A consistent estimator of the evolutionary rate

Krzysztof Bartoszek [a], Serik Sagitov [b],*

[a] Department of Mathematics, Uppsala University, 751 06, Uppsala, Sweden
[b] Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, 412 96 Gothenburg, Sweden

HIGHLIGHTS

- Estimation of the evolutionary rate without knowing the phylogenetic tree.
- We study a tree-free evolutionary model based on the Yule process and Brownian motion.
- We compute the asymptotic variance of the trait disparity for phylogenetically correlated samples.

ARTICLE INFO

ABSTRACT

We consider a branching particle system where particles reproduce according to the pure birth Yule process with the birth rate λ, conditioned on the observed number of particles to be equal to n. Particles are assumed to move independently on the real line according to the Brownian motion with the local variance σ^2. In this paper we treat n particles as a sample of related species. The spatial Brownian motion of a particle describes the development of a trait value of interest (e.g. log-body-size). We propose an unbiased estimator R_n^2 of the evolutionary rate $\rho^2 = \sigma^2/\lambda$. The estimator R_n^2 is proportional to the sample variance S_n^2 computed from n trait values. We find an approximate formula for the standard error of R_n^2 based on a neat asymptotic relation for the variance of S_n^2.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Biodiversity within a group of n related species could be quantified by comparing suitable trait values. For some key trait values like log-body-size, researchers apply the Brownian motion model proposed by Felsenstein (1985). It is assumed that the current trait values $(X_1^{(n)}, \ldots, X_n^{(n)})$ have evolved from the common ancestral state X_0 as a branching Brownian motion with the local variance σ^2. Given a phylogenetic tree describing the ancestral history of the group of species the Brownian trajectories of the trait values for sister species are assumed to evolve independently after the ancestor species splits in two daughter species. The resulting phylogenetic sample $(X_1^{(n)}, \ldots, X_n^{(n)})$ consists of identically distributed normal random variables with a dependence structure caused by the underlying phylogenetic signal.

A mathematically appealing and biologically motivated version of the phylogenetic sample model assumes that the phylogenetic

tree behind the normally distributed trait values $(X_1^{(n)}, \ldots, X_n^{(n)})$ is unknown. As a natural first choice to model the unknown species tree, we use the Yule process with birth rate λ (see Yule, 1924). Since the phylogenetic sample size is given, n, the Yule process should be conditioned on having n tips: such conditioned branching processes have received significant attention in recent years, due to e.g. Aldous and Popovic (2005), Gernhard (2008), Mooers et al. (2012), Stadler (2009, 2011), and Stadler and Steel (2012). This "tree-free" approach for phylogenetic comparative methods was previously addressed by Sagitov and Bartoszek (2012), Crawford and Suchard (2013) and Mulder and Crawford (2015) (much earlier Edwards, 1970 used a related branching Brownian process as a population genetics model).

In our work we show that a properly scaled sample variance is an unbiased and consistent estimator of the compound parameter $\rho^2 = \sigma^2/\lambda$ which we call the evolutionary rate of the trait value in question. Our main mathematical result, Theorem 2.1, gives an asymptotic expression for the variance of the phylogenetic sample variance. This result leads to a simple asymptotic formula for the estimated standard error of our estimator. Our result is in agreement with the work of Crawford and Suchard (2013) whose simulations indicate that their approximate maximum likelihood

procedure yields an unbiased consistent estimator of σ^2. This is illustrated using the example of the Carnivora order studied previously by Crawford and Suchard (2013).

The phenotype modelled by a Brownian motion is usually interpreted as the case of neutral evolution with a random noise around the ancestral state. This model was later developed into an adaptive evolutionary model based on the Ornstein–Uhlenbeck process by Felsenstein (1988), Hansen (1997), Butler and King (2004), Hansen et al. (2008), and Bartoszek et al. (2012). The tree-free setting using the Ornstein–Uhlenbeck process was addressed by Bartoszek and Sagitov (2015) where for the Yule–Ornstein–Uhlenbeck model, some phylogenetic confidence intervals for the optimal trait value were obtained via three limit theorems for the phylogenetic sample mean. Furthermore, it was shown that the phylogenetic sample variance is an unbiased consistent estimator of the stationary variance of the process.

At the end of their discussion Crawford and Suchard (2013) write that as the tree of life is refined interest in "tree-free" estimation methods may diminish. They however indicate that "tree-free" estimates may be useful to calculate starting points for simulation analysis. We certainly agree with the second statement but believe that development of "tree-free" methods should proceed alongside that of "tree-based" ones.

One of the most useful features of the "tree-free" comparative models is that they offer a natural method of tree growth allowing for study of theoretical properties of phylogenetic comparative models. This is a field receiving more and more attention in recent years (e.g. Gascuel and Steel, 2014; Ho and Ané, 2014a,b). A fundamental setup to consider is the pure-birth Yule tree. Statistical results for processes evolving on top of such a tree is developed in this work (and also Sagitov and Bartoszek, 2012; Crawford and Suchard, 2013; Bartoszek, 2014; Mossel and Steel, 2014; Bartoszek and Sagitov, 2015). Another alternative to studying estimators of parameters of such processes is the tree growth model proposed by Ané (2008), Ho and Ané (2013), and Ané et al. (2014). In this setup the total height of the tree is kept fixed and new tips are added to randomly chosen branches (however Ané et al., 2014 also discuss more general setups). These two approaches seem to be in agreement, at least up to the second moments, since e.g. they agree on the lack of consistency of estimating X_0. In Sagitov and Bartoszek (2012) we showed that under the Yule Brownian motion model $\mathrm{Var}[\overline{X}_n] \to 2\sigma^2$.

In a practical situation "tree-free" methods can be used for a number of purposes. Firstly as pointed out by Crawford and Suchard (2013) they can be useful for calculating starting points for further numerical estimation procedures or defining prior distributions in a Bayesian setting. Secondly they have to be used in a situation where the tree is actually unknown, e.g. when we are studying fossil data or nearly very difficult to infer. For example, the phylogenies of insects are characterized by multiple recent radiations. In such a situation the support of many internal nodes can be close to zero—we do not have enough signal to place them even though we do not suspect polytomies. Piwczyński et al.'s (2014) molecular phylogeny of flesh flies can serve as a model example—their Bayesian analysis could not converge, while maximum likelihood trees had very low support for many internal nodes. Thirdly "tree-free" methods can be useful for making predictive statements about future phenotypes, e.g. development of viruses. Finally they can be used for various sanity checks. If they contradict "tree-based" results this could indicate that the numerical method fell into a local maximum.

The paper has the following structure. Section 2 presents the model, the main results and an application. In Section 3 we state two lemmas characterizing the coalescent time of a Yule tree needed for the proof of Proposition 4.1. Section 4 states two further lemmas and a proposition directly yielding the assertion of Theorem 2.1. Proposition 4.1 deals with the covariances between coalescent times for randomly chosen pairs of tips from a random Yule n-tree.

2. The main results

The basic evolutionary model considered in this paper is characterized by four parameters $(\lambda, n, X_0, \sigma^2)$ and consists of two stochastic components: a random phylogenetic tree defined by parameters (λ, n) and a trait evolution process along a lineage defined by parameters (X_0, σ^2). The first component, species tree connecting n extant species, is modelled by the pure birth Yule process (Yule, 1924) with the birth (speciation) rate λ and conditioned on having n tips (Gernhard, 2008). For the second component we adapt the approach by assuming that for a given $i = 1, \ldots, n$, the current trait value $X_i^{(n)}$ has evolved from the
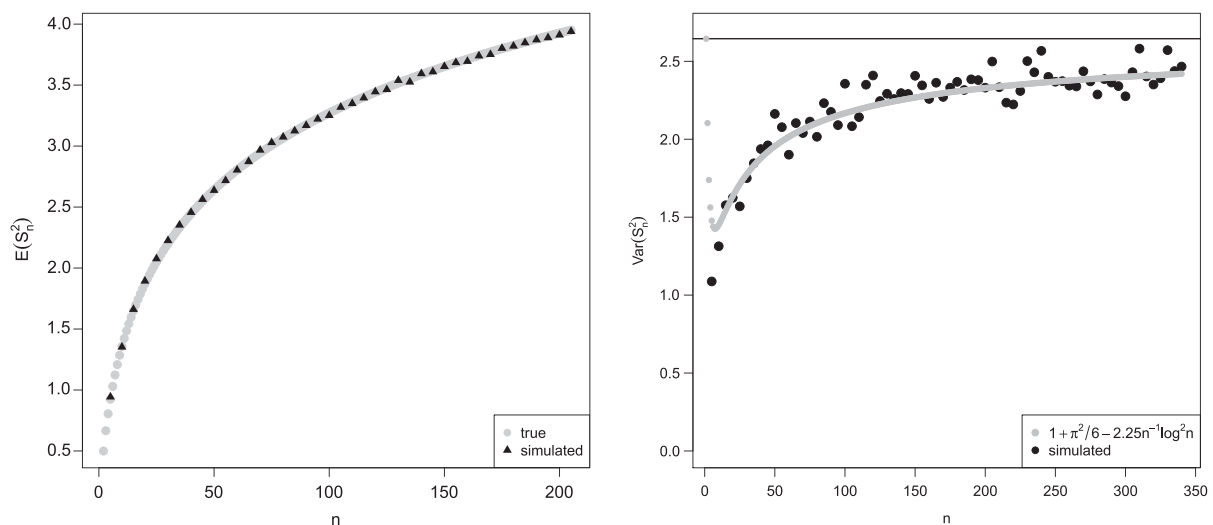

Fig. 1. Left: True and simulated values of $E\left[S_n^2\right]$, right: simulated values of $\mathrm{Var}[S_n^2]$ with limit equalling $\pi^2/6+1$. Each point comes from 10 000 simulated Yule trees and Brownian motions on top of them. Parameters used in simulations are $\lambda=1$, $X_0=0$ and $\sigma^2=1$. The grey line on the right panel fits a curve based on the convergence rate $O(n^{-1}\log n^2)$.