Received: 22 August 2019 | Revised: 23 December 2019 | Accepted: 17 January 2020

DOI: 10.1111/stan.12205

SPECIAL ISSUE ARTICLE

WILEY

Introducing a general class of species diversification models for phylogenetic trees

Francisco Richter[1,2] | **Bart Haegeman**[3] | **Rampal S. Etienne**[2] | **Ernst C. Wit**[1,4]

[1]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Groningen, The Netherlands

[2]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands

[3]Theoretical and Experimental Ecology Station, CNRS and Paul Sabatier University, Toulouse, France

[4]Institute of Computational Science, Università della Svizzera italiana (USI), Lugano, Switzerland

Correspondence
*Francisco Richter, Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Groningen, The Netherlands.
Email: f.richter@rug.nl

Phylogenetic trees are types of networks that describe the temporal relationship between individuals, species, or other units that are subject to evolutionary diversification. Many phylogenetic trees are constructed from molecular data that is often only available for extant species, and hence they lack all or some of the branches that did not make it into the present. This feature makes inference on the diversification process challenging. For relatively simple diversification models, analytical or numerical methods to compute the likelihood exist, but these do not work for more realistic models in which the likelihood depends on properties of the missing lineages. In this article, we study a general class of species diversification models, and we provide an expectation-maximization framework in combination with a uniform sampling scheme to perform maximum likelihood estimation of the parameters of the diversification process.

KEYWORDS

EM algorithm, generalized linear models, importance sampling, nonhomogeneous Poisson process, phylogenetic trees

1 | INTRODUCTION

Evolutionary relationships of species are commonly described by phylogenetic trees or, in more general scenarios, by phylogenetic networks (Ragan, 2009). A phylogenetic tree is a hypothesis on how species or other biological units have diversified over time. It is usually described by a binary tree whose nodes are ordered in time. Phylogenetic relationships can be inferred from a variety of sources such as morphology and behaviors of species, biochemical pathways, DNA, and protein sequences (Lemey, Salemi, & Vandamme, 2009), both from extant, that is, living species or from extinct species through ancient DNA or the fossil record. However, data on extinct species are often incomplete and only accurate molecular phylogenies of extant species are available. In this article, we consider such phylogenetic trees as primary observations. Even though they lack extinct lineages, they are believed to contain information on how species diversified and hence they have been used to answer fundamental questions, such as "does diversity affect diversification?" (Cornell, 2013; Etienne et al., 2012), "what is the effect of environmental and ecological interactions on evolutionary dynamics?" (Barraclough, 2015; Ezard, Aze, Pearson, & Purvis, 2011; Lewitus & Morlon, 2017), "how does biodiversity vary spatially?" (Goldberg, Lancaster, & Ree, 2011; Mittelbach et al., 2007), and "what traits play a key role in species diversification?" (FitzJohn, Maddison, & Otto, 2009; Lynch, 2009; Paradis, 2005), to name just a few.

To help to answer these questions, specific mathematical models have been developed that can infer various parameters from phylogenetic diversification pattern (Morlon, 2014). Most current approaches have started to use likelihood-based methods to perform inference on phylogenetic trees (Etienne et al., 2012; FitzJohn et al., 2009; Ricklefs, 2007; Stadler, 2011). Although statistically principled, in each of these models, a new method to compute the likelihood needs to be developed. These models often rely on describing the macroevolutionary process by coupled ordinary differential equations—the so-called *master* or *Kolmogorov equations*—and these quickly become intractable as model complexity increases, particularly due to the lack of data on extinct species (Höhna, Stadler, Ronquist, & Britton, 2011; Ricklefs, 2007).

Alternative ways to deal with Kolmogorov equations have been used since the 1950s in fields outside evolutionary biology. These methods have used point process theory (Daley & Vere-Jones, 2007; Serfozo, 1990), which does not solve Kolmogorov equations directly but employs Gillespie-type simulations that were introduced in the context of chemical reaction modeling (Gillespie, 1976, 1977). A single Gillespie simulation represents an exact sample from the probability mass function that is the solution of the system, thus allowing for stochastic optimization methods to maximize the likelihood (Tijms, 1994).

In this article, we present a first step for a general inference procedure of a general species diversification model. In Section 2, we describe a general diversification process based on a generalized linear model (GLM) description of a nonhomogeneous point process. This model can be used to describe many alternative evolutionary hypotheses. In Section 3, we introduce an expectation-maximization (EM) algorithm to optimize the likelihood under incomplete information, namely, the extinct lineages. We present a data augmentation algorithm, involving stochastic simulation combined with an importance sampler, to perform the E-step. We provide a proof-of-concept by comparing our inference with that obtained using direct likelihood calculations. In Section 4, we apply our method to the diversification of a small clade of Vangidae, consisting of a group of medium-sized birds living in Madagascar. Our aim is to discover

Rampal Etienne and Ernst C. Wit jointly contribute to last authorship.

whether the evolutionary record supports more the diversity dependence hypothesis (Etienne et al., 2012) or the phylodiversity hypothesis (Castillo, Verdú, & Valiente-Banuet, 2010), for which no direct likelihood computation exists. Finally in Section 5, we provide directions for future extensions of the method that are needed to allow evolutionary biologists to routinely apply our approach to larger phylogenetic trees to study general diversification dynamics in a unified framework.

2 | A GENERAL DIVERSIFICATION MODEL

We define a phylogenetic tree $x = (\tau, t, a)$ on a time interval $[0, T]$ as a functional object described by three components: a binary vector τ of event types (speciation or extinction), a vector of continuous event times t, and a network configuration object a, describing which species speciated or went extinct at each event time. We model the shape and structure of the tree by means of a collection of point processes, in this case, a set of dynamical nonhomogeneous Poisson processes (NHPP) where speciation and extinction of species are random events that happen within a time interval $[0, T]$. Figure 1 shows an example of a phylogenetic tree with three speciation events and one extinction event.

In this article, we assume that the process starts at time $t_0 = 0$ with a single species b_1. At this stage, the tree is subject to two Poisson processes: a potential speciation of species b_1 and a potential extinction of species b_1. Both processes are assumed to have a waiting time with time-continuous rates $\lambda_{b_1}(t)$ and $\mu_{b_1}(t)$, respectively. In the time-homogeneous case, the waiting time for the first event to occur is therefore an exponential with rate $\lambda_{b_1} + \mu_{b_1}$. More in general (Daley & Vere-Jones, 2007), the probability density for the process x to have a single species up tox time t_1 and a speciation event exactly at time t_1 is given by

$$f(t_1) = \lambda_{b_1}(t_1) e^{-\int_{t_0}^{t_1} \lambda_{b_1}(t) + \mu_{b_1}(t) dt}. \tag{1}$$

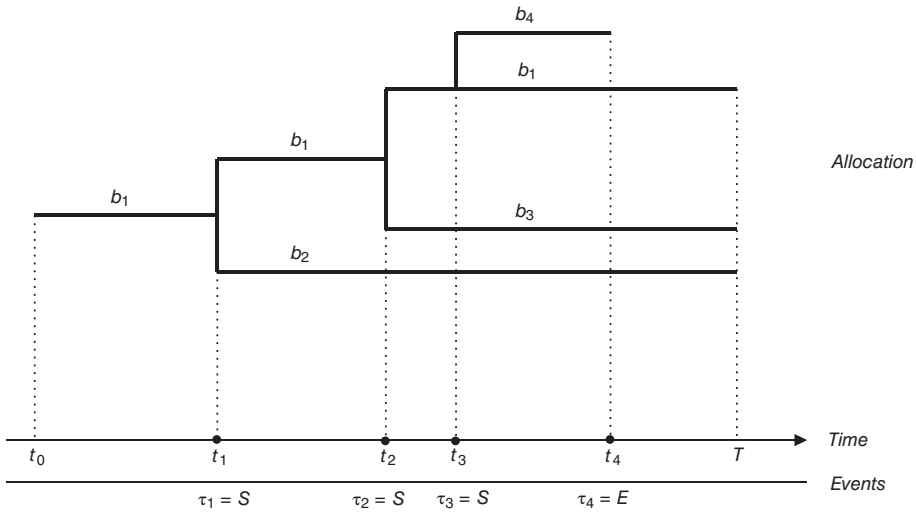


FIGURE 1 Phylogenetic tree with four events: three speciation events and one extinction event. Each branch represents a species

If indeed a speciation occurs, the process continues with four NHPPs: two potential speciations and two potential extinctions. This is repeated until the present time T, unless the tree dies out before then. We consider a general scenario where at time t each of the N_t present species b has its own speciation rate $\lambda_b(t)$ and extinction rate $\mu_b(t)$ defined as a linear function via link function h,

$$h(\lambda_b(t)) = \sum_{j=1}^{m} \beta_j c_{bjt}, \quad h(\mu_b(t)) = \sum_{j=1}^{m} \alpha_j c_{bjt}. \tag{2}$$

where c_{bjt} is one of $j = 1, \dots, m$ possible covariates of species b at time t affecting the speciation and/or extinction processes. Our entire process is therefore governed by the parameter set $\theta = \{\beta_1, \dots, \beta_m, \alpha_1, \dots, \alpha_m\}$. Typically, we will consider the logarithmic link function $h = \log$, but Equation (2) can be trivially modified by choosing for h any monotonous increasing function that maps $(0, \infty)$ onto $\mathbb{R}$. The class of statistical models satisfying these specifications are an extension of the well-known GLMs (Dobson & Barnett, 2008).

This GLM extension to phylogenetic trees spans a very broad spectrum of possibilities for evolutionary biologists to test hypotheses and integrate their species diversification data. Diversification rates can be influenced by individual attributes, typically called *traits*, environmental factors, such as average temperature, by the composition of the diversifying clade itself or of its local ecological community. In the literature, a range of models have been explored, where diversification rates are assumed to be constant (Nee, May, & Harvey, 1994), change through time (Rabosky & Lovette, 2008), depend on diversity (Etienne et al., 2012), on individual traits (Freckleton, Phillimore, & Pagel, 2008; Paradis, 2005) or other factors (Morlon, 2014). In order to test realistic models, we are interested in flexible rates that are able to change dynamically through all those factors simultaneously. For example, the speciation rate of species b at time t could also depend on other species' traits.

Mathematically, the method allows the inclusion of any set of covariates that might be interesting to incorporate for evolutionary biologists; however, full information on individual covariates, such as traits, are rarely available—especially not on the missing species. One way to deal with this is by including an extra augmentation step and simulating full information of traits on augmented trees (Hoehna et al., 2019). Another option is to use observable proxies related to, for example, trait diversity, such as different forms of phylogenetic diversity. These present interesting direction for future work.

3 | MLE INFERENCE WITH MCEM USING IMPORTANCE SAMPLING

The loglikelihood of a full tree including extinct branches $x \in \mathcal{X}$ involving a total of M events by extrapolating from (1) can easily be shown to be given by

$$\ell_x(\theta) = \sum_{i=1}^{M} \sum_{b=1}^{N_{t_i}} \left[\log \left[\lambda_b(t_i;\theta) \mathbb{1}_{Sp}(t_i, b) + \mu_b(t_i;\theta) \mathbb{1}_{Ex}(t_i, b) \right] - \int_{t_{i-1}}^{t_i} \lambda_b(t;\theta) + \mu_b(t;\theta)\, dt \right] \tag{3}$$

where $\mathbb{1}_{\mathrm{Sp}}(t_i, b) = 1$ if species b speciates at time t_i, 0 otherwise and $\mathbb{1}_{\mathrm{Ex}}(t_i, b) = 1$ if species b becomes extinct at time t_i, 0 otherwise. An additional term $-\sum_{b=1}^{N_{t_M}} \int_{t_M}^{T} \lambda_b(t;\theta) + \mu_b(t;\theta)\, dt$ has to

be added to the likelihood, if the final event time t_M does not correspond to the present T. For the case when diversification rates are stepwise constant, this reduces to the solutions in Wrenn (2012) and Reynolds (1973). When the full phylogenetic tree and the covariates at all times are given, we can directly maximize the loglikelihood function (3) to obtain the maximum likelihood estimates of the parameters (Paradis, 2005) and perform model selection to determine what factors are important for diversification. In practice, however, we almost never observe the full phylogenetic tree, but only a tree with the extant species.

3.1 | Difficulties of MLE estimation and an MCEM algorithm

Let us denote $\mathcal{Y}$ as the space of ultrametric trees (Gavryushkin & Drummond, 2016), that is, time-calibrated trees without extinct lineages and $\mathcal{X}(y)$ as the space of all full trees that, when pruning all extinct species, lead to the ultrametric tree $y \in \mathcal{Y}$. Then the log likelihood of an observed, extant species only tree y is given by the integral of the likelihood (3) over all possible full trees,

$$\ell_y(\theta) = \log \int_{\mathcal{X}(y)} \exp\left(\ell_x(\theta)\right)\, dx. \tag{4}$$

However, because of the complexity of the space $\mathcal{X}(y)$ a closed-form solution for Equation (4) is not available in most cases (Gavryushkin, Whidden, & Matsen, 2016), making inference, or in particular, direct MLE computations difficult or impossible.

A typical method for likelihood maximization under incomplete data is the application of the EM algorithm (Dempster, Laird, & Rubin, 1977), considering the information about the extinct species as a missing data problem. In the EM algorithm, a sequence $\{\theta^{(s)}\}$ of parameter values are generated by iterating the following two steps:

E-step Compute the conditional expectation $Q(\theta|\theta^{(s)}) = E_{\theta^{(s)}}(\ell_X(\theta)|Y = y)$.
M-step Choose $\theta^{(s+1)}$ to be the value of $\theta \in \Omega$ which maximizes $Q(\theta|\theta^{(s)})$.

This algorithm is run iteratively until convergence is reached. Under certain regularity conditions (Dempster et al., 1977), the point of convergence can be shown to be the MLE for the incomplete data problem, that is, maximizing $\ell_y(\theta)$.

As in the case of Equation (4), the calculation of $Q(\theta|\theta^*)$ does not have a closed-form due to the complexity of the space $\mathcal{X}(y)$, so approximations are needed. To perform this task, we use Monte Carlo integration (Wei & Tanner, 1990), where given a set of sampled trees $x_1, \ldots, x_p$ from an importance sampler distribution $g(x|y, \theta)$ we approximate $Q(\theta|\theta^*)$ by

$$\begin{aligned} Q(\theta|\theta^*) &\approx \frac{1}{p}\sum_{i=1}^{p} \ell_{x_i}(\theta)\frac{f_{X|Y}(x_i|y,\theta^*)}{g_{X|Y}(x_i|y,\theta^*)} \\ &\propto \frac{1}{p}\sum_{i=1}^{p} w_i \ell_{x_i}(\theta), \end{aligned} \tag{5}$$

where the importance weights are defined as $w_i = \frac{f_{X,Y}(x_i,y|\theta^*)}{g_{X|Y}(x_i|y,\theta^*)}$, using the law of conditional probabilities to obtain the proportional expression.

In the M-step, we optimize (5) via numerical methods and the Hessian is calculated and represents the Fisher information matrix H^{-1}. Assuming that the errors given by the EM algorithm are independent of the Monte Carlo errors, the standard errors for the MCEM algorithm are defined as

$$\mathrm{SE}(\hat{\theta}_i) = \sqrt{-H^{-1}_{i,i} + \frac{\mathrm{VMCE}}{N_{\mathrm{EM}}}}, \tag{6}$$

where $-H^{-1}_{i,i}$ corresponds to the diagonal components of the information matrix giving the EM error, VMCE is the variance of the MC error, and N_{EM} is the number of MCEM iterations considered for estimation. Note that if the EM is run long enough, the second term in (6) goes to zero, making the information matrix the decisive value for standard errors on MCEM algorithm (McLachlan & Krishnan, 2007). With the standard errors we can construct confidence intervals for the parameters and test hypotheses about the significance of covariates of interest.

3.2 | A simple importance sampler

To sample trees we propose a tree augmentation algorithm that samples independently the three components of the tree: event types, event times, and species allocations. The algorithm is shown in Figure 2.

3.2.1 | Step 1. Generate event times and number of extinctions

The number of extinct species d and $2d$ missing event times, that is, speciations and extinctions of these d missing species are sampled uniformly in the following manner:

1. Sample the number of missing species d uniformly from the discrete space $\{0, \dots, M^e\}$, where M^e is a predefined ceiling, such that the probability of more than M^e extinctions is extremely unlikely.
2. Sample $2d$ branching times uniformly from the continuous space $(0, T]$ and then sort them.

The probability of sampling a set of $2d$ unobserved event times $t^e = (t^e_1, \dots, t^e_{2d})$ for a tree of dimension d is

$$g_{\text{event times}}(d, t^e) = \frac{1}{M^e + 1}\left(\frac{1}{T}\right)^{2d}(2d)!$$

Note that this scheme samples the dimension of the tree uniformly, but the size of the space of trees grows in a factorial way with the dimension of the tree. This means that the sample size required to obtain a robust Monte Carlo approximation of the integral (4) must be large. This is a limitation of this importance sampler, and hence it is only reliable when many extinctions are unlikely.

3.2.2 | Step 2. Generate event types

We simulate a binary event chain $\tau^e = (\tau^e_1, \dots, \tau^e_{2d})$ assigning either S (speciation) or E (extinction) to each event time. This chain is subject to the rule that the number of Es up to any point in

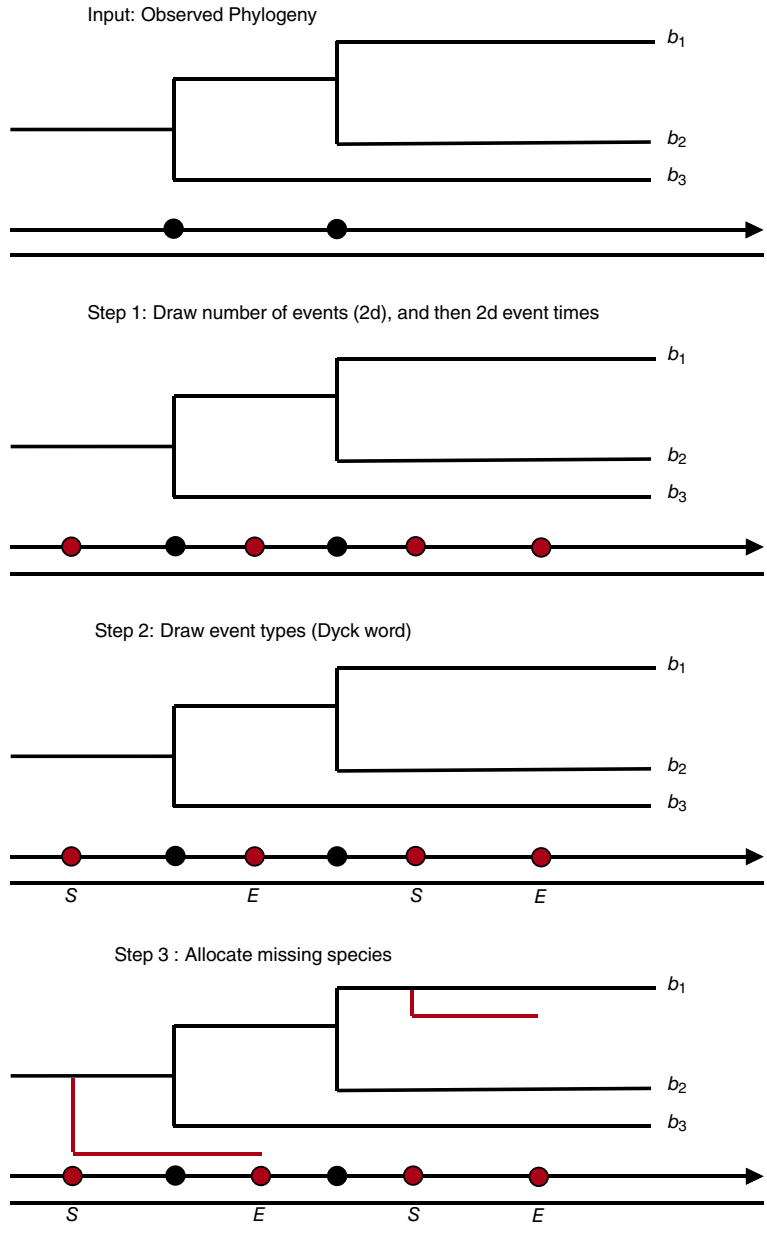


FIGURE 2 The three components of our phylogenetic tree augmentation algorithm

the chain should be less than or equal to the number of Ss in the chain up to that point. The set of allowed chains is known in the mathematical literature as the set of Dyck words and several methods for sampling Dyck words have been developed (Kasa, 2010). Furthermore, given a number of events $2d$, the number of possible Dyck words is known as the Catalan number (Zvonkin, 2014),

$$C_d = \binom{2d}{d} \frac{1}{d+1}.$$

By uniformly sampling a Dyck word τ^e of length $2d$, the probability of a specific event sequence is given by $g_{\text{events}}(\tau^e) = 1/C_d$.

3.2.3 | Step 3. Species allocation

Given the missing event times and missing event types we can perform the tree allocations by sampling a parent species of each missing speciation and by defining which species, that is, the parent species or the inserted "new species," becomes extinct at the extinction event. To sample uniformly we just need to count the number of possible trees in agreement with the event times $t^e = (t_1^e, \dots, t_{2d}^e)$ and event types. This number, $n(\tau_{2d}^e, t_{2d}^e)$, can be calculated by starting with $n(\tau_0^e, t_0^e) = 1$ and applying the following rules when going from root to tips in the phylogenetic tree:

- For each unobserved speciation event at t_i^e, that is, $\tau_i^e = S$, update $n(\tau_i^e, t_i^e)$ in the following way,

$$n(\tau_i^e = S, t_i^e) = n(\tau_{i-1}^e, t_{i-1}^e) \times \left(2N_{t_i^-}^o + N_{t_i^-}^e\right),$$

  where $N_{t^-}^o$ is the number of observed branches just before t and $N_{t^-}^e$ is the number of unobserved branches just before t. Note that events on observed branches count twice compared with those on unobserved branches. Intuitively, this accounts for the two eventualities following an unobserved speciation on an observed branch: either the first or the second daughter species is observed (the other one is unobserved), while for a speciation on an unobserved branch, both daughter species are unobserved. A more formal argument justifying the factor of two is provided by Laudanno, Haegeman, and Etienne (2019).
- For each unobserved extinction event at t_i^e, that is, $\tau_i^e = E$, update $n(\tau_i^e, t_i^e)$ in the following way,

$$n(\tau_i^e = E, t_i^e) = n(\tau_{i-1}^e, t_{i-1}^e) \times N_{t_i^-}^e.$$

As we sample uniformly, the probability for each possible allocation a^e of the d missing species at the missing event times t^e with Dyck word τ^e in the tree of extant species x_{obs} is then given by $g_{\text{allocation}}(a^e) = \frac{1}{n(\tau_{2d}^e, t_{2d}^e)}$.

3.2.4 | Sampling probability of a uniformly augmented tree

The uniform sampling probability of the augmented tree $x_{\text{unobs}} = (d, t^e, \tau^e, a^e)$ is then given by

$$g(x_{\text{unobs}}|x_{\text{obs}}, \theta) = \frac{1}{M^e + 1}\left(\frac{1}{T}\right)^{2d}(2d)!\frac{1}{C_d}\frac{1}{n(\tau_{2d}^e, t_{2d}^e)}. \tag{7}$$

From this equation, we can see how the dimension of the tree space plays an important role. For this reason, the uniform importance sampler becomes less efficient when many extinctions are likely. On the other hand, the uniform sampling scheme allows for easy implementation and quick computation, thereby making it suitable as a default sampler.

3.3 | Checking performance by comparing with direct ML

To show that the MCEM works, we compared our method to the linear diversity-dependence (LDD) diversification model for which the likelihood can be calculated directly (Etienne et al.,

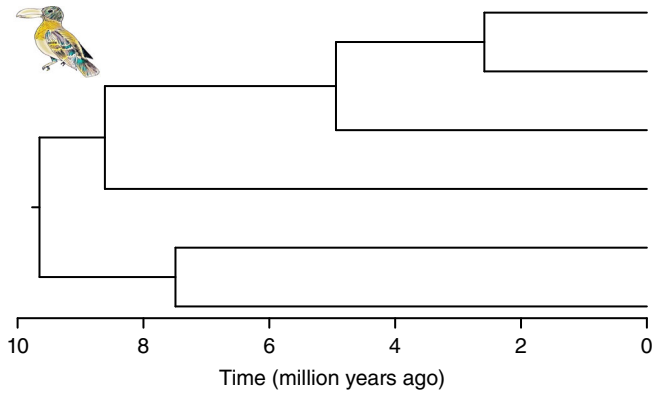


FIGURE 3 Subclade of the Malagassy Vangidae, obtained from Jønsson et al. (2012)

TABLE 1 MCEM estimation for three different samples sizes, with two replicates each

$\overline{\text{ESS}}$	Replicate	$\hat{\theta}_1$	SE($\hat{\theta}_1$)	$\hat{\theta}_2$	SE($\hat{\theta}_2$)	$\hat{\theta}_3$	SE($\hat{\theta}_3$)
37	1	1.403	0.077	−0.257	0.016	0.032	0.026
37	2	1.359	0.077	−0.249	0.016	0.031	0.026
373	3	1.709	0.098	−0.307	0.020	0.046	0.031
372	4	1.713	0.098	−0.309	0.020	0.046	0.031
2970	5	1.932	0.127	−0.336	0.026	0.056	0.033
2987	6	1.892	0.121	−0.328	0.025	0.056	0.033
MLE		1.937		−0.326		0.060	

Note: The first column is the mean of the effective sample size over the 1,000 iterations considered. The last row is the MLE directly calculated by computing the likelihood (Etienne et al., 2012). Estimated values are for the linear DD model with $\theta_1 = \lambda_0$, $\theta_2 = (\mu_0 - \lambda_0)/K$ and $\theta_3 = \mu_0$.

2012). In this model, speciation rates depend on diversity of the phylogenetic tree at that point. We consider the diversification model with rates

$$\lambda_b(t) = \lambda_0 - (\lambda_0 - \mu_0)\frac{N_t}{K}, \quad \mu_b(t) = \mu_0$$

where N_t is the number of extant species (diversity) at time t and $\theta = \{\lambda_0, \frac{\mu_0 - \lambda_0}{K}, \mu_0\}$ are model parameters. This model is a special case of our general modeling framework, defined in (2). We perform the MCEM routine on a clade of Malagassy birds, the so-called Vangidae clade shown in Figure 3, which has been analyzed in Jønsson et al. (2012). We replicated the routine several times with different sample sizes to observe the impact of sample size on estimation and the robustness of the method.

In Table 1 we show six replicates corresponding to three pairs with different sample size orders. We drop the first 1,000 iterations as burn-in, and use the next 1,000 MCEM iterations for parameters estimation, reporting the mean value and the standard error from Equation (6). We observe that for small sample sizes (Replicates 1 and 2), the estimation is poor. For the cheapest setup, the mean effective sample size (ESS) is approximately 37 and this does not seem enough to sample in spaces with a substantial number of missing species. In this scenario, the MCEM estimates are not robust. As sample size increases, we see that inference becomes more and more accurate and matches the MLE procedure by Etienne et al. (2012).

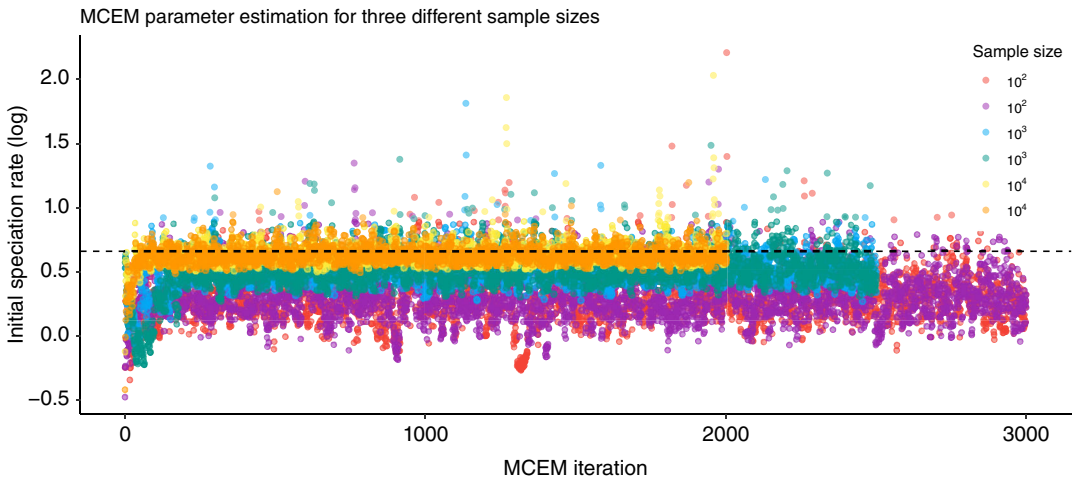


FIGURE 4 MCEM applied to the tree of Figure 3 under the LDD diversification model. Evolution of the estimate of the first parameter, the initial speciation rate $\theta_1 = \lambda_0$ through EM iterations. We plot six replicates: two for three different sample sizes. For better visualization we cut higher sample sizes at iteration 2,000 and 2,500

These replicates are also summarized in Figure 4, where we show a visualization of the dynamical MCEM parameter estimation for $\log \lambda_0$ corresponding to the logarithm of the initial speciation rate at stem age. The dashed black line indicates the true MLE. We see in all six cases that estimations go quickly to the true MLE with a stable behavior after a couple of 100 iterations. To visually compare biases and variation through different sample sizes, we show the replicates for small sample sizes until the 2,000th and 2,500th MCEM iteration. We clearly see that for higher sample sizes, bias and variation decrease.

Note that the ESS is between 30% and 40% in these cases. An efficient importance sampler with 100% ESS is a priority for future publications in order to apply the method to larger phylogenetic trees.

4 | DIVERSITY-DEPENDENCE: DIVERSITY OR PHYLODIVERSITY?

Phylodiversity is defined as the total branch length of extant species of a tree, and it has been proposed as an alternative to diversity in conservation ecology (Faith, 1992). Figure 5 shows phylodiversity and diversity through time for a simple example tree.

As an illustration of the flexibility of our method, we now consider a model similar to diversity-dependence introduced in the previous section, but with dependence on phylodiversity P_t instead of N_t. Diversity-dependence has been detected in a Vangidae clade (Jønsson et al., 2012) and we would like to extend the analysis to check if phylodiversity-dependence (LPD) is a more suitable factor in diversification of Vangidae than diversity-dependence (LDD). In addition to these two models, which both assume linear dependence of speciation rate on diversity or phylodiversity, we consider the exponential diversity dependence (EDD) and exponential phylogenetic diversity (EPD) models. The exponential models use the log-link function common in the statistical literature, rather than the identity link suggested by the evolutionary biology literature. Table 2 shows the parameter definitions for the four models tested on the phylogenetic tree of the Vangidae.

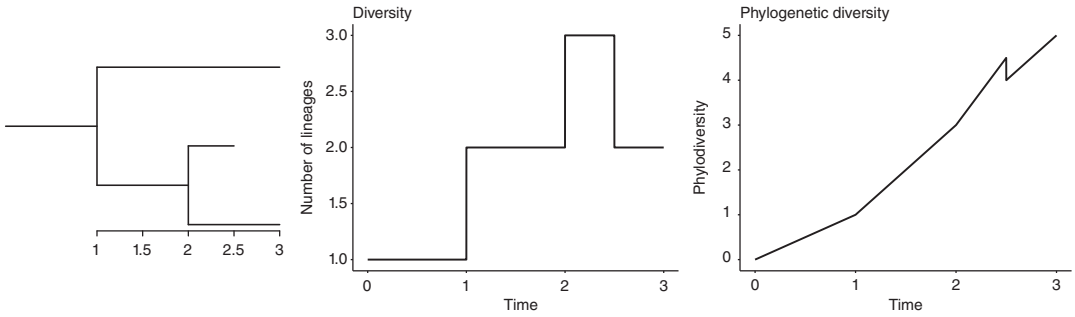


FIGURE 5 Example of a simple tree with one extinction. The two panels on the right show the difference between diversity and phylodiversity through time

TABLE 2 Four diversity-dependent diversification models, where speciation rate depends on diversity or phylodiversity, either linearly or exponentially

Model	$\lambda_b(t)$	θ_1	θ_2	θ_3
LDD	$\lambda_0 - (\lambda_0 - \mu_0)\frac{N_t}{K}$	λ_0	$-(\lambda_0 - \mu_0)\frac{1}{K}$	μ_0
LPD	$\lambda_0 - (\lambda_0 - \mu_0)\frac{P_t}{K}$	λ_0	$-(\lambda_0 - \mu_0)\frac{1}{K}$	μ_0
EDD	$\lambda_0 e^{-aN_t}$	$\ln(\lambda_0)$	$-a$	μ_0
EPD	$\lambda_0 e^{-aP_t}$	$\ln(\lambda_0)$	$-a$	μ_0

Note: All models assume constant extinction rate and have three parameters to be estimated.
Abbreviations: EPD, exponential phylogenetic diversity; LDD, linear diversity-dependence; LPD, linear phylodiversity-dependence.

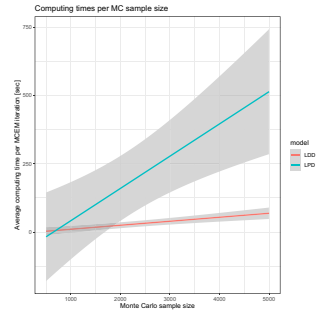
We performed the MCEM routine for each of the four diversification models, obtaining the ML estimates of the parameters and calculating Monte Carlo estimation for the likelihood function and the corresponding AIC values (Wit, Heuvel, & Romeijn, 2012). Interestingly, we found that phylodiversity models do not performs better than ordinary diversity models, but there is an improvement of the exponential diversity-dependence model over the linear DD model. Table 3 shows the inference results for each of the four diversification models.

To get an idea of the computational cost of the method we include, next to Table 3, a plot of computing times (for one MCEM iteration) as a function of Monte Carlo sample size for PD and DD models starting at their respective MLE values reported in the table. The values are average of 100 replicates performed in an ordinary computer. From the plot we can see that for our example tree, each iteration takes a couple of minutes for large Monte Carlo sample size, which means that the whole routine should take few hours at most. We also see that the computing times increases linearly with the MC sample size.

We conclude that the best model in this analysis is an EDD model with parameters $\theta_1 = 2.58(0.96), \theta_2 = -1.02(0.25), \theta_3 = 0.04(0.03)$, suggesting an exponential decreasing speciation rate with an exponential decay constant close to 1, given by θ_2. We found an initial speciation rate of approximately 4.85 species per million years which decreases until 0.03 at the present time. This indeed suggests that the diversification process of this Vangidae clade in Madagascar has slowed down dramatically over the past 10 million years. Moreover, the extinction rate of 0.04 species per million years suggests that the clade has now reached a stable diversification behavior, whereby any further speciations will tend to be offset by extinctions.

TABLE 3 Parameter estimation of the four diversity-dependent models of Table 2 when applied to the Vanga tree of Figure 3, including Monte Carlo approximations of the loglikelihood and AIC

Model	θ_1	θ_2	θ_3	Loglikelihood	AIC
LDD	1.94	−0.33	0.06	−11.36	28.72
LPD	0.31	−0.01	0.04	−14.37	34.74
EDD	2.58	−1.02	0.04	−11.19	28.37
EPD	−0.28	−0.04	0.13	−13.44	32.89



Note: Next to the table we see the plot of average computing times per MCEM iteration (in seconds) for DD and PD models at their respective MLE.
Abbreviations: EPD, exponential phylogenetic diversity; LDD, linear diversity-dependence; LPD, linear phylodiversity-dependence.

5 | DISCUSSION

We have presented a flexible method for testing a broad variety of diversification models in phylogenetic analysis and provided some simple examples. This is a first step toward a robust general methodology to identify potential factors in diversification processes from phylogenetic trees.

The unobserved extinct species turn the inference problem naturally into a problem that can be approached by means of an EM algorithm. Given the complexity of the E-step, a Monte Carlo importance sampler has been proposed, involving a uniform importance sampler. Given the computational simplicity both in terms of sampling and calculation of uniform samplers, this may be a convenient option for small sized trees, where more sophisticated importance samplers, involving the underlying nonhomogenous Poisson processes, would not necessarily improve efficiency. As in the case of Vangidae clade where a few missing species are likely, we found that the uniform importance sampler leads to accurate estimation. However, the performance of our uniform importance sampler deteriorates as the dimension of the phylogenetic tree increases. In order to apply this method on high-dimensional trees, a more efficient importance sampler should be carefully chosen. This we will leave for future work.

Current approaches perform inference by means of likelihood maximization, which requires that formulas for the likelihood must be derived on a case-by-case basis. Here, we consider a general class of models that include an augmentation step inside an EM algorithm, thereby avoiding direct likelihood calculation and thus allowing inference for a wide variety of diversification models.

In principle, in cases when full information of covariates is still missing after the augmentation step, extensions of the augmentation procedure are possible. However, this is beyond the scope of the current article.

Moreover, to increase efficiency alternatives to MCEM algorithms may be considered, such as the stochastic approximation version of the EM algorithm (SAEM; Delyon, Lavielle, & Moulines, 1999) or a Bayesian approach (Richardson & Green, 1997). In both cases, the algorithm could make use of the previous MC samples, thereby improving efficiency at some computational cost.

Even though in this article we only refer to the context of a diversification process of ecological species, a phylogenetic tree is used in many other fields to describe other kinds of processes, such as language evolution (Greenhill, Atkinson, Meade, & Gray, 2010) and cultural diversification

(Mace & Holden, 2005). Therefore, the method that we have developed in this article is potentially useful for inferring the underlying driving process of such branching processes.

ACKNOWLEDGEMENTS

This work is part of the research program Mathematics for Planet Earth with project number 657.014.005, which is financed by the Dutch Research Council (NWO). F.R. and E.W. would also like to acknowledge the contribution of the COST Action CA15109.

ORCID

Francisco Richter https://orcid.org/0000-0002-0924-4613
Ernst C. Wit https://orcid.org/0000-0002-3671-9610

REFERENCES

Barraclough, T. G. (2015). How do species interactions affect evolutionary dynamics across whole communities? *Annual Review of Ecology, Evolution, and Systematics*, *46*, 25–48.

Castillo, J. P., Verdú, M., & Valiente-Banuet, A. (2010). Neighborhood phylodiversity affects plant performance. *Ecology*, *91*, 3656–3663.

Cornell, H. V. (2013). Is regional species diversity bounded or unbounded? *Biological Reviews*, *88*, 140–165.

Daley, D. J., & Vere-Jones, D. (2007). *An introduction to the theory of point processes: Volume II: General theory and structure*. Berlin, Germany: Springer Science & Business Media.

Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, *27*, 94–128.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B Methodological*, *39*(1), 1–38.

Dobson, A. J., & Barnett, A. (2008). *An introduction to generalized linear models*. Boca Raton, FL: CRC Press.

Etienne, R. S., Haegeman, B., Stadler, T., Aze, T., Pearson, P. N., Purvis, A., & Phillimore, A. B. (2012). Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1732), 1300–1309.

Ezard, T. H., Aze, T., Pearson, P. N., & Purvis, A. (2011). Interplay between changing climate and species ecology drives macroevolutionary dynamics. *Science*, *332*, 349–351.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, *61*, 1–10.

FitzJohn, R. G., Maddison, W. P., & Otto, S. P. (2009). Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic Biology*, *58*, 595–611.

Freckleton, R. P., Phillimore, A. B., & Pagel, M. (2008). Relating traits to diversification: A simple test. *The American Naturalist*, *172*, 102–115.

Gavryushkin, A., & Drummond, A. J. (2016). The space of ultrametric phylogenetic trees. *Journal of Theoretical Biology*, *403*, 197–208.

Gavryushkin, A., Whidden, C., & Matsen, F. (2016). *The combinatorics of discrete time-trees: Theory and open problems*. bioRxiv, 063362.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, *22*, 403–434.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, *81*, 2340–2361.

Goldberg, E. E., Lancaster, L. T., & Ree, R. H. (2011). Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. *Syst Biol*, *60*, 451–465.

Greenhill, S. J., Atkinson, Q. D., Meade, A., & Gray, R. D. (2010). The shape and tempo of language evolution. *Proceedings of the Royal Society B: Biological Sciences*, *277*, 2443–2450.

Hoehna, S., Freyman, W. A., Nolen, Z., Huelsenbeck, J., May, M. R., & Moore, B. R. (2019). A Bayesian approach for estimating branch-specific speciation and extinction rates. *bioRxiv*, 555805.

Höhna, S., Stadler, T., Ronquist, F., & Britton, T. (2011). Inferring speciation and extinction rates under different sampling schemes. *Molecular Biology and Evolution*, *28*, 2577–2589.

Jønsson, K. A., Fabre, P.-H., Fritz, S. A., Etienne, R. S., Ricklefs, R. E., Jørgensen, T. B., … Irestedt, M. (2012). Ecological and evolutionary determinants for the adaptive radiation of the madagascan vangas. *Proceedings of the National Academy of Sciences*, *109*, 6620–6625.

Kasa, Z. (2010) Generating and ranking of Dyck words. arXiv preprint arXiv:1002.2625.

Laudanno, G., Haegeman, B., & Etienne, R. S. (2019). Additional analytical support for a new method to compute the likelihood of diversification models. bioRxiv, 693176.

Lemey, P., Salemi, M., & Vandamme, A.-M. (2009). *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing*. Cambridge, MA: Cambridge University Press.

Lewitus, E., & Morlon, H. (2017). Detecting environment-dependent diversification from phylogenies: A simulation study and some empirical illustrations. *Systematic Biology*, *67*, 576–593.

Lynch, V. J. (2009). Live-birth in vipers (viperidae) is a key innovation and adaptation to global cooling during the cenozoic. *Evolution: International Journal of Organic Evolution*, *63*, 2457–2465.

Mace, R., & Holden, C. J. (2005). A phylogenetic approach to cultural evolution. *Trends in Ecology & Evolution*, *20*, 116–121.

McLachlan, G., & Krishnan, T. (2007). *The EM algorithm and extensions* (Vol. *382*). New York, NY: John Wiley & Sons.

Mittelbach, G. G., Schemske, D. W., Cornell, H. V., Allen, A. P., Brown, J. M., Bush, M. B., … Sax, D. F. (2007). Evolution and the latitudinal diversity gradient: Speciation, extinction and biogeography. *Ecology Letters*, *10*, 315–331.

Morlon, H. (2014). Phylogenetic approaches for studying diversification. *Ecology Letters*, *17*, 508–525.

Nee, S., May, R. M., & Harvey, P. H. (1994). The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *344*, 305–311.

Paradis, E. (2005). Statistical analysis of diversification with species traits. *Evolution*, *59*, 1–12.

Rabosky, D. L., & Lovette, I. J. (2008). Explosive evolutionary radiations: Decreasing speciation or increasing extinction through time? *Evolution*, *62*, 1866–1875.

Ragan, M. A. (2009). Trees and networks before and after Darwin. *Biology Direct*, *4*, 43.

Reynolds, J. F. (1973). On estimating the parameters of a birth-death process. *Australian Journal of Statistics*, *15*, 35–43.

Richardson, S., & Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B Statistical Methodology*, *59*, 731–792.

Ricklefs, R. E. (2007). Estimating diversification rates from phylogenetic information. *Trends in Ecology & Evolution*, *22*, 601–610.

Serfozo, R. F. (1990). Point processes. *Handbooks in Operations Research and Management Science*, *2*, 1–93.

Stadler, T. (2011). Inferring speciation and extinction processes from extant species data. *Proceedings of the National Academy of Sciences*, *108*, 16145–16146.

Tijms, H. C. (1994). *Stochastic models: An algorithmic approach* (Vol. *994*). Chichester, England: John Wiley & Sons.

Wei, G. C., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*, 699–704.

Wit, E., Heuvel, E. V. D., & Romeijn, J.-W. (2012). All models are wrong: An introduction to model uncertainty. *Statistica Neerlandica*, *66*, 217–236.

Wrenn, F. (2012). *General birth-death processes: Probabilities, inference, and applications* (Doctoral dissertation). UCLA.

Zvonkin, A. K. (2014). Enumeration of weighted plane trees. arXiv preprint arXiv:1404.4836.

How to cite this article: Richter F, Haegeman B, Etienne RS, Wit EC. Introducing a general class of species diversification models for phylogenetic trees. *Statistica Neerlandica*. 2020;74:261–274. https://doi.org/10.1111/stan.12205