A CONSISTENCY LEMMA IN STATISTICAL PHYLOGENETICS

MIKE STEEL

Abstract. This short note provides a simple formal proof of a folklore result in statistical phylogenetics concerning the convergence of bootstrap support for a tree and its edges.

1. Definitions and preliminaries

In this note T will refer to any rooted or unrooted phylogenetic tree, and $T^{-\rho}$ will refer to the unrooted tree obtained from T by suppressing the root vertex ρ if it has one (i.e. if T is unrooted then $T^{-\rho} = T$). Let θ be a vector of continuous parameters – including the branch lengths of T, along with possibly other continuous parameters required to specify a model of character evolution on T. Let Θ denote the set of values θ may take. Branch lengths, in particular, are assumed to be strictly positive and finite; and in general Θ will be some open subset of Euclidean space. Consider any stochastic process (e.g. Markov process, or mixture of Markov processes) which assigns to each pair (T, θ) a probability distribution $\mathbf{s} = \mathbf{s}(T, \theta)$ on discrete, finite-state characters at the tips of the tree. We assume throughout that the map $\theta \mapsto \mathbf{s}(T, \theta)$ is continuous. Such models are central to statistical phylogenetics and methods for reconstructing phylogenetic trees from aligned genetic (e.g. DNA) sequences. A *tree reconstruction method* ψ is any method that reconstructs a set of one or more unrooted phylogenetic trees from any given distribution $\hat{\mathbf{f}}$ of site pattern frequencies. Suppose we generate k sites i.i.d. from (T, θ), and let $\hat{\mathbf{s}}$ be the random variable equal to the resulting proportion of site patterns (character types). The method ψ is a *statistically consistent* estimator of the unrooted topology of T if the probability that $\psi(\hat{\mathbf{s}}) = \{T^{-\rho}\}$ converges to 1 as $k \to \infty$[1]. Suppose that ψ satisfies the following condition:

(*) For every tree T for which $T^{-\rho}$ is fully-resolved (i.e. binary), and each $\theta \in \Theta(T)$ a value $\epsilon = \epsilon_{(T,\theta)} > 0$ exists for which the following inequality holds for every probability distribution $\hat{\mathbf{f}}$ on site patterns: $\|\hat{\mathbf{f}} - \mathbf{s}(T, \theta)\| < \epsilon \Rightarrow \psi(\hat{\mathbf{f}}) = \{T^{-\rho}\}$.

Here $\|\cdot\|$ denotes any of the usual norms in Euclidean space. Condition (*) implies the statistical consistency of ψ for inferring $T^{-\rho}$ since the i.i.d. assumption ensures that $\hat{\mathbf{s}}$ converges in probability to $\mathbf{s}(T, \theta)$ as k grows, and so:

$$\mathbb{P}(\psi(\hat{\mathbf{s}}) = \{T^{-\rho}\}) \geq \mathbb{P}(\|\hat{\mathbf{s}} - \mathbf{s}(T, \theta))\| < \epsilon_{(T,\theta)}) \to 1, \text{ as } k \to \infty.$$

Not only does condition (*) imply that $\psi(\mathbf{s}(T, \theta)) = \{T^{-\rho}\}$ whenever $T^{-\rho}$ is fully-resolved but (*) also implies the stronger condition that for any tree T' that has a different unrooted topology (fully-resolved or non-fully-resolved) from the fully-resolved tree T we have:

$$\inf_{\theta' \in \Theta(T')} \|\mathbf{s}(T, \theta) - \mathbf{s}(T', \theta')\| > 0, \tag{1}$$

a strong 'identifiablity' condition, referred to as 'no touching' in [3].

Condition (*) is a type of local stability condition. It applies, for example, to distance-based tree reconstruction applied to (statistically consistent) 'corrected distances' derived from the

Date: January 26, 2015.

[1]There is a slightly stronger definition involving almost sure convergence rather than convergence in probability, and the results here can be extended to that setting also.

characters, provided that the distance-reconstruction method has a positive 'safety radius', which holds for many (but not all) distance-based methods, including the popular Neighbor-Joining method [1]. Condition (*) also applies to MLE (maximum likelihood estimation) for models which satisfy (1) – such models include the general time-reversible (GTR) Markov processes and its submodels (e.g. Jukes-Cantor type models) and certain extensions of these models. Here MLE treats θ as 'nuisance parameters' to be optimized as part of the search for the MLE tree; given a vector $\hat{\mathbf{f}}$ as input, MLE selects the tree(s) T' maximizing $\sup_{\theta \in \Theta(T')} \mathbb{P}(\hat{\mathbf{f}}|\mathbf{s}(T', \theta))$. The proof that Condition (*) holds for models satisfying (1) follows from standard analytic arguments based on the continuity of the map $\theta \mapsto \mathbb{P}(\hat{\mathbf{f}}|\mathbf{s}(T', \theta))$ (see e.g. [2] or [3]).

2. Result

Given $\hat{\mathbf{s}}$ derived from k i.i.d. site patterns, let $\hat{\mathbf{s}}^*$ denote the frequency of site patterns obtained by taking an i.i.d. sample of k site patterns using probability distribution $\hat{\mathbf{s}}$. Thus $\hat{\mathbf{s}}^*$ is the distribution of site patterns in a bootstrap sample from the original data. The *bootstrap support of an edge* e of an unrooted phylogenetic tree T', is the expected proportion of such bootstrap samples for which a tree, sampled uniformly at random from $\psi(\hat{\mathbf{s}}^*)$, has an edge that induces the same split of the leaf taxa as e does in T' (it is a random variable by its dependence on $\hat{\mathbf{s}}$, and since ψ can return more than one tree). The *bootstrap support for* T' is the random variable $\mathbb{P}(\psi(\hat{\mathbf{s}}^*) = \{T'\}|\hat{\mathbf{s}})$, the expected proportion of bootstrap samples for which ψ returns the single tree T'. The following result was motivated by a question from T. Warnow (pers. comm.).

Lemma 1. *Suppose k sites are generated i.i.d. by $\mathbf{s}(T, \theta)$. Under the sufficient condition (*) for statistical consistency, the bootstrap support of every edge e of $T^{-\rho}$ converges in probability to 1 as $k \to \infty$. Moreover, the bootstrap support for $T^{-\rho}$ converges in probability to 1 as $k \to \infty$.*

Proof. Clearly it suffices to prove the second assertion in the lemma, since, by definition, the bootstrap support for any edge e of $T^{-\rho}$ is at least $\mathbb{P}(\psi(\hat{\mathbf{s}}^*) = \{T^{-\rho}\}|\hat{\mathbf{s}})$. Let $X = X(\hat{\mathbf{s}})$ be the 0/1 random variable which takes the value 1 precisely if $\psi(\hat{\mathbf{s}}^*) = \{T^{-\rho}\}$, and which is 0 otherwise. Let Y denote the expected bootstrap support for $T^{-\rho}$ given $\hat{\mathbf{s}}$; thus $Y = \mathbb{P}(\psi(\hat{\mathbf{s}}^*) = \{T^{-\rho}\}|\hat{\mathbf{s}}) = \mathbb{E}[X|\hat{\mathbf{s}}]$ (i.e. the conditional expectation of X given $\hat{\mathbf{s}}$). Notice that:

$$\mathbb{E}[Y] = \mathbb{E}[E[X|\hat{\mathbf{s}}]] = \mathbb{E}[X] = \mathbb{P}(\psi(\hat{\mathbf{s}}^*) = \{T^{-\rho}\}). \tag{2}$$

Now, as k grows, $\hat{\mathbf{s}} \xrightarrow{p} \mathbf{s}$, and $\hat{\mathbf{s}}^* - \hat{\mathbf{s}} \xrightarrow{p} \mathbf{0}$; thus $\hat{\mathbf{s}}^* \xrightarrow{p} \mathbf{s}$. Consequently, by Condition (*), $\mathbb{P}(\psi(\hat{\mathbf{s}}^*) = \{T^{-\rho}\})$ converges to 1 as $k \to \infty$, and so, by (2), $\lim_{k\to\infty} \mathbb{E}[Y] = 1$. Finally, since Y takes values in the interval $[0, 1]$, and the expected value of Y converges to 1 as $k \to \infty$, it follows that (for the bootstrap support for $T^{-\rho}$) we have $Y \xrightarrow{p} 1$ as $k \to \infty$, as required. □

Note that the empirical bootstrap support for an edge (or for a tree) given $\hat{\mathbf{s}}$, converges in probability to the (expected) bootstrap support value defined here, as the number N of independent bootstrap replicates becomes large; hence our results are also relevant for empirical bootstrap support for large N.

References

[1] Atteson, K. (1999). The performance of neighbor-joining methods of phylogeny reconstruction. *Algorithmica* 25(2–3): 251–278.

[2] Chang, J.T. (1996). Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.* 137: 51–73.

[3] Steel, M. (2011). Can we avoid 'SIN' in the House of 'No Common Mechanism'? *Systematic Biology* 60(1): 96–109.

Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand.
E-mail address: mike.steel@canterbury.ac.nz