consistency of maximum likelihood in phylogenetics (Chang, 1996), and which curtailed an unproductive debate in the literature about whether the detailed theoretical assumptions of Wald's original theorem applied.

2. A general result

Consider the general problem of identifying a discrete parameter lying in an arbitrary finite set A from a sequence of independent and identically distributed (i.i.d.) observations that take values in an arbitrary finite set U. Suppose further that the probability distribution on U is determined not just by the discrete parameter $a \in A$ but also by some additional (nuisance) parameters. In this paper, we will assume that these additional parameters are continuous, and we denote the parameter space associated with each discrete parameter $a \in A$ by $\Theta(a)$. We assume throughout that $\Theta(a)$ is an open subset of some Euclidean space.

In the usual phylogenetic setting, A is the set of fully resolved (binary) phylogenetic tree topologies on a given leaf set, U is the set of possible site patterns, and the parameter set $\Theta(a)$ specifies, for the tree topology a the branch lengths of the tree each of which lies in the range $(0, \infty)$, and possibly other parameters relevant to the model. Thus, if we are only concerned with branch lengths, and trees are unrooted, then $\Theta(a) = (0, \infty)^{2n-3}$ where n is the number of leaves of tree a. The trees in A may be either rooted or unrooted, and for reconstruction we estimate the same type of

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.08.012

tree (thus in the rooted case, the branch lengths are assumed to be ultrametric).

Returning to the general set-up, let $p_{(a,\theta)}$ denote the probability distribution on some finite set U determined by the discrete-continuous parameter pair (a,θ). Suppose we have a discrete (prior) probability distribution π on A, and, for each $a\in A$, a continuous (prior) probability distribution on $\Theta(a)$ with a probability density function $f_a(\theta)$. We will suppose that the following conditions hold for all $a\in A$:

(C1) $\pi(a)>0$;
(C2) the density $f_a(\theta)$ is continuous, bounded and nonzero on $\Theta(a)$;
(C3) the function $\theta\mapsto p_{(a,\theta)}(u)$ is continuous and nonzero on $\Theta(a)$ for each $u\in U$;
(C4) for all $\theta\in\Theta(a)$, and all $b\neq a$, we have: $\inf_{\theta'\in\Theta(b)} d(p_{(a,\theta)}, p_{(b,\theta')})>0$.

In (C4) and henceforth, d denotes the L_1 metric – that is, for any two probability distributions p,q on U: $d(p,q):=\sum_{u\in U}|p(u)-q(u)|$.

In the phylogenetic setting, if π is any of the usual nonzero priors on binary phylogenetic trees (e.g. the uniform 'proportional to distinguishable arrangements' or PDA distribution, or the Yule distribution), then condition (C1) is satisfied. If we take the usual exponential prior on branch lengths then condition (C2) is satisfied. For all Markov processes on trees, condition (C3) holds (the nonzero condition holds, since in any tree with pendant edges of positive lengths all site patterns have a strictly positive probability). Finally, for all models for which identifiability holds (e.g. the general time-reversible (GTR) model or any submodel down to the highly restrictive Jukes–Cantor model), condition (C4) holds (see e.g. Steel and Székely, 2009; a specific lower bound on d for the two-state symmetric model is provided via Lemma 7.3 of Steel and Székely, 2007).

Now, suppose we are given a sequence $\mathbf{u}=(u_1,\ldots,u_k)\in U^k$ generated i.i.d. by some unknown pair (a,θ) and we wish to identify the discrete parameter (a) from $\mathbf{u}$ given prior densities on A and the continuous parameters. The *maximum a-posteriori* (MAP) estimator selects the element $b\in A$ that maximizes the posterior probability of b given $\mathbf{u}$ – that is, it maximizes $\pi(b)\mathbb{E}_{\theta'}[\mathbb{P}(\mathbf{u}|b,\theta')]$, where

$$\mathbb{P}(\mathbf{u}|b,\theta')=\prod_{i=1}^{k} p_{(b,\theta')}(u_i), \tag{1}$$

which is the probability of generating the sequence of i.i.d. observations $(u_1,\ldots,u_k)$ from the underlying parameters (b,θ'), and where $\mathbb{E}_{\theta'}$ refers to taking expectation with respect to the prior probability distribution on $\Theta(b)$.

Let $P(a,\theta,k)$ denote the probability that, for a sequence $u_1,\ldots,u_k$ generated i.i.d. by (a,θ), the MAP estimator correctly selects a. The following theorem establishes a sufficient condition for the statistical consistency of the MAP estimator in this context.

Theorem 1. *Provided conditions* (C1)–(C4) *hold for all* $a\in A$, *then*

$$\lim_{k\to\infty} P(a,\theta,k)=1$$

for all $a\in A$, *and* $\theta\in\Theta(a)$.

Proof. Our proof relies on a general but technical lemma, the proof of which we defer to the Appendix. □

Lemma 2. *For any* $\epsilon_1,\epsilon_2>0$ *there exists a value* $\delta>0$ *for which the following holds*: *for any finite set* U, *and any four probability distributions* p,q,r,s *on* U *that satisfy the three conditions*:

(i) $d(p,q)\geq\epsilon_1$;
(ii) *for all* $u\in U$ *with* $r(u)>0$, $p(u)\geq\epsilon_2$ *and* $q(u)>0$;
(iii) $d(p,r)<\delta$ *and* $d(p,s)<\delta$;

the quantity $Q=\sum_{u\in U:r(u)>0} r(u)\log(s(u)/q(u))$ *is well defined* (*i.e. logarithms are applied to positive quantities*) *and* $Q\geq\frac{1}{3}\epsilon_1^2$.

2.1. Application to the proof of Theorem 1

To apply Lemma 2 we need to define the quantities mentioned by it, and we will do this in the order p,s then q,r followed by ϵ_1 and ϵ_2. Notice first that the statement of Lemma 2 is sufficiently general to allow (but not require) for q,r and s to depend on the data (i.e. to be random variables), as will be the case in our application of the lemma. This causes no problem for the argument, as we remark at the end of the proof.

We suppose throughout that the sequence $\mathbf{u}=u_1,\ldots,u_k$ is generated i.i.d. by (a,θ_0) where θ_0 is any particular element of $\Theta(a)$. Then the MAP estimator will correctly select a from $\mathbf{u}$ if and only if the *Bayes Factor* defined by

$$BF_{a/b}=\frac{\pi(a)\mathbb{E}_{\theta}[\mathbb{P}(\mathbf{u}|a,\theta)]}{\pi(b)\mathbb{E}_{\theta'}[\mathbb{P}(\mathbf{u}|b,\theta')]}$$

is strictly greater than 1 for all $b\neq a$. By the Bonferroni inequality, it suffices to show that for each $b\neq a$ the probability that $\mathbf{u}$ is such that $BF_{a/b}>1$ tends to 1 as k grows. To achieve this we first observe that $BF_{a/b}=(\pi(a)/\pi(b))\cdot R_{a/b}$ where

$$R_{a/b}:=\frac{\mathbb{E}_{\theta}[\mathbb{P}(\mathbf{u}|a,\theta)]}{\mathbb{E}_{\theta'}[\mathbb{P}(\mathbf{u}|b,\theta')]}, \tag{2}$$

and where $\pi(a)/\pi(b)$ is finite and strictly positive by (C1). Thus, it suffices to show that, for each $b\neq a$ and for every finite constant M, the inequality $R_{a/b}>M$ holds with a probability that tends to 1 as $k\to\infty$. We will establish this inequality by providing an explicit lower bound to the numerator of $R_{a/b}$ and an explicit upper bound to the denominator of $R_{a/b}$, and showing that, with probability tending to 1 as k grows, their ratio exceeds M.

Before describing the lower bound, observe that we can re-write Eq. (1) as follows:

$$\mathbb{P}(\mathbf{u}|b,\theta)=\prod_{u\in U} p_{(b,\theta)}(u)^{n_u}, \tag{3}$$

where, for each $u\in U$,

$n_u:=|\{i: u_i=u\}|$.

For the lower bound on the numerator of $R_{a/b}$, consider the subset N_τ of $\Theta(a)$ consisting of a closed ball centered on θ_0 and of radius $\tau>0$. Note that we can always select a sufficiently small value of $\tau>0$ for which $N_\tau\subset\Theta(a)$ by the assumption that $\Theta(a)$ is an open subset of some Euclidean space. Letting $\mu(N_\tau)=\int_{N_\tau} f_a(\theta)\,d\theta>0$ we have

$$\mathbb{E}_{\theta}[\mathbb{P}(\mathbf{u}|a,\theta)]=\int_{\Theta(a)}\mathbb{P}(\mathbf{u}|a,\theta)f_a(\theta)\,d\theta\geq\int_{N_\tau}\mathbb{P}(\mathbf{u}|a,\theta)f_a(\theta)\,d\theta,$$

and so

$$\mathbb{E}_{\theta}[\mathbb{P}(\mathbf{u}|a,\theta)]\geq\mu(N_\tau)\cdot\inf_{\theta\in N_\tau}\{\mathbb{P}(\mathbf{u}|a,\theta)\}. \tag{4}$$

2.2. Lower bound and the distributions p and s

Let $p=p_{(a,\theta_0)}$ (the generating probability distribution on the true parameters) and let s be the probability distribution of the form $p_{(a,\theta)}$ that minimizes $\mathbb{P}(\mathbf{u}|a,\theta)$ when θ is restricted to N_τ; such a distribution s exists from the compactness of N_τ and the continuity condition of (C3). Then, from (3) we have: $\inf_{\theta\in N_\tau}\{\mathbb{P}(\mathbf{u}|a,\theta)\}=\prod_{u\in U}s(u)^{n_u}$. Applying this to (4) gives

$$\mathbb{E}_{\theta}[\mathbb{P}(\mathbf{u}|a,\theta)]\geq\mu(N_\tau)\cdot\prod_{u\in U}s(u)^{n_u}. \tag{5}$$