University of Groningen

Inferring the drivers of species diversification

Richter Mendoza, Francisco

DOI:
10.33612/diss.167307789

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Richter Mendoza, F. (2021). *Inferring the drivers of species diversification: Using statistical network science*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen. https://doi.org/10.33612/diss.167307789

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 30-11-2023

4

LINEAGE-DEPENDENT PHYLOGENETIC DIVERSITY AS A DRIVER OF SPECIES DIVERSIFICATION

The 'art' of building a good model is to capture the essential features of the biology without burdening the model with non-essential details.

Darren J. Wilkinson

Richter F, Wit EC, Etienne RS. In preparation.

ABSTRACT

Modelling species diversification processes and performing statistical inference on these processes using phylogenetic trees is an active area of research. It requires the development of novel quantitative tools to study the influence of ecological factors on (macro)evolutionary processes. The Yule model or constant-rate birth-death models are still widely used, because of their simplicity that allows fast computation of the likelihood, i.e., the probability of the phylogenetic tree given the diversification model parameters.

The development of more complex species diversification models that consider additional factors typically involves the computation of the likelihood for the diversification model, via the master equation. These more complex models consider species interactions and need to integrate across all such interactions, due to the lack of information about these interactions in the past — which are unlikely to ever become available.

A promising alternative is to use a proxy for (past) species interactions. Species diversity is one such proxy, and it has indeed been possible to compute the likelihood for models in which diversification rates depend on diversity. However, this proxy assumes that all species interact in the same way. To accommodate variation in these interactions, we propose to use phylogenetic diversity as a proxy, because phylogenetic diversity between species, defined as the time to the most common ancestor, represents the niche distance among species.

In this chapter, we integrate per-species phylogenetic distance into diversity-dependent diversification models, the results of which we will call lineage-dependent diversification models. We show that these models cover a broad range of topologies, consistent with those of real trees. In addition, we develop a stochastic gradient descent framework that will enable parameter estimation for these models. In summary, with the minimal modification of phylodiversity dependence we expand diversity-dependent diversification models to represent a much broader range of models that can mimic complex topological characteristics of phylogenetic trees.

4.1. INTRODUCTION

Studying the mechanisms underlying species diversification has been an active area of research (Ragan, 2009) and particularly during the last three decades since large-scale DNA sequencing and phylogenetic analysis has been possible (Reynolds, 1973; Nee *et al.*, 1994; Morlon, 2014). Larger and more accurate phylogenies continue to appear (Jetz *et al.*, 2012; Upham *et al.*, 2019; Ramírez-Barahona *et al.*, 2020; Condamine *et al.*, 2019; Hedges *et al.*, 2015b) and species diversification models are more and more sophisticated in order to capture and study multiple hypotheses on how species diversified (Morlon, 2014; Ricklefs, 2007; Etienne and Apol, 2009). In 1925 Yule published a mathematical characterisation of a process where species diversifies with a constant rate without extinction (Yule, 1925). In 1948, Kendall generalised Yule's results by allowing for extinction and time dependent speciation and extinction rates (Kendall *et al.*, 1948). In 1994 Nee et al. presented the likelihood for the time-dependent birth-death process given a phylogenetic tree (Nee *et al.*, 1994), which typically does not contain extinct species. In the last 20 years, a large number of species diversification models have been developed, including diversity-dependent (Etienne *et al.*, 2012b), state-dependent (Maddison *et al.*, 2007; Herrera-Alsina *et al.*, 2019; FitzJohn *et al.*, 2009; Paradis, 2008; Ng and Smith, 2014), and (paleo-)environment-dependent (Condamine *et al.*, 2019; Lewitus and Morlon, 2017) diversification rates. Still, these models have only scratched the surface of all possible diversification processes and more inference methods are needed (Rabosky and Goldberg, 2015).

Maximum likelihood approaches have become a standard to compare various macroevolutionary scenarios using reconstructed phylogenies (Nee, 2006), even though this comparison may have limitations on identifiability (Louca and Pennell, 2020). The design of stochastic birth-death-type species diversification models (SDM) lends itself well for easy testing of hypotheses. Within SDM we can identify two nested classes of models. One class considers global diversification rates, i.e. all species have the same probability to speciate or become extinct. A more general class of SDM considers diversification rates that can differ between species. We will call these models lineage-dependent diversification (LDD) models. Models that assume a global rate for all lineages (lineage-independent diversification (LID) models) are by far the most used and are generally assumed to be a good starting point for analysis. Current LDD diversification models range from simply assuming a shift in the rates (Laudanno *et al.*, 2020c; Rabosky, 2014; Höhna *et al.*, 2019; Maliet *et al.*, 2019), or dependence on a dynamic state (Maddison *et al.*, 2007).

Despite the development of sophisticated (LDD) models, simple constant-rate birth-death models are still commonly used, even though their predictions on temporal (Phillimore and Price, 2008) and topological (Heard, 1996; Mooers *et al.*, 2007; Purvis *et al.*, 2011; Shao, 1990) properties deviate from those in empirical phylogenies. One of the reasons for the limited use of LDD models is that likelihood calculation is much more complicated (Laudanno *et al.*, 2020b) than for lineage-independent diversification (LID) models. However, LDD models are the next generation models that are needed to incorporate more complex ecological interactions, such as niche differentiation and/or facilitative interactions (Barraclough, 2015; Fox, 2005; Olave *et al.*, 2020; Bairey *et al.*, 2016; Roy *et al.*, 2020). Current diversity-dependent diversification models consider such interactions by

simply accounting for the role of species richness on diversification, and for more than a decade diversity-dependence models have already been extensively used in macroevolutionary analysis, detecting clade-level "carrying capacities" and studying the influence of species richness on macroevolutionary processes. In the previous chapter, we generalised diversity-dependent diversification models by allowing diversification rates to depend on phylogenetic diversity, and hence not only the number of species but also their distinctiveness is taken into account. However, this model still assumes that all species are equally likely to diversify, and is therefore an LID model. Current inference procedures mostly use the branching times of the trees as their only input. With LDD models we can take into account topology as well. Per-species phylogenetic distance, defined as the time of the most common ancestor among two species, has not been included in phylogenetic analysis for macroevolutionary studies while it serves as one of the most common proxies for ecological similarities.

In this manuscript, we present and study a LDD model, the lineage-dependent phylodiversity-dependent (LDPD) model and study its effect on macroevolutionary processes. First, in Section 4.2, we discuss the relationship between tree shape and evolutionary advantage, the state of the art of LDD models, and current challenges and we propose a general LDD model that satisfies several desired biological and mathematical properties that makes it a powerful tool for quantitative macroecological and phylogenetic analysis. Then, in Section 4.3, we introduce the lineage-dependent phylodiversity-dependent models and analyse how these models can help capture proper tree shapes and balance levels observed in current phylogenies. We describe the Phylodiversity Matrix, as a dynamical matrix that captures the genetic distance among pairs of species. Finally, in section 4.4 we provide a methodology (stochastic gradient descent) for parameter estimation and derive the required equations for the LDPD model. We use the data augmentation algorithm introduced in the previous chapter to approximate the gradient of the likelihood. We discuss potential directions, advantages and limitations of the method.

4.2. MODE AND TEMPO IN EVOLUTIONARY PROCESSES AND REAL PHYLOGENIES

Mathematically the diversification process is characterised by two components, time and balance. Biologically, they represent the tempo and mode of the macro-evolutionary dynamics. Several statistics or measurements have been designed to describe both components. The gamma index describes the distribution of the waiting times between events throughout the process (Pybus and Harvey, 2000; Fordyce, 2010). It is especially useful to capture diversification rate decreases compared with higher rates in the past. The ρ-metric introduced in Pigot *et al.* (2010) is an alternative to the gamma-statistic providing values between -1 and 1 indicating speedup in speciation rates towards 1. Regarding the topology of the tree, the Colless index (Colless, 1982) is probably the most used statistic for characterising tree balance; however, dozens of other indices to summarise the shape of the tree have been developed. One example is the Sackin's index (M. Coronado *et al.*, 2020), which computes the sum of the number of ancestors for each tip of the tree; Another example is the Cophenetic index that computes the sum of the

depths of the least common ancestor of every pair of leaves of the tree. The Cophenetic index is a proportion of the MPD index in (Mazel *et al.*, 2016), which calculates the mean pairwise distance between all species in the clade, that is, the mean of the per-species phylogenetic diversity. In this manuscript, we focus on Gamma and Colless, but our analysis can be easily replicated with other metrics.

In figure 4.1, we show the distribution of tempo and mode of 64 Mammal phylogenies represented by the Gamma (GI) and Colless index (CI). We use a normalised version of the CI (PDA normalisation), where high values represent highly unbalanced trees and values close to zero represent highly balanced trees. We use the R packages `ape` and `treeshape` to calculate gamma and Colless indices respectively.

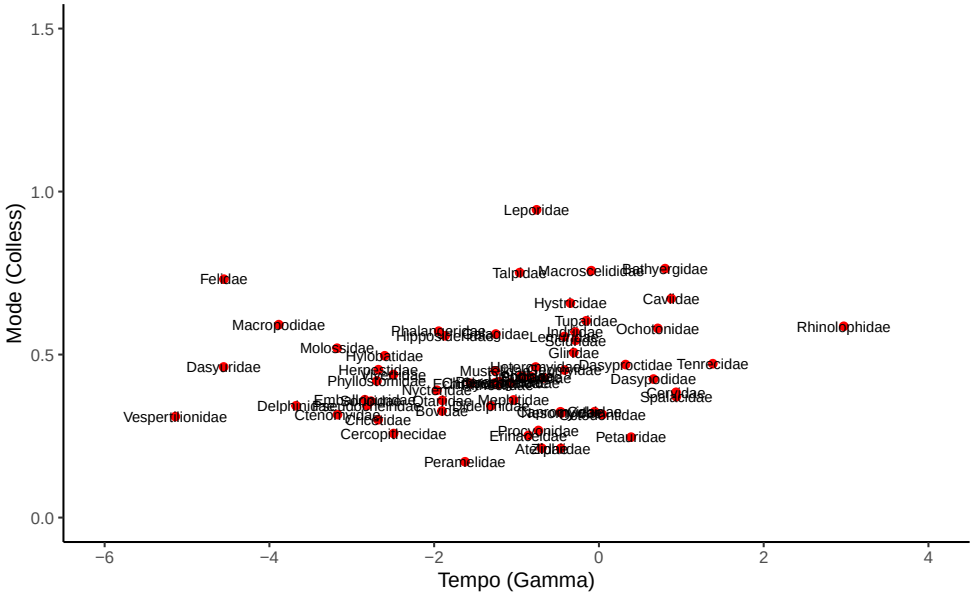


Figure 4.1 | Distribution of tempo and mode for 64 mammal phylogenies, characterised by the Gamma and Colless index.

Many studies have analysed and reported gamma and Colless indexes for standard species diversification models (Mooers and Heard, 1997), but this has not been done extensively for diversity-dependent diversification models. We performed a simulation study to study how well such models capture these tree features. We used the standard linear diversity dependence model (LDD) (Etienne *et al.*, 2012b) for four sets of parameters. In Figure 4.2, we plot the four extreme phylogenies with maximum and minimum indices. To be consistent with observed phylogenies, we computed the MLE for the four extreme phylogenies described above, and used these to simulate phylogenies under the LDD model and computed the Gamma and Colless distributions. Figure 4.3 shows that both extremes in the mode (Colless) and one in the tempo (gamma) cannot be mimicked by the simulations.

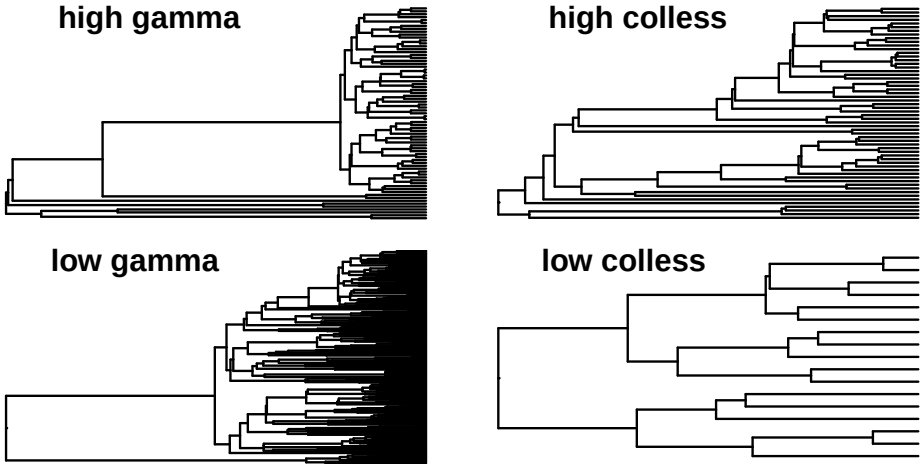


Figure 4.2 | Example of the 4 most extreme mammal phylogenies. Rhinolophidae is the clade with maximum gamma index, Leporidae is the clade with maximum Colless index, Vespertilionidae is the clade with minimum gamma index and Peramelidae is the clade with minimum Colless index.

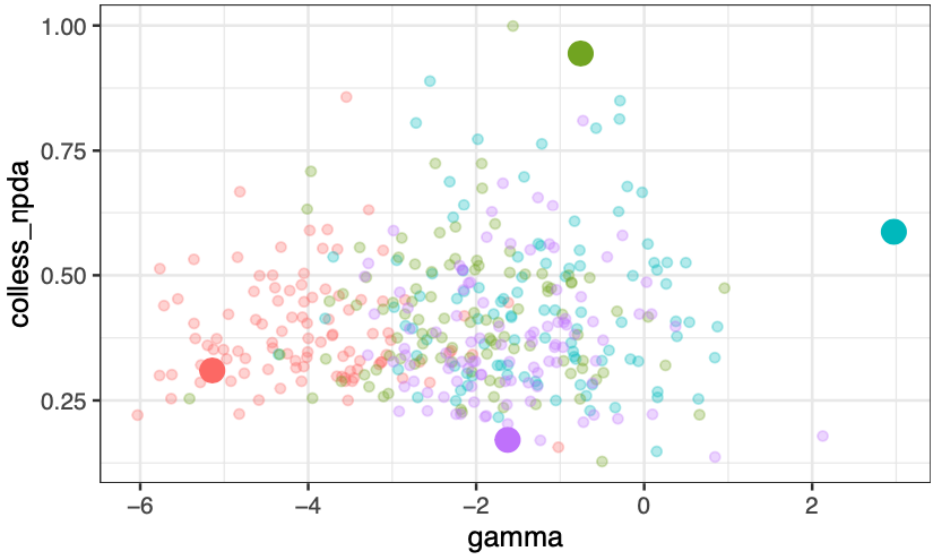


Figure 4.3 | Comparison of LDD simulations with empirical phylogenies. Each small point represents the Colless vs gamma coordinate of a simulated tree under the LDD model. Each colour represents a parameter combination for which we used the maximum likelihood estimators for the four trees shown in Figure 4.2. The Colless and Gamma statistics for these four trees are shown as large circles.

4.3. THE PHYLOGENETIC-DIVERSITY MATRIX IN LID MODELS

The development of LDD models has only just started; the vast majority of developed and used SDM are LID models (Morlon, 2014), especially because if their computational simplicity (Slowinski and Guyer, 1989). The few LDD models (e.g. (Oliveira *et al.*, 2020)) do not take into account ecological interactions and other essential properties of the diversification process. Here we aim to generalise diversity-dependent diversification models in order to keep them flexible enough to capture mode and tempo, even in extreme phylogenetic trees. We thus search for a model which: (1) incorporate time-varying carrying capacities (Marshall and Quental, 2016), (2) has heritable rates (Caron and Pie, 2020), (3) has the flexibility to promote speciation for younger species for unbalanced trees and promotes speciation in older clades for balanced trees (Jones, 2011), (4) considers community interactions among lineages, (5) considers the dynamical nature of niche diversity (Smaldino *et al.*, 2019) as the ecological role that an organism plays in an ecosystem changes.

4.3.1. PHYLOGENETIC DIVERSITY

Phylogenetic diversity or phylodiversity is an ideal candidate for capturing interactions between species, because it is associated with functional diversity (Oliveira *et al.*, 2020), character diversity and other ecological features, although there is still some controversy about this association (Mazel *et al.*, 2018; Tucker *et al.*, 2016). Here we use a per-species phylogenetic diversity index, so we can model LID models where the speciation rate of each species is proportional to the phylogenetic distance of this species holds to the rest of the species in the clade. For this purpose we define a phylogenetic diversity matrix, known also as phylogenetic distance matrix although this term is not only restricted in the literature to the process here defined.

Let $\mathscr{S}_t$ be the set of all species in the phylogenetic tree at time t. We define the phylogenetic diversity matrix $P(t)$ as a dynamic matrix, with dynamic dimension $|\mathscr{S}_t| \times |\mathscr{S}_t|$, that takes into account the phylogenetic distance between species. The entries of the matrix are defined as the times to the most recent common ancestor for each pair of species,

$$P_{ij}(t) = \text{time to most recent common ancestor of species } i \text{ and } j.$$

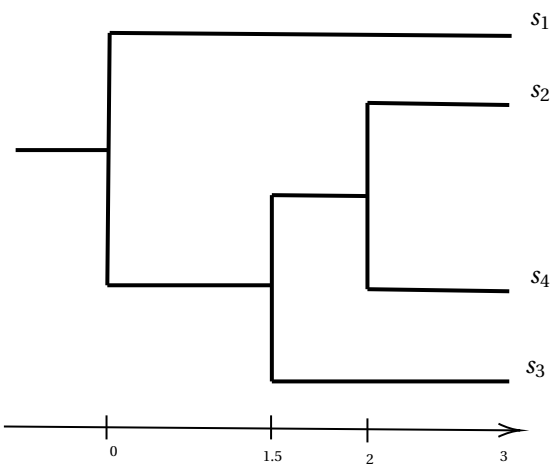
Figure 4.4 shows a simple tree as an example with calculations of the phylogenetic diversity matrix at three different times.

We then define for each species s the mean phylogenetic diversity (Mazel *et al.*, 2016) as

$$P_{s,t} = \frac{1}{|\mathscr{S}_t|} \sum_{s' \in \mathscr{S}_t} P_{s,s'}(t);$$

which is proven to be closely related to Faith's phylogenetic diversity (Faith, 1992) which is widely used in both macroevolution and ecology. We define the overall mean phylogenetic diversity as

$$PDM_t = \frac{1}{|\mathscr{S}_t|} \sum_{s \in \mathscr{S}_t} P_{s,t}$$



$P(1) =$

	s_1	s_2
s_1	0	1.1
s_2	1.1	0

$P(1.6) =$

	s_1	s_2	s_3
s_1	0	1.6	1.6
s_2	1.6	0	0.1
s_3	1.6	0.1	0

$P(2.5) =$

	s_1	s_2	s_3	s_4
s_1	0	3	3	3
s_2	3	0	1.5	1
s_3	3	1.5	0	1.5
s_4	3	1	1.5	0

Figure 4.4 | Phylogenetic diversity matrix at three different time points

which represents the average distance that each species has with the rest of the species in the phylogeny. It describes how phenotypically distinct it is from the other extant species.

These quantities have both a robust biological meaning and elegant and convenient mathematical properties. Note that, between branching times (i.e in periods when no speciation or extinction happens) we have that

$$P_{s,t_i+t} = P_{s,t_i} + t; \qquad PDM_{t_i+t} = PDM_t + t \tag{4.1}$$

Our definition of PDM_t entails

$$\sum_s \left(PDM_t - P_{t,s}\right) = 0 \tag{4.2}$$

We use these properties to develop fast and efficient inference algorithms.

4.3.2. THE LID MODELS

We here propose a generalisation of the diversity dependence model (Etienne *et al.*, 2012b) which considers differences among species introduced in the speciation rate,

$$\lambda_{t,s} = \lambda_0 + \beta_N(2 - N_t) + \beta_P\left(PDM_t - P_{t,s}\right); \qquad \lambda_0 > 0, \beta_N > 0, \beta_P > 0 \tag{4.3}$$

This model considers a speciation rate that linearly decreases with the number of species as in the usual LDD model (Etienne *et al.*, 2012b) and adds a LID effect which gives a speciation advantage to species with on average shorter distance to other species. This model, called the lineage-dependent phylodiversity-dependent diversification (LDPD) model thus assumes that species that speciate faster will produce species that speciate fast

as well (Caron and Pie, 2020), which will lead to more unbalanced trees. If we change the constraint $\beta_P > 0$ to $\beta_P < 0$, the model assumes that species that are more phylogenetically distant to other species are more likely to speciate (Nyman, 2010), resulting in more balanced trees.

Note that the properties (4.1, 4.2) imply that the overall speciation rate does not accelerate or decrease relative to the LDD overall speciation rate but only creates differences between the rates of different species.

To analyse how the LDPD model can capture balance and tempo we performed a simulation study. We fixed parameters $\lambda_0 = 0.5, \beta_N = -0.05, \mu_0 = 0.1$ and we varied $\beta_P = 0, 0.001, 0.05, 0.1, 0.5$. We performed 100 simulations for each parameter value. Figure 4.5 shows the distribution of mode and tempo of the simulated data. We can see that by changing β_P we can cover the different balance values found in empirical trees, while the distribution of gamma remains wide. This simulation shows the flexibility of the model presented here, especially in relationship with topology. Note that the scale in both indices is larger than for the mammal phylogenies, which suggest that with this model, we can cover balance and tempo observed in nature.

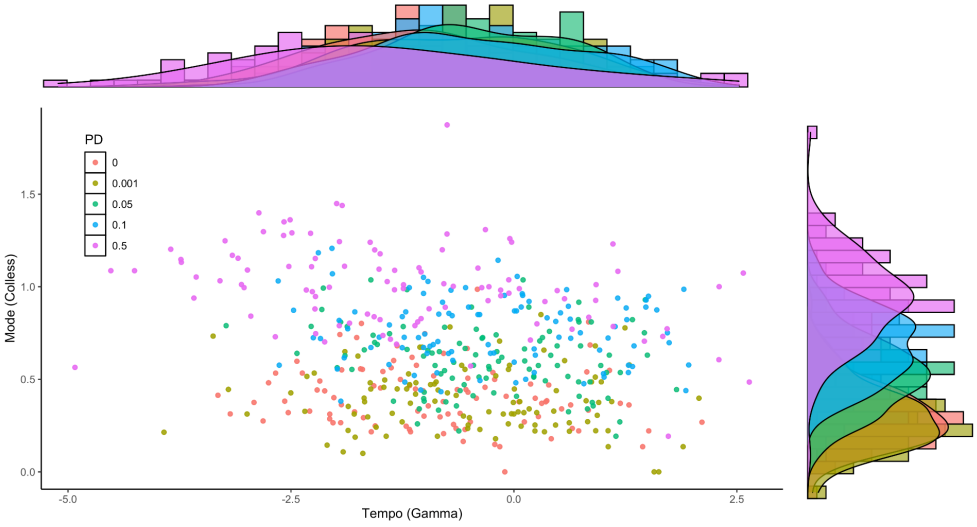


Figure 4.5 | Distribution of tempo and mode, characterised by the Gamma and Colless index, for 5 different values for the PD effect. We used $\lambda_0 = 0.5, \beta_N = -0.05, \mu_0 = 0.1$ and we constrain simulation to a crown time of 15My.

4.4. PARAMETER ESTIMATION

In previous chapters we developed an MCEM algorithm for likelihood optimisation. In this chapter, we propose a different approach where the EM optimisation is replaced by a stochastic gradient descent method (Robbins and Monro, 1951; Chen *et al.*, 2014).

The aim is to maximise the likelihood function

$$f(y|\theta) = \int_{x\in\mathscr{X}(y)} f(x,y|\theta)\,dx = \int_{x\in\mathscr{X}(y)} \frac{f(x,y|\theta)}{g_\theta(x)} g_\theta(x)dx = \mathbb{E}_{x\sim g_\theta}\left[\frac{f(x,y|\theta)}{g_\theta(x)}\right]$$

where y is the observed phylogenetic tree and x is a variable describing all full trees that are in agreement with y. The distribution or importance sampler g_θ can be, for instance, the uniform sampler introduced in Chapter 2 or the efficient emphasis algorithm developed in Section 3.3.3.

Thus, the maximum likelihood estimator is

$$\hat{\theta} = \arg\max_\theta \mathbb{E}_{x\sim g_\theta}\left[\frac{f(x,y|\theta)}{g_\theta(x)}\right]$$

To maximise this function, we propose a Stochastic Gradient Descent (SGD), which iteratively computes

$$\theta_i = \theta_{i-1} - \eta\, G(\theta)$$

where η is a step size (also known as the learning rate in the machine learning literature) and $G(\theta)$ is the gradient of the likelihood function

$$\begin{aligned} G(\theta) &= \nabla\, \mathbb{E}_{x\sim g_\theta}\left[\frac{f(x,y|\theta)}{g_\theta(x)}\right] \\ &= E_{x\sim g_\theta}\frac{f(x,y|\theta)}{g_\theta(x)}\left[\frac{\partial \log f(x,y|\theta)}{\partial\theta} - \frac{\partial \log g_\theta(x)}{\partial\theta}\right] \end{aligned}$$

This gradient can typically not be calculated analytically, but we can use an unbiased Monte-Carlo estimator,

$$\widehat{G(\theta)} \approx \frac{1}{n}\sum_{x_i\sim g_\theta} \frac{f(x_i,y|\theta)}{g_\theta(x_i)}\left[\frac{\partial \log f(x_i,y|\theta)}{\partial\theta} - \frac{\partial \log g_\theta(x_i)}{\partial\theta}\right],$$

where n is the number of sampled trees from the DAA developed in section 3.3.3. Thus, we compute the next-step iteration of the SGD as

$$\theta_i = \theta_{i-1} - \eta\,\widehat{G(\theta)}.$$

This can be evaluated by observing that in the case of species diversification processes, the loglikelihood function is

$$\log f(x|\theta) = \sum_{i\in\mathscr{H}_{spe}} \log\left(\lambda_{t_i,s_i^*|\theta}\right) + \sum_{i\in\mathscr{H}_{ext}} \log\left(\mu_{t_i,s_i^*|\theta}\right) - \int_{t_0}^{t_p}\sum_{s\in\mathscr{S}_t}(\lambda_{t,s|\theta}+\mu_{t,s|\theta})dt$$

where $\mathscr{H}_{spe}$ is the set of indices where the i-th event is speciation and $\mathscr{H}_{ext}$ is the set of indices where the i-th event is an extinction, t_i is the i-th event time and s_i^* is the species that performed an action (speciated or became extinct) at time t_i.

The sampling probability, under the emphasis data augmentation algorithm, is

$$\log g(x|\theta) = \sum_{i\in\mathscr{M}} \left[\log\left(\sum_{s\in\mathscr{S}_{t_i}} \lambda_{t_i,s|\theta} \right) - \log\left(N^e_{t_i^-} + 2N^o_{t_i^-} \right) + \log(\mu_0) \; - \mu_0(t^e_i - t_i) \right] - \int_{t_0}^{t_p} \left[\sum_{s\in\mathscr{S}_r} \lambda_{r,s|\theta} \left(1 - e^{-\mu_0(t_p - r)}\right) \right] dr$$

where $N^e_{t_i^-}$ is the number of missing species just before time t and $N^o_{t_i^-}$ is the number of extant species just before t, $\mathscr{M}$ is the set of indexes corresponding to missing speciations, and t^e_i is the extinction time of the species that speciated at time t_i.

We are interested in the logarithm of the ratio

$$\log r(x|\theta) = \log f(x|\theta) - \log g(x|\theta).$$

In the case of constant extinction rate $\mu_{t_i,s^*_i|\theta} = \mu_0$ there are several simplifications and the log of the ratio is

$$\log r(x|\theta) = \sum_{\mathscr{H}_{spe}} \log\left(\lambda_{t_i,s^*_i|\theta}\right) - \sum_{i=1}^{p} N^o_{t_i}\mu_0(t_i - t_{i-1}) + \sum_{i\in\mathscr{M}}^{p} \left[\log\left(N^e_{t_i^-} + 2N^o_{t_i^-}\right) - \log\left(\sum_{s\in\mathscr{S}_{t_i}} \lambda_{t_i,s|\theta} \right) \right] + \int_{t_0}^{t_p} \sum_{s\in\mathscr{S}_t} \lambda_{t,s|\theta} e^{-\mu_0(t_p - t)} dt$$

Thus, for the LDPD model we have

$$\begin{aligned} \log r(x|\theta) = & \sum_{i\in\mathscr{H}_{spe}} \log\left(\lambda_0 + \beta_N(2 - N_{t_i}) + \beta_P P'_{t_i,s^*}\right) + \\ & \sum_{i\in\mathscr{M}}^{p} \left[\log\left(N^e_{t_i} + 2N^o_{t_i}\right) - \log\left(N_{t_i}(\lambda_0 + \beta_N(2 - N_{t_i}))\right)\right] + \\ & \sum_{i=1}^{p} N^o_{t_i}\mu_0(t_i - t_{i-1}) + N_{t_i}(\lambda_0 + \beta_N(2 - N_{t_i}))\frac{e^{-\mu_0 t_p}}{\mu_0}[e^{\mu_0 t_i} - e^{\mu_0 t_{i-1}}] \end{aligned} \tag{4.4}$$

where $P'_{t,s} = \left(PDM_t - P_{t,s}\right)$. Thus, the gradients with respect to the various parameters is calculated with the partial derivatives

$$\begin{aligned} \frac{\partial \log r(x|\theta)}{\partial_{\mu_0}} = & \sum_{i=1,\ldots,p} N^o_{t_i}(t_i - t_{i-1}) + \\ & \frac{1}{\mu_0^2} N_{t_i}\left[\lambda_0 + (2 - N_{t_i})\beta_N\right]\left[e^{\mu_0(t_i - t_p)}[\mu_0(t_i - t_p) - 1] - e^{\mu_0(t_{i-1} - t_p)}[\mu_0(t_{i-1} - t_p) - 1]\right], \end{aligned} \tag{4.5}$$

$$\frac{\partial \log r(x|\theta)}{\partial_{\lambda_0}} = \sum_{i \in \mathscr{H}_{spe}} \left[\frac{1}{\lambda_0 + \beta_N(2 - N_{t_i}) + \beta_P P'_{t_i,s^*}} \right] + \sum_{i \in \mathscr{M}_x} \left[\frac{-N_{t_i}}{N_{t_i}(\lambda_0 + \beta_N(2 - N_{t_i}))} \right] + \sum_{i \in \{1,\dots,p\}} \left[N_{t_i} \frac{e^{-\mu_0 t_p}}{\mu_0} \left[e^{\mu_0 t_i} - e^{\mu_0 t_{i-1}} \right] \right], \tag{4.6}$$

$$\frac{\partial \log r(x|\theta)}{\partial_{\beta_N}} = \sum_{i \in \mathscr{H}_{spe}} \left[\frac{(2 - N_{t_i})}{\lambda_0 + \beta_N(2 - N_{t_i}) + \beta_P P'_{t_i,s^*}} \right] + \sum_{i \in \mathscr{M}_x} \left[\frac{-N_{t_i}}{N_{t_i}(\lambda_0 + \beta_N(2 - N_{t_i}))} \right] + \sum_{i \in \{1,\dots,p\}} \left[N_{t_i} \frac{e^{-\mu_0 t_p}}{\mu_0} \left[e^{\mu_0 t_i} - e^{\mu_0 t_{i-1}} \right] \right], \tag{4.7}$$

and

$$\frac{\partial \log r(x|\theta)}{\partial_{\beta_P}} = \sum_{i \in \mathscr{H}_{spe}} \left[\frac{P'_{t_i,s^*}}{\lambda_0 + \beta_N(2 - N_{t_i}) + \beta_P P'_{t_i,s^*}} \right]$$

Thus, we have an explicit form to compute the stochastic gradient descent step and perform optimisation. The method can be used for any kind of model, but gradients need to be calculated in every case.

4.5. SUMMARY

Species diversification models can be used to quantify the relationship of different ecological variables with species diversification processes. Most current implemented species diversification models assume that all species are equally probable to speciate or become extinct, which does not allow to quantify the effect of the topology on the processes.

We have presented a generalised diversity-dependence model that preserves the relationship between speciation rate and species richness of previously studied models but adds an ecological advantage to species that are either more or less phylogenetically distant to the other species in the clade. With simulations we have shown that this model is flexible enough to capture a large variety of topologies. We propose this model as a standard alternative to current diversity-dependent diversification models. This model can be also complemented with the model of the previous chapter, which also takes into account dynamical carrying capacities.

Finally, we have presented an estimation method based on a stochastic gradient descent method and provide the corresponding equations to use it for the LDPD model.