

ASYMPTOTICS OF ORNSTEIN-UHLENBECK TREE MODELS: THEORY AND COMPUTATION

by

Lam Si Tung Ho

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2014

Date of final oral examination: 07/11/2014

The dissertation is approved by the following members of the Final Oral Committee:

Cécile Ané, Associate Professor, Statistics and Botany

Bret Larget, Professor, Statistics and Botany

Michael Newton, Professor, Statistics and Biostatistics & Medical Informatics

Sebastien Roch, Associate Professor, Mathematics and Statistics

Jun Zhu, Professor, Statistics and Entomology

© Copyright by Lam Si Tung Ho 2014
All Rights Reserved

ACKNOWLEDGMENTS

First and foremost, I would like to deeply thank my advisor, professor Cécile Ané, because without her guidance, this thesis would not have been possible. I appreciate all our fruitful discussions from which I learned a lot. Her passion for research is contagious and I am infected by that. I also thank Dr. Ané for her invaluable advice on research and career development.

I am thankful to professor Bret Larget, professor Michael Newton, professor Sebastien Roch, and professor Jun Zhu to be my committee members. I would like to thank them for their time, questions, and comments. Especially, I wish to thank professor Sebastien Roch for giving me insightful suggestions leading to results in Chapter 3.

I would like to express my gratitude to professor Rick Nordheim. He brought statistics into my life. Without him, I would not have pursued a PhD in statistics. I want to thank him for his constant support, help, and encouragement. I also would like to thank professor Duong Minh Duc for encouraging me to study abroad.

Last but not least, I would like to thank my parents, my wife and my younger sister for their endless support and unconditional love. I also wish to thank all my friends for their friendship and support.

CONTENTS

Contents ii

Abstract iv

1 Introduction 1

2 Microergodicity 7

2.1 *Introduction and overview of main results* 7

2.2 *Microergodicity under hierarchical autocorrelation* 13

2.3 *Different convergence rates of ML estimators for different microergodic parameters* 22

2.4 *Discussion* 26

3 Consistency and $\sqrt{n}$ -consistency on trees with unbounded height 28

3.1 *Introduction* 28

3.2 *Definitions and statements of results* 31

3.3 *Estimating μ* 43

3.4 *Convergence rate of a new estimator for α and γ* 63

4 Practical inference difficulties and challenges with model selection 75

4.1 *Introduction* 75

4.2 *Limitations due to phylogenetic dependence* 77

4.3 *(Un)identifiability of selection regimes* 80

4.4 *(Non)-Microergodicity of selection parameters* 85

4.5 *Power to detect shifts in the selection optimum* 89

4.6 *Conclusion* 98

5 A linear-time algorithm for trait evolution 100

5.1 *Introduction* 100

5.2 *Algorithm* 103

5.3	<i>The (generalized) 3-point structure</i>	110
5.4	<i>Gaussian trait models with the 3-point structure</i>	113
5.5	<i>Applications to non-Gaussian models</i>	117
5.6	<i>Further Applications</i>	122
5.7	<i>Discussion</i>	123
6	Discussion and Future Developments	126
A	Appendix A	127
A.1	<i>Spectral decomposition of the OU covariance matrix on symmetric trees</i>	127
A.2	<i>Supporting lemmas and technical proofs</i>	128
B	Appendix B	138
B.1	<i>Identifiability of multiple selection optima</i>	138
B.2	<i>Supplementary Figures</i>	139
C	Appendix C	145
C.1	<i>Algorithm for multiple traits with different phylogenetic signals</i>	145
C.2	<i>Algorithm for multiple traits with different within and between-species covariances</i>	146
	References	147

ABSTRACT

Tree models arise in evolutionary biology when sampled species are related to each other according to a phylogenetic tree. Tips of this tree represent contemporary species while internal nodes represent their ancestors. Researchers are often interested in studying the evolution of a phenotype using only observations from the present (at the tips). In this context, the Ornstein-Uhlenbeck (OU) process has been widely used to model trait evolution under the presence of natural selection, whereby a trait is attracted to a selection optimum μ with selection strength α . The primary goal of this thesis is to study the asymptotics of OU tree models theoretically and computationally.

In the theoretical part, I report several intrinsic inference difficulties of OU tree models including non-identifiability and non-microergodicity. A well-defined and practical model parametrization is proposed to avoid non-identifiability. Situations where non-microergodicity occurs are also identified. Moreover, I provide a necessary and sufficient condition for the consistency of the maximum likelihood estimator of μ and establish a phase transition on its convergence rate. I also propose a novel estimation method for α , which achieves an optimal convergence rate under natural assumptions.

In the computational part, I develop a linear-time algorithm which is applicable to many Gaussian models including OU processes with varying selection parameters along the tree, as well as non-Gaussian models like phylogenetic logistic regression and phylogenetic generalized linear (mixed) models. The traditional calculation methods are computationally heavy, requiring complex and time-consuming matrix calculations, and causing failure or inaccurate results on very large trees. The new algorithm for phylogenetic regression is implemented in the R package `phylolm`.

1 INTRODUCTION

Since Darwin, it is widely accepted that biological species share an evolutionary history described by a phylogenetic tree. In this tree, tips represent present-day species, internal nodes represent past ancestors, and the length of an edge represents the amount of time either between two speciation events (internal edge) or from the last speciation event to the present (external edge). Researchers often want to make inferences about the history of evolution based on data from the present. Because related species are more similar than distant species, data across species violate the usual independent assumption. Thus, several evolutionary models have been developed to take into account this relatedness. These models, which we call tree models, assume that traits evolved along the phylogeny according to a Markov process. Tree models are special cases of graphical models where a tree is a directed acyclic graph such that every branch is a directed edge starting from an ancestor to one of its descendant, and such that data are observed at leaves only, not at internal nodes.

For continuous traits, the two most commonly used tree models are the Brownian Motion (BM) model and the Ornstein-Uhlenbeck (OU) model. The BM model is useful for neutral evolution where the trait is influenced only by random drift ([Felsenstein, 1985](#)) while the OU model is more appropriate to model evolution under the presence of natural selection. The OU model assumes that the trait evolves along the tree according to an OU process, which is attracted toward a “selection optimum” μ while still being influenced by random drift ([Hansen, 1997](#)). Such process is defined by the following stochastic equation ([Ikeda and Watanabe, 1981](#))

$$dY_t = -\alpha(Y_t - \mu)dt + \sigma dB_t,$$

where Y_t is the response variable, μ is the selection optimum, $\alpha \geq 0$ is the selection strength, $\sigma \geq 0$ is the scale parameter, and B_t is a standard BM. At an internal node (a speciation event), the process splits into several independent

processes starting from the value of this internal node. Each process represents the trait evolution along one descendant branch of the node. In the case when $\alpha = 0$, the action of natural selection disappears and the trait evolves as a BM process. Therefore, the BM model is a special case of the OU model, and a test to select between them can be used to detect the presence of natural selection. My thesis is dedicated to study the asymptotics of the OU tree models both theoretically and computationally.

Chapter 2, which was published in the Annals of Statistics (Ho and Ané, 2013a), considers a scenario where μ and α are constant. In this scenario, the trait value at the root y_0 can either be random or fixed. If y_0 is random, it is natural to assume that it follows the stationary distribution, which is a Gaussian distribution with mean μ and variance $\sigma^2/(2\alpha)$. In this case, the distribution of the observations at the tips $(Y_i)_{i \in \text{tips}}$ is Gaussian with mean μ and variance

$$\frac{\sigma^2}{2\alpha} \mathbf{V}, \text{ with } V_{ij} = e^{-\alpha d_{ij}}$$

where d_{ij} is the tree distance between tips i and j , i.e. the length of the path between i and j . If instead the model is conditional on the ancestral value y_0 at the root, the Gaussian process has mean $(1 - e^{-\alpha t_{ii}})\mu + e^{-\alpha t_{ii}}y_0$ for tip i and variance matrix

$$\frac{\sigma^2}{2\alpha} \mathbf{V}, \text{ with } V_{ij} = e^{-\alpha d_{ij}}(1 - e^{-2\alpha t_{ij}})$$

where t_{ii} is the distance from the root to tip i and t_{ij} is the shared time from the root to tips i and j . To study trait evolution, biologists are often interested in estimating the selective optimum μ , the selection strength α , and the variance rate σ^2 using the maximum likelihood estimator (MLE) (see e.g. Butler and King, 2004; Cooper and Purvis, 2010; Beaulieu et al., 2012). However, very few studies have focused on the statistical properties of this method on tree models, and the usual asymptotic properties have been taken for granted in empirical studies for more than twenty five years. In Chapter 2, we investigate the microergodicity of parameters under the OU random root model. A function

$f(\mu, \alpha, \sigma^2)$ is microergodic if the distribution $P_{(\mu, \alpha, \sigma^2)}$ of the whole observable process $(Y_i)_{i \geq 1}$ at the tips of the limit infinite tree is such that $P_{(\mu_1, \alpha_1, \sigma_1^2)}$ and $P_{(\mu_2, \alpha_2, \sigma_2^2)}$ are orthogonal for any values $f(\mu_1, \alpha_1, \sigma_1^2) \neq f(\mu_2, \alpha_2, \sigma_2^2)$. The concept of microergodicity was introduced in spatial statistics (Stein, 1999), and it is a necessary condition for the existence of any consistent estimator (Zhang, 2004). We prove that if the tree height is bounded, then the stationary mean (or selection optimum) μ is not microergodic. As a consequence, there is no hope to construct any consistent estimator for μ . Furthermore, we prove that the selective strength α and the variance rate σ^2 are not microergodic in some scenarios. However, we give a general and mild sufficient condition ensuring the microergodicity of both α and σ^2 : if the internal node ages do not cluster tightly. For example, if the internal node ages have at least two limit points, then both α and σ^2 are microergodic. If not, σ^2 is still microergodic as long as zero is a limit point of the set of internal node ages. This is the case when the tree has many “young” internal nodes.

The non-microergodicity of μ presented in Chapter 2 is restricted to trees with bounded height. This assumption can be satisfied in practice when biologists are only interested in studying a certain family of species. On the other hand, tree growth processes where the tree height is unbounded are also of interest mathematically (for instance, the Yule (pure birth) process Yule, 1925) and can be applied to biological studies across several families. Chapter 3, which is submitted to the Journal of Mathematical Biology (Ané et al., 2014), considers the scenario when the tree height is unbounded. We provide a necessary and sufficient condition for the MLE of μ to be consistent. To be specific, conditional on the tree and α , the MLE of μ is consistent if and only if for any $t > 0$, the number of species living at time t away from the present increases to infinity with the number of present-day species. A phase transition on the convergence rate of the MLE is also established: its $\sqrt{n}$ -consistency is lost if the tree is too bushy relative to the selection strength α . The bushiness of the tree is measured by standard notions used in probability theory on trees, such as the growth and the branching number (see e.g. Peres, 1999). The growth only represents how fast

new species arise while the branching number also takes into account the shape of the tree. For the covariance parameters, we propose a novel estimation method which is constructed from two sets of independent contrasts. By using contrasts, we can get rid of the nuisance parameter μ . This is very similar to the idea of restricted maximum likelihood. We prove that this estimation method achieves $\sqrt{n}$ -consistency under natural assumptions on the tree. Our results can be applied to trees generated from the Yule process. In particular, the MLE of μ is consistent conditional on the tree and α , regardless of the speciation rate λ . Moreover, if $\lambda > 2\alpha$, the MLE of μ is $\sqrt{n}$ -consistent. On the other hand, the $\sqrt{n}$ -consistency is lost when $\lambda < 2\alpha$. Finally, the assumptions for the $\sqrt{n}$ -consistency of our estimators for α and γ are satisfied asymptotically under the Yule process.

In Chapter 4, which is submitted to *Methods in Ecology and Evolution* (Ho and Ané, 2014), we consider more general OU models where the selection optimum μ has different values $(\mu_1, \mu_2, \dots, \mu_p)$ on different parts of the tree (see Butler and King, 2004, for more details). We show that if the value at the root y_0 is fixed, non-identifiability occurs very often for these OU models and causes the MLE of (y_0, μ) to be non-unique. Two practical re-parametrization methods are suggested to solve this problem. Unfortunately, identifiability cannot guarantee the consistency of the MLE due to the correlation between observations (see Chapters 2 and 3). We illustrate this issue using simulations on several large trees from across the tree of life. We also investigate carryover effects of the non-microergodicity on the problem of detecting changes in the selection optimum μ . When the positions of the changes are known, our simulations show that the power to detect these changes depends little on the sample size. On the other hand, the power depends significantly on the effect size, which is the magnitude of the shifts in μ relative to the standard deviation at the tips. In the scenario where the numbers and positions of the shifts are unknown, we face an issue of selecting from a huge number of possible models. Indeed, the number of possible models having m changes along an n -taxon bifurcating tree is $\frac{(2n-2)!}{m!(2n-2-m)!}$ (choosing m positions for m shifts among $2n - 2$ edges). Our simulations show that the traditional stepwise procedures using AIC or BIC

fail to select the correct model and grossly overestimate the number of changes. We propose a phylogenetic adaptation of the modified BIC which is suggested by [Zhang and Siegmund \(2007\)](#) for the traditional change-point problem. In particular, we use the following penalty for models having m shifts on an n -taxon tree:

$$3 \log n + (2m - 1) \log n + \sum_{i=0}^m \log n_i,$$

where n_0 is the number of observations which are not influenced by any shift and n_i ($i = 1, 2, \dots, m$) is the number of observations influenced by shift i . Notice that $n_0 + n_1 + \dots + n_m = n$. This modified BIC penalizes more severely each additional change in the model. Its penalty is roughly between 2 to 3 times greater than the usual BIC penalty as the number of changes gets larger. Our results show that this new criterion avoids favoring overly complex models and performs much better than both AIC and BIC.

Besides theoretical limitations mentioned in previous Chapters, the OU tree models are challenged by large phylogenetic covariance matrices. Traditional methods for calculating the likelihood of the OU model often use generic methods such as Cholesky decomposition to compute the inverse and the determinant of the covariance matrix. Such computations do not raise any issue for small trees with dozens of taxa. However, [Hadfield and Nakagawa \(2010\)](#) pointed out that these brute-force methods can be time-consuming and cause failure on large trees with thousands of species. Chapter 5, which was published in *Systematic Biology* ([Ho and Ané, 2013b](#)), develops a linear time algorithm for efficient likelihood calculations on very large trees. Our algorithm can be applied to a broad class of models whose covariance matrix $\mathbf{V}$ satisfies a 3-point condition: symmetric, non-negative entries ($V_{ij} \geq 0$), and for any i, j, k (not necessarily distinct), the two smallest of V_{ij} , V_{ik} and V_{jk} are equal. Most of the Gaussian trait evolution models including the BM model and the OU models belong to this class. Furthermore, non-Gaussian methods such as phylogenetic logistic regression and phylogenetic Poisson regression can also be sped up by our algorithm. To illustrate the gain in computing time of our method,

we performed several simulations on the 4507-species mammal tree (Bininda-Emonds et al., 2007). This algorithm has been implemented in the R package `phylolm` <http://cran.r-project.org/web/packages/phylolm/>, with functions for phylogenetic linear regression and phylogenetic logistic regression.

2 MICROERGODICITY

Abstract Hierarchical autocorrelation in the error term of linear models arises when sampling units are related to each other according to a tree. The residual covariance is parametrized using the tree-distance between sampling units. When observations are modeled using an Ornstein-Uhlenbeck (OU) process along the tree, the autocorrelation between two tips decreases exponentially with their tree distance. These models are most often applied in evolutionary biology, when tips represent biological species and the OU process parameters represent the strength and direction of natural selection. For these models, we show that the mean is not microergodic: no estimator can ever be consistent for this parameter, and provide a lower bound for the variance of its MLE. For covariance parameters, we give a general sufficient condition ensuring microergodicity. This condition suggests that some parameters may not be estimated at the same rate as others. We show that, indeed, maximum likelihood estimators of the autocorrelation parameter converge at a slower rate than that of generally microergodic parameters. We showed this theoretically in a symmetric tree asymptotic framework and through simulations on a large real tree comprising 4507 mammal species.

Funding This work was funded in part by the National Science Foundation (DEB-0830036 and DMS-1106483).

2.1 Introduction and overview of main results

2.1.1 Motivation

This work is motivated by the availability of very large data sets to compare biological species, and by the current lack of asymptotic theory for the models that are used to draw inference from species comparisons. For instance, [Cooper and Purvis \(2010\)](#) studied the evolution of body size in mammals using data from 3,473 species whose genealogical relationships are depicted by their family tree in

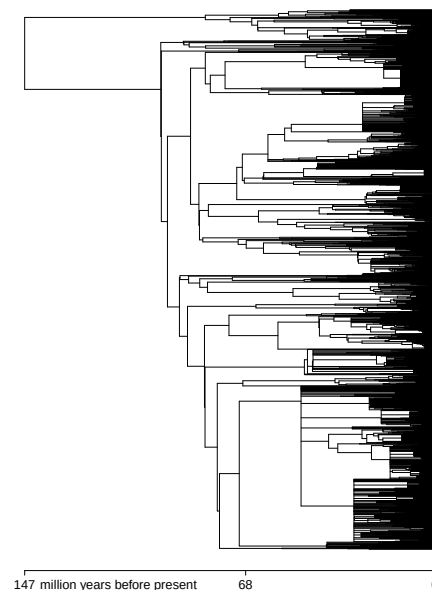


Figure 2.1: Family tree of 4,507 mammal species (Bininda-Emonds et al., 2007). Branch lengths indicate estimated diversification times on the horizontal axis. The Cretaceous/Tertiary mass extinction event marked the extinction of dinosaurs 65.5 million years ago. Cooper and Purvis (2010) used body mass data available for 77% of these species to infer the mode of evolution: neutral evolution (BM) versus natural selection (OU).

figure 2.1. Even from this abundance of data, Cooper and Purvis found a lack of power to discriminate between a model of neutral evolution versus a model with natural selection. To model neutral evolution, body size is assumed to follow a Brownian motion (BM) along the branches of the tree, with observations made on present-day species at the tips of the tree. To model natural selection, body size is assumed to follow an Ornstein-Uhlenbeck (OU) process, whose parameters represent a selective body size (μ) and a selection strength (α). The lack of power observed by Cooper and Purvis suggests a non-standard asymptotic behavior of the model parameters, which is the motivation for our work.

2.1.2 Tree structured autocorrelation

Hierarchical autocorrelation, as depicted in the mammalian tree, arises whenever sampling units are related to each other through a vertical inheritance pattern, like biological species, genes in a gene family, or human cultures. In the genealogical tree describing the relatedness between units, internal nodes represent ancestral unobserved units (like species or human languages). Branch lengths measure evolutionary time between branching events and define a distance between pairs of sampling units. This tree and its branch lengths can be used to parametrize the expected autocorrelation. For doing so, the BM and the OU process are the two most commonly used models. They are defined as usual along each edge in the tree. At each internal node, descendant lineages inherit the value from the parent edge just prior to the branching event, thus ensuring continuity of the process. Conditional of their starting value, each lineage then evolves independently of the sister lineages. BM evolution of the response variable (or of error term) along the

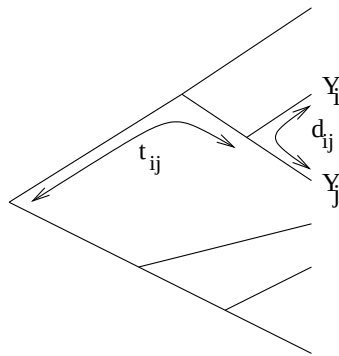


Figure 2.2: Correlation (or residual correlation) between observations at tips i and j are parametrized in the OU model as a function of the tree distance d_{ij} between i and j and of the length t_{ij} of their shared path from the root. For instance, [Cooper and Purvis \(2010\)](#) considered body mass (Y) across 3,473 mammal species ($i, j = 1 \dots 3473$).

tree results in normally distributed errors and in a covariance matrix governed by the tree, its branch lengths, and a single parameter σ^2 . The covariance between two tips i and j is simply $\sigma^2 t_{ij}$, where t_{ij} is the shared time from the root of

the tree to the tips (figure 2.2). Under the more complex OU process, changes towards a value μ are favored over changes away from this value, making the OU model appropriate to address biological questions about the presence or strength of natural selection. This model is defined by the following stochastic equation (Ikeda and Watanabe, 1981) $dY_t = -\alpha(Y_t - \mu)dt + \sigma dB_t$ where Y is the response variable (such as body size), α is the selection strength and B_t is a BM process. In what follows, μ is called the “mean” even though it is not necessarily the expectation of the observations. It is the mean of the stationary distribution of the OU process, and it is the mean at the tips of the tree if the state at the root has mean μ . In the biology literature, μ is called the “optimal” value or “adaptive optimum” in reference to the action of natural selection, but this terminology could cause confusion here with likelihood optimization. The parameter α measures the strength of the pull back to μ . High α values result in a process narrowly distributed around μ , as expected under strong natural selection if the selective fitness of the trait is maximized at μ and drops sharply away from μ . Simple mathematical models of natural selection at the level of individuals result in the OU process for the population mean (Lande, 1979; Hansen and Martins, 1996). If $\alpha = 0$, the OU process reduces to a BM with no pull towards any μ value, as if the trait under consideration does not affect fitness. While some applications focus on the presence of natural selection ($\alpha \neq 0$) such as Cooper and Purvis (2010), other applications are interested in models where μ takes different values ($\mu_1, \dots, \mu_p$) along different branches in the tree, to model different adaptation regimes (e.g. Butler and King, 2004). Other applications assume a randomly varying μ along the tree, varying linearly with explanatory variables (Hansen et al., 2008). In our work, we develop an asymptotic theory for the simple case of a constant μ over the whole tree. The covariance between two observed tips depends on how the unobserved response at the root is treated. It is reasonable to assume that this value y_0 at the root is a random variable with the stationary Gaussian distribution with mean μ and variance $\gamma = \sigma^2/(2\alpha)$. With this assumption, the observed process $(Y_i)_{i \in \text{tips}}$ is Gaussian with mean μ

and variance matrix

$$\gamma \mathbf{V}, \text{ with } V_{ij} = e^{-\alpha d_{ij}} \quad (2.1)$$

where d_{ij} is the tree distance between tips i and j , i.e. the length of the path between i and j . Therefore, the strength α of natural selection provides a direct measure of the level of autocorrelation. If instead we condition on the response value y_0 at the root, the Gaussian process has mean $(1 - e^{-\alpha t_{ii}})\mu + e^{-\alpha t_{ii}}y_0$ for tip i and variance matrix

$$\gamma \mathbf{V}, \text{ with } V_{ij} = e^{-\alpha d_{ij}}(1 - e^{-2\alpha t_{ij}}) \quad (2.2)$$

where, again, t_{ii} is the distance from the root to tip i and t_{ij} is the shared time from the root to tips i and j (Figure 2.2).

2.1.3 Main results and link to spatial infill asymptotics

In contrast to autocorrelation in spatial data or time series, hierarchical autocorrelation has been little considered in the statistics literature, even though tree models have been used in empirical studies for over 25 years. The usual asymptotic properties have mostly been taken for granted. Recently, [Ané \(2008\)](#) showed that the maximum likelihood (ML) estimator of location parameters is not consistent under the BM tree model as the sample size grows indefinitely, proving that the basic consistency property should not be taken for granted. However, [Ané \(2008\)](#) did not consider the more complex OU model, for which the ML estimator admits no analytical formula.

In the spatial infill asymptotic framework when data are collected on a denser and denser set of locations within a fixed domain, σ^2 can be consistently estimated but α cannot under an OU spatial autocorrelation model in dimension $d \leq 3$ ([Zhang, 2004](#)). Recently, α has been proved to be consistently estimated under OU model when $d \geq 5$ ([Anderes, 2010](#)). We uncover here a similar asymptotic behavior under the OU tree model. Just like in infill asymptotics, the tree structure implies that all sampling units may remain within a bounded distance of each

other, and that the minimum correlation between any pair of observations does not go down to zero with indefinitely large sample sizes. It is therefore not surprising that some properties may be shared between these two autocorrelation frameworks. Under infill asymptotics, microergodic parameters can usually be consistently estimated (see [Zhang and Zimmerman, 2005](#)) while non-microergodic parameters cannot (e.g. α). A parameter is microergodic when two different values for it lead to orthogonal distributions for the complete, asymptotic process ([Stein, 1999](#)).

In section [2.2](#), we prove that the mean μ is non-microergodic under the OU autocorrelation framework, and we provide a lower bound for the variance of the MLE of μ . We also give a sufficient condition for the microergodicity of the OU covariance parameters α and σ^2 (or γ) based on the distribution of internal node ages. The microergodic covariance parameter under spatial infill asymptotics with OU autocorrelation, σ^2 , is recovered as microergodic if 0 is a limit point of the sequence of node ages, i.e. with dense sampling near the tips. Our condition for microergodicity suggests that some parameters may not be estimated at the same rate as others. In section [2.3](#), we illustrate this theoretically for a symmetric tree asymptotic framework, where we show that the REML estimator of α converges at a slower rate than that of the generally microergodic parameter. We also illustrate that the ML estimate convergence rate of α is slower than that of σ^2 , through simulations on a large 4507-species real tree showing dense sampling near the tips.

In most of this work, we only consider ultrametric trees, that is, trees in which the root is at equal distance from all the tips. This assumption is very natural for real data. We also focus on model [\(2.1\)](#), because the model matrix is not of full rank under model [\(2.2\)](#) on an ultrametric tree.

2.1.4 Other tree models in spatial statistics

Trees have already been used for various purposes in spatial statistics. When considering different resolution scales, the nesting of small spatial regions into

larger regions can be represented by a tree. The data at a coarse scale for a given region is the average of the observations at a finer scale within this region. For instance, [Huang et al. \(2002\)](#) use this “resolution” tree structure to obtain consistent estimates at different scales, and otherwise use a traditional spatial correlation structure between locations at the finest level. In contrast, the tree structure in our model is the fundamental tool to model the correlation between sampling units, with no constraint between values at different levels. Trees have also been used to capture the correlation among locations along a river network ([Cressie et al., 2006](#); [Hoef et al., 2006](#); [Hoef and Peterson, 2010](#), and discussion). A river network can be represented by a tree with the associated tree distance. To ensure that the covariance matrix is positive definite, moving average processes have been introduced, either averaging over upstream locations or over downstream locations, or both. There are two major differences between our model and these river network models. First, the correlation among moving averages considered in [Cressie et al. \(2006\)](#) and [Hoef and Peterson \(2010\)](#) decreases much faster than the correlation considered in this work. Most importantly, any location along the river is observable, while observations can only be made at the leaves of the tree in our framework.

2.2 Microergodicity under hierarchical autocorrelation

The concept of microergodicity was formalized by [Stein \(1999\)](#) in the context of spatial models. This concept was especially needed in the infill asymptotic framework, when some parameters cannot be consistently estimated even if the whole process is observed. Specifically, consider the complete process $(Y_s)_{s \in S}$ where S is the space of all possible observation units. In spatial infill asymptotics, S can be the unit cube $[0, 1]^d$. In our hierarchical framework, we consider a sequence of nested trees converging to a limit tree, which is the union of all nodes and edges of the nested trees. In this case, S is the set of all tips in the

limit tree. Consider a probability model $(P_\theta)_{\theta \in \Theta}$ on $(Y_s)_{s \in S}$. A function $f(\theta)$ of the parameter vector is said to be microergodic if for all $\theta_1, \theta_2 \in \Theta$, $f(\theta_1) \neq f(\theta_2)$ implies that P_{θ_1} and P_{θ_2} are orthogonal. If a parameter is not microergodic, then there is no hope of constructing any consistent estimator for it (see [Zhang, 2004](#), for an excellent explanation). In spatial infill asymptotics with OU correlation in dimension $d \leq 3$, α and γ are not microergodic even though σ^2 is ([Zhang, 2004](#)) and the MLE of σ^2 is strongly consistent ([Ying, 1991](#)). Also note that the microergodicity of (γ, α) is equivalent to the microergodicity of both γ and α .

2.2.1 Theory of equivalent Gaussian measures

We recall here the theory of equivalent Gaussian measures, which we apply to Ornstein-Uhlenbeck tree models in the next section. We consider two Gaussian measures P_k ($k = 1, 2$) on the σ -algebra $\mathcal{U}$ generated by a sequence of random variables $(Y_j)_{j=1}^\infty$, a linearly independent basis for both $\mathcal{H}_1$ and $\mathcal{H}_2$ where $\mathcal{H}_k$ is the Hilbert space generated by $(Y_j)_{j=1}^\infty$ with linear product: $\langle Y_{j_1}, Y_{j_2} \rangle = \text{cov}_k(Y_{j_1}, Y_{j_2})$ for $k = 1$ or 2 . The entropy distance between equivalent Gaussian measures P_1 and P_2 on the σ -algebra $\mathcal{U}' \subset \mathcal{U}$ is defined as twice the symmetrized Kullback-Leibler divergence:

$$r(\mathcal{U}') = - \left[\mathbb{E}_{P_1} \log \frac{P_2(dw)}{P_1(dw)} + \mathbb{E}_{P_2} \log \frac{P_1(dw)}{P_2(dw)} \right].$$

We will use the following properties proved in [Ibragimov and Rozanov \(1978\)](#):

$$r(\mathcal{U}') \leq r(\mathcal{U}'') \quad \text{for } \mathcal{U}' \subset \mathcal{U}''. \quad (2.3)$$

Consider non singular Gaussian measures P_1 and P_2 on the σ -algebra $\mathcal{U}_n$ generated by $(Y_j)_{j=1}^n$. Let $r_n = r(\mathcal{U}_n)$. Then $(r_n)_{n=1}^\infty$ is non-decreasing and:

$$P_1 \perp P_2 \Leftrightarrow r_n \rightarrow \infty, \quad \text{and} \quad P_1 \equiv P_2 \Leftrightarrow r_n \rightarrow r < \infty. \quad (2.4)$$

We now recall how to calculate r_n as described in [Stein \(1999\)](#) (see also [Ibragimov and Rozanov, 1978](#)). Consider a new basis $(Y_{1,n}, \dots, Y_{n,n})$ obtained by linearly

transforming $(Y_1, \dots, Y_n)$ such that this new basis is centered orthonormal under P_1 : $\mathbb{E}_1 Y_{j,n} = 0$ and $\text{cov}_1(Y_{j_1,n}, Y_{j_2,n}) = \delta_{j_1,j_2}$ is 1 if $j_1 = j_2$ and is 0 otherwise, and such that $\text{cov}_2(Y_{j_1,n}, Y_{j_2,n}) = \sigma_{j_1,n}^2 \delta_{j_1,j_2}$ for some $\sigma_{j_1,n}^2$. Also set $m_{j,n} = \mathbb{E}_2 Y_{j,n}$. Then

$$r_n = \frac{1}{2} \sum_{j=1}^n (\sigma_{j,n}^2 + 1/\sigma_{j,n}^2 - 2 + m_{j,n}^2 + m_{j,n}^2/\sigma_{j,n}^2) .$$

[Rao and Varadarajan \(1963\)](#) take a similar approach using the Hellinger distance instead of the entropy distance r_n . They show that the following condition is sufficient for the orthogonality of P_1 and P_2 .

$$\lim_{n \rightarrow \infty} \sum_{j=1}^n (\sigma_{j,n}^2 - 1)^2 = \infty. \quad (2.5)$$

2.2.2 Microergodicity of Ornstein-Uhlenbeck tree models

We say that $\mathbb{T}$ is a subtree of tree $\mathbb{T}'$ if we can get $\mathbb{T}$ by removing some branches from $\mathbb{T}'$. We consider a nested sequence of trees $(\mathbb{T}_n)_{n=1}^\infty$ such that $\mathbb{T}_{n-1}$ is a subtree of $\mathbb{T}_n$ for every n . This is to ensure that the observations $(Y_j)_{j=1}^n$ at the tips of $\mathbb{T}_n$ provide a well-defined infinite sequence $(Y_n)_{n \geq 1}$. One essential assumption is that trees are ultrametric, that is, the distance from the root to leaf nodes of tree $\mathbb{T}_n$ is assumed to be the same for all tips. This is equivalent to saying that the tree distances between tips define an ultrametric metric. This assumption comes in naturally. If the distance from the root to all tips is constant, models (2.1) and (2.2) predict equal variances and equal means at the tips, which are reasonable assumptions. Ultrametric trees arise in most applications when tips are extant species sampled at the present time and branch lengths represent time calibrated in millions of years for instance. Define $\mathcal{J}^{\mathbb{T}_n}$ as the set of all internal nodes of tree $\mathbb{T}_n$ (including the root) and $\mathcal{J} = \bigcup_{n=1}^\infty \mathcal{J}^{\mathbb{T}_n}$. Let $(T_i)_{i \in \mathcal{J}}$ be the sequence of node ages. The age of a node is the distance from the node to any of its descendant tip. This is well defined on ultrametric trees. $\mathcal{J}^{\mathbb{T}_n}$ is a subset of $\mathcal{J}^{\mathbb{T}_{n+1}}$ so $(T_i)_{i \in \mathcal{J}}$ is a well defined infinite sequence. In most of what

follows, we will assume that

- (C) $(\mathbb{T}_n)_{n=1}^\infty$ is a nested sequence of ultrametric trees and the sequence of internal node ages $(T_i)_{i \in \mathcal{I}}$ is bounded.

Without loss of generality, we can assume that all trees are bifurcating because a multifurcating tree can be made into a bifurcating tree with some zero branch lengths. With this assumption $\mathcal{I}^{\mathbb{T}_n}$ contains $n - 1$ internal nodes. This is equivalent to counting nodes and their ages with multiplicity, where an internal node having d descendants contributes his age $d - 1$ times.

Theorems 2.1 and 2.3 below state general results on the microergodicity of parameters in OU tree models. Our main tool is the equivalence (2.4) applied to $r_n = r(\mathbb{T}_n)$, the entropy distance between P_{θ_1} and P_{θ_2} for two parameter sets $\theta_k = (\mu_k, \alpha_k, \gamma_k)$, $k = 1, 2$, on the σ -algebra generated by $(Y_j)_{j=1}^n$.

2.2.3 Microergodicity of the mean μ

Theorem 2.1. *Under OU model (2.1) and condition (C), μ is not microergodic.*

The theorem follows directly from (2.4) and the boundedness of $(T_i)_{i \in \mathcal{I}}$ once the following upper bound is established:

$$r(\mathbb{T}) \leq (\mu_1 - \mu_2)^2 / (\gamma_1 e^{-2\alpha_1 T}) \quad (2.6)$$

if $\alpha_1 = \alpha_2$ and $\gamma_1 = \gamma_2$, where T is the age of the root of $\mathbb{T}$ (appendix A.2.2). One consequence is that there is no consistent estimator for μ . To illustrate this, we consider the MLE of μ and provide a lower bound for its variance. We let t be the length of the shortest branch stemming from the root and k the number of daughters of the root (figure 2.3).

Theorem 2.2. *Assume OU model (2.1) on an ultrametric tree. Let $\hat{\mu}$ be the MLE of μ conditional on some possibly wrong value α_* of α . Then*

$$\text{var}(\hat{\mu}) \geq \frac{\sigma^2}{2\alpha} e^{-2\alpha t} \left(1 + \frac{e^{2\alpha t} - 1}{k} \right). \quad (2.7)$$

The equality holds if and only if α is known ($\alpha_* = \alpha$) and the tree is a star tree with the root as unique internal node, in which case $k = n$ and $t = T$. If T is bounded as the sample size n grows and $\alpha > 0$, then $\hat{\mu}$ is not consistent.

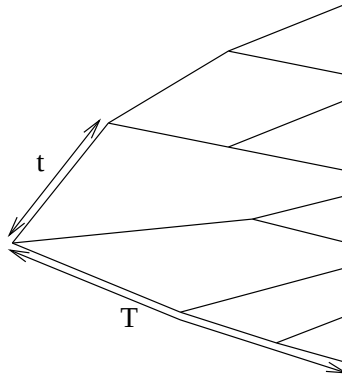


Figure 2.3: Ultrametric tree with all tips at equal distance T from the root. The root has $k = 3$ children here, and t is the minimum distance from the root to its children.

The second part of the theorem follows directly from the lower bound (2.7). Note that $\hat{\mu}$ is Gaussian with mean μ . Therefore, the lower bound of its variance implies that $\hat{\mu}$ cannot converge to μ . Hence, it is not consistent.

The assumption that $\alpha > 0$ is trivial. When $\alpha = 0$, the OU process reduces to a BM where μ has no influence on the process. In that case, μ is no longer a parameter in the model. As expected, the lower bound on the variance of $\hat{\mu}$ is heavily influenced by the actual value of the correlation parameter α . The precision of $\hat{\mu}$ is weakest when autocorrelation is strong, i.e. when α is small, for a given value of $\gamma = \sigma^2/(2\alpha)$.

The ultrametric assumption is necessary. If the tree is not ultrametric, model (2.2) predicts unequal variances and most importantly unequal means at the tips. Such trees can carry more information about μ . Consider for instance the star tree in figure 2.4, in which all tips are directly connected to the root, by a branch of length t_1 for half of the tips and of length t_2 for the other half of the tips. If $t_1 \neq t_2$ the variance of $\hat{\mu}$ goes to 0 as the sample size grows (see appendix

A.2.2), thus providing a counterexample to theorem 2.2 when the ultrametric assumption is violated.

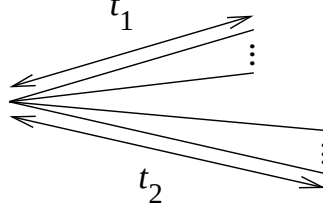


Figure 2.4: Example of a non-ultrametric tree on which μ can be consistently estimated.

Proof of Theorem 2.2. To prove (2.7), we note that $\hat{\mu} = (\mathbf{1}^t V_{\alpha_*}^{-1} \mathbf{1})^{-1} \mathbf{1}^t V_{\alpha_*}^{-1} Y$, where $\mathbf{1}$ is a vector of ones. This estimator is unbiased and has variance $\frac{\sigma^2}{2\alpha} (\mathbf{1}^t V_{\alpha}^{-1} \mathbf{1})^{-1}$ when $\alpha_* = \alpha$ is known. Its variance is larger when α is unknown, by the Gauss-Markov theorem. For this reason, we only need to prove the following lemma (which is done in the appendix A.2.2).

Lemma 2.1. *For all $\alpha > 0$, $(\mathbf{1}^t V_{\alpha}^{-1} \mathbf{1})^{-1} \geq e^{-2\alpha T} + \frac{1}{k}(e^{-2\alpha(T-t)} - e^{-2\alpha T})$ with equality if the tree is a star with k branches stemming from the root.*

Theorem 2.2 can be applied to any tree growth asymptotic framework, so long as T is bounded. For instance, both conditions are met almost surely with $k = 2$ under the coalescent model (Kingman, 1982a,b). Even if these conditions do not hold asymptotically, (2.7) provides a finite-sample upper bound on the estimator's precision. This inequality can be used, for instance, under the Yule model of tree growth (Yule, 1925; Aldous, 2001) if we let both T and n increase indefinitely.

2.2.4 Microergodicity of the autocorrelation parameter (γ, α)

Theorem 2.3. *Under OU model (2.1) and condition (C),*

(a) Let t_0 be a limit point of $(T_i)_{i \in \mathcal{I}}$. Then $f_{t_0}(\gamma, \alpha)$ is microergodic, where

$$f_t(\gamma, \alpha) = \begin{cases} \gamma(1 - e^{-2\alpha t}), & t > 0 \\ \gamma\alpha, & t = 0. \end{cases}$$

(b) If $\sum_{i \in \mathcal{I}} (T_i - t)^2 = \infty$ for all $t \geq 0$ then (γ, α) is microergodic. Note that this condition is satisfied if $(T_i)_{i \in \mathcal{I}}$ has 2 or more limit points.

Proof of Theorem 2.3. The key idea is to reduce the tree for a lower bound of $r(\mathbb{T}_n)$. We will consider subtrees that provide independent contrasts, sufficient to ensure microergodicity. Our constructive proof could be used to construct estimators based on a restricted set of contrasts, but we do not pursue this here. Let $i \in \mathcal{I}$ be an arbitrary internal node, and Y_1^i, Y_2^i be two leaves having i as their most recent common ancestor. Let p_i be the path connecting Y_1^i and Y_2^i . We define $C_i^{p_i} = Y_1^i - Y_2^i$ as a contrast with respect to internal node i and path p_i . For convenience, we define $T_{C_i^{p_i}} = T_i$. The following lemma is proved in appendix A.2.2.

Lemma 2.2. We have that $C_i^{p_i} \sim N(0, 2\gamma(1 - e^{-2\alpha T_i}))$. Also, $C_{i_1}^{p_{i_1}}$ and $C_{i_2}^{p_{i_2}}$ are independent if their paths p_{i_1} and p_{i_2} do not intersect.

Proof of part (a) We denote $\mathcal{I}_S^\mathbb{T} = \{i: T_i \in S, i \in \mathcal{I}^\mathbb{T}\}$ the set of internal nodes of $\mathbb{T}$ whose ages lie in S . Let (γ_1, α_1) and (γ_2, α_2) such that $f_{t_0}(\gamma_1, \alpha_1) \neq f_{t_0}(\gamma_2, \alpha_2)$. Denote

$$g(t) = \frac{1}{2} \left(\frac{f_t(\gamma_1, \alpha_1)}{f_t(\gamma_2, \alpha_2)} + \frac{f_t(\gamma_2, \alpha_2)}{f_t(\gamma_1, \alpha_1)} - 2 \right), t \in [0, T^*]$$

and let $\delta = g(t_0)/2 > 0$. Note that g is continuous at t_0 , so there exists $\epsilon_\delta > 0$ such that $g(t) \geq g(t_0) - \delta$ for all t satisfying $|t - t_0| < \epsilon_\delta$. We now use lemma A.1 (in appendix A.2.1) to select a large set $\mathcal{C}_n$ of independent contrasts with respect to internal nodes whose ages are in $(t_0 - \epsilon_\delta, t_0 + \epsilon_\delta)$ such that $|\mathcal{C}_n| \geq \frac{1}{2} |\mathcal{I}_{(t_0 - \epsilon_\delta, t_0 + \epsilon_\delta)}^\mathbb{T}|$. Let $r(\mathcal{C}_n)$ be the entropy distance between P_{θ_1} and P_{θ_2} on the σ -algebra generated

by $\mathcal{C}_n$. By (2.3) and direct calculation,

$$r(\mathbb{T}_n) \geq r(\mathcal{C}_n) = \sum_{C \in \mathcal{C}_n} g(T_C) \geq |\mathcal{C}_n|(g(t_0) - \delta) = \delta |\mathcal{C}_n| \geq \frac{\delta}{2} |\mathcal{J}_{(t_0 - \epsilon_\delta, t_0 + \epsilon_\delta)}^{\mathbb{T}_n}|.$$

Clearly $|\mathcal{J}_{(t_0 - \epsilon_\delta, t_0 + \epsilon_\delta)}^{\mathbb{T}_n}| \rightarrow \infty$ if t_0 is a limit point of $(T_i)_{i \in \mathcal{J}}$. Therefore $f_{t_0}(\gamma, \alpha)$ is microergodic.

Proof of part (b) First, we consider the case when $(T_i)_{i \in \mathcal{J}}$ has two different limit points t_1 and t_2 . By part (a), $f_{t_1}(\gamma, \alpha)$ and $f_{t_2}(\gamma, \alpha)$ are microergodic. So, (γ, α) is microergodic by the following lemma (proved in appendix A.2.2):

Lemma 2.3. *Assume there exists $t_1 \neq t_2$ such that both $f_{t_1}(\gamma_1, \alpha_1) = f_{t_1}(\gamma_2, \alpha_2)$ and $f_{t_2}(\gamma_1, \alpha_1) = f_{t_2}(\gamma_2, \alpha_2)$. Then $(\gamma_1, \alpha_1) = (\gamma_2, \alpha_2)$.*

We now turn to the case when $(T_i)_{i \in \mathcal{J}}$ has only one limit point t_0 . We already know that $f_{t_0}(\gamma, \alpha)$ is microergodic, so we may assume that $f_{t_0}(\gamma_1, \alpha_1) = f_{t_0}(\gamma_2, \alpha_2)$, that is $g(t_0) = 0$. Denote $\mathcal{J}_{(t_0, \infty)} = \bigcup_{n=1}^{\infty} \mathcal{J}_{(t_0, \infty)}^{\mathbb{T}_n}$ and $\mathcal{J}_{[0, t_0]} = \bigcup_{n=1}^{\infty} \mathcal{J}_{[0, t_0]}^{\mathbb{T}_n}$. The condition in (b) implies that $\sum_{i \in \mathcal{J}_{(t_0, \infty)}} (T_i - t_0)^2 = \infty$ or $\sum_{i \in \mathcal{J}_{[0, t_0]}} (T_i - t_0)^2 = \infty$ or both. We now use lemma A.2 (appendix A.2.1) to select, for each n , a large set $\mathcal{C}_n$ of independent contrasts such that $\lim_n \sum_{C \in \mathcal{C}_n} (T_C - t_0)^2 = \infty$. Again, by (2.3) we have $r(\mathbb{T}_n) \geq r(\mathcal{C}_n) = \sum_{C \in \mathcal{C}_n} g(T_C)$, which we approximate below. If $t_0 = 0$, by Taylor expansion there exists $c(\alpha, T^*)$ such that $|e^{-2\alpha x} - 1 + 2\alpha x - 2\alpha^2 x^2 + \frac{4}{3}\alpha^3 x^3| \leq c(\alpha, T^*)x^4$ for every x satisfying $|x| < T^*$. Similarly, if $t_0 > 0$ there exists $c(\alpha, T^*)$ such that $|e^{-2\alpha x} - 1 + 2\alpha e^{-2\alpha t_0}(x - t_0) - 2\alpha^2 e^{-2\alpha t_0}(x - t_0)^2| \leq c(\alpha, T^*)x^3$ for every x satisfying $|x - t_0| < T^*$. In both cases, we can then write

$$r(\mathcal{C}_n) = \frac{1}{2} \sum_{C \in \mathcal{C}_n} (h_{t_0}(\alpha_1) - h_{t_0}(\alpha_2))^2 (T_C - t_0)^2 + o(T_C - t_0)^2$$

where $o(T_C - t_0)^2$ is uniform in n , $h_0(\alpha) = \alpha$ and $h_t(\alpha) = 2\alpha e^{-2\alpha t} / (1 - e^{-2\alpha t})$ for $t > 0$. Therefore $r(\mathcal{C}_n) \rightarrow \infty$ unless $(\gamma_1, \alpha_1) = (\gamma_2, \alpha_2)$. Hence (γ, α) is microergodic. $\square$

Theorem 2.3 part (b) gives a very general sufficient condition ensuring the microergodicity of (γ, α) . Unfortunately, it is not a necessary condition in general. To prove so, we consider the particular case when $\mathbb{T}_n$ is a symmetric tree, i.e. a tree in which each internal node is the parent of subtrees of identical shapes (see figure 2.5). We give below 3 examples in which $(T_i)_{i \in \mathcal{J}}$ has only one limit point t_0 and the condition in theorem 2.3 part (b) is violated. Two examples illustrate the non-microergodicity of (γ, α) , one in which $t_0 > 0$ and one in which $t_0 = 0$. In the last example the condition in (b) is violated, yet (γ, α) is microergodic.

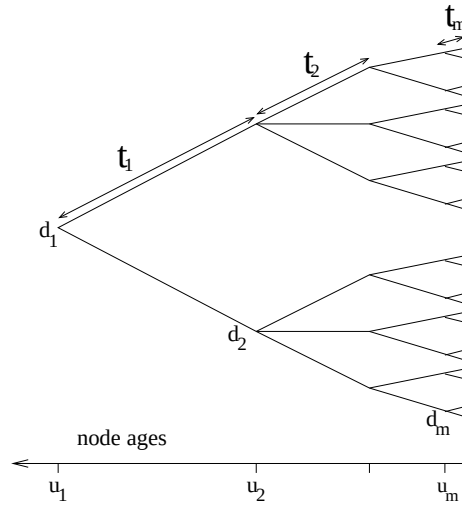


Figure 2.5: Symmetric trees with $m = 4$ levels.

Theorem 2.4. Consider the OU model (2.1) on symmetric trees with m levels and whose internal nodes at level i have d_i descendants along branches of length t_i .

- (a) Increasing node degrees. Consider a nested sequence of symmetric trees with a fixed number of levels m and fixed branch lengths $t_1, t_2, \dots, t_m$. Assume that the number of descendants d_m at the last level goes to infinity, but all other $d_1, \dots, d_{m-1}$ are fixed, so that $t_m > 0$ is the only limit point of $(T_i)_{i \in \mathcal{J}}$. Then (γ, α) is not microergodic.

(b) Dense sampling near the tips, or at distance t_0 from the tips. Consider a nested sequence of symmetric trees with a growing number of levels m , $d_k = d$ descendants at all levels $k \geq 1$ and such that the age of nodes at level k is $u_k = q^k + t_0$ for some $0 < q < 1$. Suppose that $dq^2 < 1$, to guarantee the violation of the condition in theorem 2.3 (b).

(i) If $t_0 = 0$ then (γ, α) is not microergodic.

(ii) If $t_0 > 0$ then (γ, α) is microergodic.

We discuss here the key ingredients of the proof. The technical details are provided in appendix A.2.2. Note that node ages are counted with multiplicity. Here u_i is the age of the $d_1 \dots d_{i-1}$ internal nodes at level i , with multiplicity $d_i - 1$ for each. Hence in part (a) $u_m = t_m$ is the only limit point. For a symmetric tree, the eigenvalues of the covariance matrix are $\gamma \lambda_k(\alpha)$ with multiplicity $d_1 \dots d_{k-1}(d_k - 1)$, where $\lambda_k(\alpha) = \sum_{i=k}^m d_{i+1} \dots d_m (e^{-2\alpha u_{i+1}} - e^{-2\alpha u_i})$ (appendix A.1). In (a), only the multiplicity of the smallest eigenvalue increases to infinity when the tree grows. If (γ_1, α_1) and (γ_2, α_2) share the same smallest eigenvalue, i.e. if $\gamma_1 \lambda_m(\alpha_1) = \gamma_2 \lambda_m(\alpha_2)$, then insufficient information is gained to distinguish between $P_{(\gamma_1, \alpha_1)}$ and $P_{(\gamma_2, \alpha_2)}$ when the tree grows. In (b), the eigenvalue with the largest multiplicity is also the smallest, $\gamma \lambda_m(\alpha) = \gamma(1 - e^{-2\alpha u_m})$. It converges to 0 when $t_0 = 0$ and to $\gamma(1 - e^{-2\alpha t_0}) > 0$ when $t_0 > 0$, yielding too little information in (i) when $t_0 = 0$, but more information to distinguish between $P_{(\gamma_1, \alpha_1)}$ and $P_{(\gamma_2, \alpha_2)}$ in (ii) when $t_0 > 0$.

2.3 Different convergence rates of ML estimators for different microergodic parameters

Section 2.2 suggests that the different parameters may not be estimated at the same rate. Indeed, if t_0 is the only limit point of internal node ages, then theorem 2.3 showed that $f_{t_0}(\gamma, \alpha)$ is microergodic regardless of whether condition in (b) is

satisfied or not. Therefore, the ML or REML estimate of $f_{t_0}(\gamma, \alpha)$ is expected to converge to the true value at a faster rate than the estimate of other parameters. In particular, for $t_0 = 0$ the ML estimate of σ^2 is expected to converge at a faster rate than that of α , which might not even be consistent. Here we identify cases with unequal convergence rates both theoretically and empirically.

2.3.1 Faster convergence of the REML estimator for $f_{t_0}(\gamma, \alpha)$ than for α and γ

We focus here on the symmetric tree growth model from Theorem 2.4 part (a) with nodes of increasing degrees, but we consider here the case when $\tilde{n} = n/d_m = d_1 \dots d_{m-1}$ increases indefinitely to ensure the microergodicity of γ and α . We show that the REML estimator of (γ, α) is consistent and asymptotic normally distributed. We further show that $f_{t_m}(\gamma, \alpha)$, which is microergodic regardless of the growth of $\tilde{n}$, is estimated at a faster rate than α or γ , which have stronger requirements to be microergodic.

Theorem 2.5. *Consider the asymptotic growth model from above with OU model (2.1). Denote $\nu = \gamma(1 - e^{-2\alpha t_m})$. Then the REML estimator $(\hat{\gamma}, \hat{\alpha})$ is consistent and $\begin{pmatrix} \sqrt{\tilde{n}}(\hat{\gamma} - \gamma) \\ \sqrt{\tilde{n}}(\hat{\alpha} - \alpha) \end{pmatrix} \xrightarrow{d} N\left(\mathbf{0}, \begin{pmatrix} 8\nu^2 & 0 \\ 0 & \nu_\alpha \end{pmatrix}\right)$.*

Moreover, if $n/\tilde{n} = d_m$ converges to infinity, then $\sqrt{\tilde{n}}(\hat{\gamma} - \gamma, \hat{\alpha} - \alpha)^t$ converges to a centered normal distribution and the asymptotic correlation between $\log \hat{\gamma}$ and $\log(1 - e^{-2\hat{\alpha} t_m})$ is -1 .

The proof in appendix A.2.2.1 gives the expression for ν_α . With increasing node degrees at m levels, the age of nodes at the last level t_m is the only limit point of $(T_i)_{i \in \mathcal{I}}$ if $\tilde{n}$ is bounded. The growth of $\tilde{n}$ ensures at least 2 limit points and the consistency of all parameters. Our results show that the rate of convergence is $\tilde{n}^{-1/2}$ for both $\hat{\alpha}$ and $\hat{\gamma}$. However, only one limit point (t_m) is required for the consistent estimation of $\nu = f_{t_m}(\gamma, \alpha)$, which is microergodic regardless of $\tilde{n}$.

Accordingly, the convergence rate of $\hat{v}$ is $n^{-1/2}$, which can be much faster than $\tilde{n}^{-1/2}$.

2.3.2 Simulations on a very large real tree

In this section we use simulations to investigate the properties of the MLE of the OU parameters on a real tree, comprising 4507 mammal species from [Bininda-Emonds et al. \(2007\)](#). Figure 2.6 shows the distribution of node ages for this

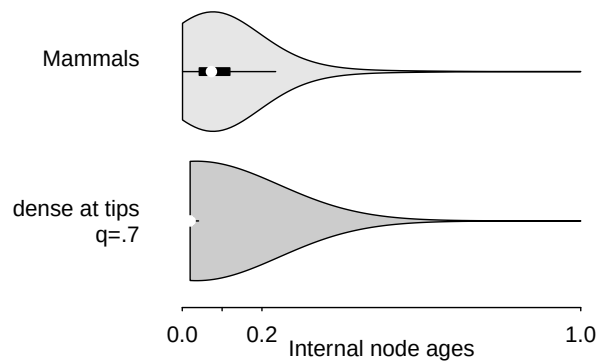


Figure 2.6: Distribution of node ages in the mammal tree (top) and in a symmetric tree (bottom) of similar size $n = 2^{12} = 4096$ with $d = 2$ at each level, levels being added near the tips at ages q^m . The value $q = .7 \approx 2^{-1/2}$ is the largest at which α and γ are not microergodic.

tree, and for a symmetric tree with dense sampling near the tips described in Theorem 2.4 (b), on which α and γ are not microergodic. Both distributions show a high density of very young nodes. Under the symmetric tree asymptotics with 0 as the only limit point, σ^2 is microergodic while (γ, α) might not be. Note that this is also the behavior under spatial infill asymptotics in dimension $d \leq 3$. For real trees like this mammal tree, therefore, we expect the MLE of σ^2 to converge quickly, and the MLE of α to converge more slowly or not at all. For various sample sizes from 10 to 4507 (full tree), we simulated data from the OU model with $\mu = 0$, $\gamma = 1$ and $\alpha = 0.1$, so $\sigma^2 = 0.2$. We created 20 sequences of six nested trees from 4507 to 10 leaves by randomly selecting subsets of leaves, conditional

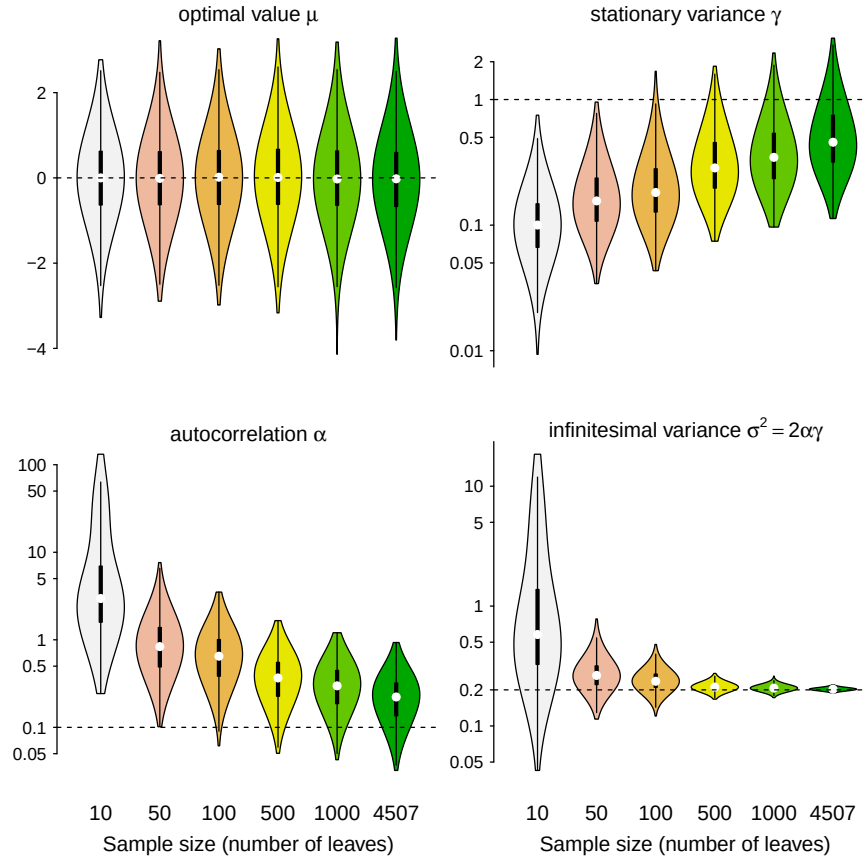


Figure 2.7: Violin plots showing the distribution of the MLE of μ , γ , α , and $\sigma^2 = 2\gamma\alpha$ on trees subsampled from the mammal phylogeny in [Bininda-Emonds et al. \(2007\)](#) with 2000 simulations at each sample size. The true values were $\mu = 0$, $\gamma = 1$, $\alpha = 0.1$, and $\sigma^2 = 0.2$.

on the root being the only common ancestor of the selected leaves to guarantee that all trees have the same height. Trees were all rescaled by the same factor to have height 1. For each tree, we simulated 100 data sets and computed the MLEs $\hat{\mu}$, $\hat{\gamma}$, and $\hat{\alpha}$. As expected, these simulations show that $\hat{\sigma}^2$ converges quickly to the true value while $\hat{\alpha}$ and $\hat{\gamma}$ do not (Figure 2.7). A strong bias is apparent for $\hat{\gamma}$ and $\hat{\alpha}$ even at the largest sample size (4507). Moreover, the correlation between $\log \hat{\alpha}$ and $\log \hat{\gamma}$ converges very fast to -1 (table 2.1). Also, the lower bound for

the variance of $\hat{\mu}$ is very close to the true variance (table 2.1). Therefore, this lower bound can be useful in practice at finite sample sizes.

Sample size	10	50	100	500	1000	4507
$\text{cor}(\log \hat{\alpha}, \log \hat{\gamma})$	-0.44	-0.927	-0.9674	-0.9938	-0.9971	-0.9993
$\text{var}(\hat{\mu})$	0.9007	0.8455	0.8499	0.8853	0.8789	0.8851
Lower bound (2.7)	0.8517	0.8472	0.8469	0.8468	0.8468	0.8468

Table 2.1: Correlation between $\log \hat{\alpha}$ and $\log \hat{\gamma}$ and variance of $\hat{\mu}$ from simulations. Last line: value of theoretical bound (2.7) for $\text{var}(\hat{\mu})$, averaged over 20 simulation subtrees.

2.4 Discussion

We considered an Ornstein-Uhlenbeck model of hierarchical autocorrelation and showed that the location parameter, here the mean μ , is not microergodic. We provided the lower bound for the variance of its ML estimator. In practice, these results could have important implications when scientists use OU hierarchical autocorrelation to detect a location shift, i.e. a change in μ along a branch of the tree (e.g. [Butler and King, 2004](#); [Lavin et al., 2008](#); [Monteiro and Nogueira, 2011](#)). Often times, the OU model is used with multiple adaptive optima whose placements on the tree are not fully known. Our results suggest that the power to detect such shifts may be low and mostly influenced by the effect size rather than by the sample size. An open question is whether the location of such shifts on the tree can be identified consistently with a growing number of tips.

We provide a general sufficient condition for the covariance parameters to be microergodic. Properties of infill asymptotics were recovered when 0 is the only limit point of internal node ages, i.e. when new nodes were added closer and closer to already existing tips. In this case, σ^2 is necessarily microergodic. This asymptotics can be appropriate for coalescent trees or when many species diverged recently from a moderate number of genera. We assume here the idealized situation with no error in the tree structure (topology and branch

lengths) and no data measurement error, leaving this for future work. With measurement error, the covariance matrix becomes $\gamma \mathbf{V}_\alpha + \sigma_e^2 \mathbf{I}$. The error variance σ_e^2 is called a nugget effect in spatial statistics. Measurement error with tree-structured correlation is rarely accounted for in applications (but see [Ives et al., 2007](#)).

For a general tree growth model, by using independent contrasts we can construct a consistent estimator for $f_{t_0}(\gamma, \alpha)$ where t_0 is any limit point of $(T_i)_{i \in \mathcal{I}}$. If $(T_i)_{i \in \mathcal{I}}$ has at least two limit points, then by lemma 2.3, we can construct a consistent estimator for (γ, α) . This proposed estimator is based on a restricted set of well-chosen contrasts, but it uses fewer contrasts and thus less information than the conventional REML estimator. We conjecture that if (γ, α) is microergodic, the REML estimator of (γ, α) is also consistent and asymptotically normal.

The microergodicity results suggest that parameters may not all be estimated at the same rate. Indeed, we show that the REML of α converges at a slower rate than $n^{-1/2}$ under a symmetric tree asymptotic framework. Similarly, our simulations suggest that the mammalian tree with 4507 species shares features similar to those under infill asymptotics (in low dimension) and under dense sampling near the tips of symmetric trees, where σ^2 can be consistently estimated but α and γ cannot. On the real tree, the MLE of σ^2 converges quickly to the true value while that of α and γ do not. This behavior may explain a lack of power to discriminate between a model of neutral evolution ($\alpha = 0$) versus a model with natural selection ($\alpha \neq 0$), as observed in [Cooper and Purvis \(2010\)](#). It would be interesting to know if most real trees share the “dense tip” asymptotic behavior, or how frequently a “dense root” asymptotic is applicable instead. Our results point to the distribution on node ages as indicative of the most appropriate asymptotic regime.

3 CONSISTENCY AND $\sqrt{n}$ -CONSISTENCY ON TREES WITH UNBOUNDED HEIGHT

Abstract Diffusion processes on trees are commonly used in evolutionary biology to model the joint distribution of continuous traits, such as body mass, across species. Estimating the parameters of such processes from tip values presents challenges because of the intrinsic correlation between the observations produced by the shared evolutionary history, thus violating the standard independence assumption of large-sample theory. For instance, [Ho and Ané \(2013a\)](#) recently proved that the mean (also known in this context as selection optimum) of an Ornstein-Uhlenbeck process on a tree cannot be estimated consistently from an increasing number of tip observations if the tree height is bounded. Here, using a fruitful connection to the so-called reconstruction problem in probability theory, we study the convergence rate of parameter estimation in the unbounded height case. For the mean of the process, we provide a necessary and sufficient condition for the consistency of the maximum likelihood estimator (MLE) and establish a phase transition on its convergence rate in terms of the growth of the tree. In particular we show that a loss of $\sqrt{n}$ -consistency (i.e., the variance of the MLE becomes $\Omega(n^{-1})$, where n is the number of tips) occurs when the tree growth is larger than a threshold related to the phase transition of the reconstruction problem. For the covariance parameters, we give a novel, efficient estimation method which achieves $\sqrt{n}$ -consistency under natural assumptions on the tree.

Funding This work was funded in part by NSF grant DMS-1106483.

3.1 Introduction

Analysis of data collected from multiple species presents challenges because of the intrinsic correlation produced by the shared evolutionary history. This dependency structure can be modeled by assuming that the traits of interest

evolved along a phylogeny according to a stochastic process. Two commonly used processes for continuous traits, such as body mass, are Brownian motion (BM) and the Ornstein-Uhlenbeck (OU) process. BM is used to model neutral evolution, with no favored direction (see e.g. [Felsenstein, 1985](#)). On the other hand, the OU process can account for natural selection using two extra parameters: a “selection optimum” μ towards which the process is attracted and a “selection strength” α ([Hansen, 1997](#)). The OU process has a stationary distribution, which is Gaussian with mean μ and variance $\gamma = \sigma^2/2\alpha$. The presence of natural selection can be detected by testing whether $\alpha > 0$ (e.g. [Harmon et al., 2010](#)). Changes in μ across different groups of organisms are used to correlate changes in selection regime with changes in behavior or environmental conditions (see e.g. [Butler and King, 2004](#); [Bartoszek et al., 2012](#)). For instance, the optimal body size μ might be different for terrestrial animals than for birds and bats. In practice, μ , α and the infinitesimal variance σ^2 (or stationary variance γ) are estimated from data on extant species. In other words, only data at the tips of the tree are available. The process at internal nodes and edges is unobserved. Also, the tree is reconstructed independently from external and abundant data, typically from DNA sequences. In practice there can be some uncertainty about a few nodes in the tree, but we assume here that the tree is known without error.

The OU process on a tree has been used extensively in practice (see e.g. [Butler and King, 2004](#); [Cooper and Purvis, 2010](#); [Brawand et al., 2011](#); [Rohlf et al., 2014](#)), but very few authors have studied convergence rates of available estimators. Recently [Ho and Ané \(2013a\)](#) showed that if the tree height is bounded as the sample size goes to infinity, no estimator for μ can ever be consistent. This is because μ is not “microergodic”: the distribution P_μ of the whole observable process $(Y_i)_{i \geq 1}$ at the tips of the tree is such that P_{μ_1} and P_{μ_2} are not orthogonal for any values $\mu_1 \neq \mu_2$, if the tree height is bounded. This boundedness assumption does not hold for common models of evolutionary trees however, such as the pure-birth (Yule) process ([Yule, 1925](#)). We consider here the case of an unbounded tree height. We study the consistency and convergence rates of several estimators, including some novel estimators, using tools from

the literature on the reconstruction problem in probability theory. In particular we relate the convergence rates of these estimators to the growth rate of the phylogeny. This connection is natural given that the growth rate (and the related branching number) is known to play an important role in the analysis of a variety of stochastic processes including random walks, percolation and ancestral state reconstruction on trees (Peres, 1999). In particular we leverage a useful characterization of the variance of linear estimators in terms of electrical networks.

Main results Section 3.3 presents the asymptotic properties of two common estimators for μ : the sample mean and the maximum likelihood estimator (MLE). Conditional on the tree, the MLE $\hat{\mu}_{\text{ML}}$ is known to be the best linear unbiased estimator for μ assuming that α is known. (The assumption of known α is proved not to be restrictive for our convergence rate results if α can be well estimated.) In fact, we give an example when $\hat{\mu}_{\text{ML}}$ performs significantly better than the sample mean, which is not consistent in that particular case. In one of our main results, we identify a necessary and sufficient condition for the consistency of $\hat{\mu}_{\text{ML}}$. We also derive a phase transition on its convergence rate, which drops from $\sqrt{n}$ -consistency (i.e. the variance is $O(n^{-1})$) to a lower rate, n being the number of samples (i.e. tip observations). This phase transition depends on the growth rate of the tree. Tree growth measures the rate at which new leaves arise as the tree height increases (see Section 3.2 for a formal definition). Roughly, when the growth rate is below 2α , we show that $\sqrt{n}$ -consistency holds. This is intuitive as a lower growth rate means lower correlations between the leaf states. On the other hand, when the growth rate is above 2α implying a sample size $n \gg e^{2\alpha T}$, i.e. when the tree is sufficiently “bushy,” then the “effective sample size” is reduced to $n^{\text{eff}} = e^{2\alpha T}$ and the $\sqrt{n}$ -consistency of $\hat{\mu}_{\text{ML}}$ is lost.

In Section 3.4, we provide novel, efficient estimators for the other two parameters, α and γ , which achieve $\sqrt{n}$ -consistency and do not require the knowledge of μ . Interestingly, the $\sqrt{n}$ -consistency in this case is not affected by growth rate, unlike the case of the MLE for μ .

Our main results are stated formally and further discussed in Section 3.2, after

necessary definitions.

Related work [Bartoszek and Sagitov \(2012\)](#) obtained a corresponding phase transition for the convergence rate of the sample mean to estimate μ , assuming a Yule process for the tree. Phase transitions for the convergence rate of some U-statistics have also been obtained for the OU model when the tree follows a supercritical branching process ([Adamczak and Miłoś, 2011a,b](#)). A main difference between these studies and our work is that we assume that the tree is known. Even though tree-free estimators are the only practical options when the tree is unknown, this situation is now becoming rare due to the ever-growing availability of sequence data for building trees. For instance [Crawford and Suchard \(2013\)](#) acknowledge that “as evolutionary biologists further refine our knowledge of the tree of life, the number of clades whose phylogeny is truly unknown may diminish, along with interest in tree-free estimation methods.”

As we mentioned, related phase transitions have been obtained for other processes on trees. For instance, the growth rate of the tree determines whether the state at the root can be reconstructed better than random for a binary symmetric channel on a binary tree (see e.g. [Evans et al., 2000](#), and references therein). In a recent result, [Mossel and Steel \(2014\)](#) established a transition for ancestral state reconstruction by majority rule for the binary symmetric model on a Yule tree at the same critical point as above. Note that majority rule is a tree-free estimator like the sample mean in [Bartoszek and Sagitov \(2012\)](#), but adapted to discrete traits. In the context of the OU model, [Mossel et al. \(2013\)](#) obtained a phase transition for estimating the ancestral state at the root, with the same critical growth rate we derive in our results.

3.2 Definitions and statements of results

In this section, we state formally and further explain our main results. First, we define our model and describe the setting in which our results are proved.

Notation For a vector $\mathbf{v}$ and matrix A , $\mathbf{v}'$ and A' denote the transposes. We let $\mathbf{0}$ and $\mathbf{1}$ denote the all-zeros and all-ones vectors respectively (with the size dictated by the context).

3.2.1 Model

Our main model is a stochastic process on a species tree $\mathbb{T}$. Let $\mathbb{T} = (\mathcal{E}, \mathcal{V})$ be a finite tree with leaf set $\mathcal{L} = \{1, \dots, n\}$ and root ρ . The leaves typically correspond to extant species. We think of the edges of $\mathbb{T}$ as being oriented away from the root. To each edge (or branch) $b \in \mathcal{E}$ of the tree is associated a positive length $|b| > 0$ corresponding to the time elapsed between the endpoints of b . For any two vertices $u, v \in \mathcal{V}$, we denote by d_{uv} the distance between u and v in $\mathbb{T}$, that is, the sum of the branch lengths on the unique path between u and v . We assume that the species tree is *ultrametric*, that is, that the distance from the root to every leaf is the same. It implies that, for any two tips $i, j \in \mathcal{L}$, d_{ij} is twice the time to the most recent common ancestor of i and j from the leaves. We let T be the height of $\mathbb{T}$, that is, the distance between the root and any leaf, and we define $t_{ij} = T - \frac{d_{ij}}{2}$. *Throughout we assume that the species tree is known.*

We consider an Ornstein-Uhlenbeck (OU) process on $\mathbb{T}$. That is, on each branch of $\mathbb{T}$, we have a diffusion

$$dY_t = -\alpha(Y_t - \mu)dt + \sigma dB_t,$$

where B_t is a standard Brownian motion (BM). In the literature on continuous traits, Y_t is known as the response variable, μ is the selection optimum, $\alpha > 0$ is the selection strength, $\sigma > 0$ is the scale parameter. We assume that the root value follows the stationary Gaussian distribution $\mathcal{N}(\mu, \gamma)$, where $\gamma = \frac{\sigma^2}{2\alpha}$. At each branching point, we run the process independently on each descendant edge starting from the value at the branching. Equivalently, the observations $\mathbf{Y} = (Y_\ell)_{\ell \in \mathcal{L}}$ at the tips of the tree are Gaussian with mean μ and variance matrix $\Sigma = \gamma \mathbf{V}_{\mathbb{T}}$ where

$$(\mathbf{V}_{\mathbb{T}})_{ij} = e^{-\alpha d_{ij}}. \quad (3.1)$$

We assume throughout that α , μ and σ are the same on every branch of $\mathbb{T}$. We will specify below whether these parameters are known, depending on the context.

Parameter estimators Our interest lies in estimating the parameters of the model, given $\mathbb{T}$, from a sample of $\mathbf{Y}$. In addition to proposing new estimators for α and σ , we study common estimators of μ . In particular we consider the empirical average at the tips

$$\bar{Y} = \frac{1}{n} \sum_{\ell \in \mathcal{L}} Y_\ell.$$

Also, writing the log-likelihood as

$$C - \frac{1}{2}(\mathbf{y} - \mu\mathbf{1})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mu\mathbf{1}) = C - \frac{1}{2}\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\mu^2\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1} + \mu\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{y},$$

where C does not depend on μ (and $\boldsymbol{\Sigma}$ is symmetric, positive definite), we note that the MLE of μ given the tree and α is

$$\hat{\mu}_{\text{ML}} = (\mathbf{1}'\mathbf{V}_{\mathbb{T}}^{-1}\mathbf{1})^{-1}\mathbf{1}'\mathbf{V}_{\mathbb{T}}^{-1}\mathbf{Y},$$

which is the well-known generalized least squares estimator for the linear regression problem

$$\mathbf{Y} = \mathbf{1}\mu + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is multivariate normal with known covariance matrix $\boldsymbol{\Sigma}$ (see e.g. [Anderson, 1984](#)). Note that the mean squared error is given by

$$\text{Var}_{\mathbb{T}}[\hat{\mu}_{\text{ML}}] = (\mathbf{1}'\mathbf{V}_{\mathbb{T}}^{-1}\mathbf{1})^{-2}\mathbf{1}'\mathbf{V}_{\mathbb{T}}^{-1}\boldsymbol{\Sigma}(\mathbf{V}_{\mathbb{T}}^{-1})'\mathbf{1} = \gamma(\mathbf{1}'\mathbf{V}_{\mathbb{T}}^{-1}\mathbf{1})^{-1}.$$

We drop the $\mathbb{T}$ in $\text{Var}_{\mathbb{T}}$ when the tree is clear from the context.

The estimators $\bar{Y}$ and $\hat{\mu}_{\text{ML}}$ are both linear estimators. More generally, for any weight vector $\boldsymbol{\theta} = (\theta_\ell)_{\ell \in \mathcal{L}}$ with $\boldsymbol{\theta}'\mathbf{1} = 1$, the following

$$Y_{\boldsymbol{\theta}} = \sum_{\ell \in \mathcal{L}} \theta_\ell Y_\ell,$$

is an unbiased estimator of μ . It is useful to think of the MLE in this context as an unbiased linear estimator minimizing the mean squared error (that is, a best linear unbiased estimator), which follows from the Gauss-Markov Theorem. Indeed, for any $\boldsymbol{\theta} = (\theta_\ell)_{\ell \in \mathcal{L}}$ with $\boldsymbol{\theta}'\mathbf{1} = 1$,

$$\text{Var}[Y_{\boldsymbol{\theta}}] = \boldsymbol{\theta}'\boldsymbol{\Sigma}\boldsymbol{\theta},$$

which is minimized if $\boldsymbol{\Sigma}\boldsymbol{\theta} = \lambda\mathbf{1}$ for some λ , which leads to the optimal choice

$$\boldsymbol{\theta}^* = (\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{1}$$

and $Y_{\boldsymbol{\theta}^*} = \hat{\mu}_{\text{ML}}$.

Lemma 3.1 (Variational formulation). *The MLE of μ given α and $\mathbb{T}$ minimizes the mean squared error among all linear unbiased estimators.*

Example 3.2 (Star tree). *Let $\mathbb{T}$ be a star tree with n leaf edges of length T emanating from the root. By symmetry, $\mathbf{1}$ is an eigenvector of $\boldsymbol{\Sigma}$ with eigenvalue $\lambda = \gamma[1 + (n-1)e^{-2\alpha T}]$. Hence, $\mathbf{1}$ is also an eigenvector of $\boldsymbol{\Sigma}^{-1}$ with eigenvalue λ^{-1} and*

$$\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1} = n\lambda^{-1},$$

so that $\boldsymbol{\theta}^* = \mathbf{1}/n$, that is, $\hat{\mu}_{\text{ML}} = \bar{Y}$, and

$$\text{Var}[\hat{\mu}_{\text{ML}}] = \frac{1}{n^2} \cdot n \cdot \lambda = \gamma \left[e^{-2\alpha T} + \frac{1 - e^{-2\alpha T}}{n} \right]. \quad (3.2)$$

3.2.2 Asymptotic setting

Our results are asymptotic. Specifically, we consider sequences of trees $\mathcal{T} = (\mathbb{T}_k)_{k \geq 1}$ with *fixed* parameters α, μ, σ . For $k \geq 1$, let n_k be the number of leaves in $\mathbb{T}_k$ and T_k be the height of $\mathbb{T}_k$. As before, we denote the leaf set of $\mathbb{T}_k$ as $\mathcal{L}_k = [n_k]$.

Assumption 1 (Unboundedness). *Throughout we assume that $n_k \leq n_{k+1}$ and $T_k \leq T_{k+1}$, and that $n_k \rightarrow +\infty$ and $T_k \rightarrow +\infty$ as $k \rightarrow +\infty$.*

For such a sequence of trees and a corresponding sequence of estimators, say X_k , we study various asymptotic properties of X_k . In this context, the following definitions are needed.

Definition 1 (Notions of convergence). *We say that $(X_k)_{k \geq 1}$ converges in probability to X if*

$$\forall \epsilon > 0, \lim_{k \rightarrow \infty} \mathbb{P}[|X_k - X| \geq \epsilon] = 0.$$

We denote this as $|X_k - X| = o_p(1)$. Moreover, we write $|X_k - X| = o_p(a_k)$ if $a_k^{-1}|X_k - X| = o_p(1)$. We say that $(X_k)_k$ is bounded in probability if for any $\delta > 0$, there exists $M_\delta > 0$ such that

$$\forall \delta > 0, \sup_k \mathbb{P}[|X_k| \geq M_\delta] < \delta.$$

We denote this as $|X_k| = O_p(1)$. Moreover, we write $|X_k| = O_p(a_k)$ if $a_k^{-1}|X_k| = O_p(1)$.

Definition 3.3 (Consistency). *Let $(X_k)_k$ be a sequence of estimators for a parameter x . We say that $(X_k)_k$ is consistent for x if $|X_k - x| = o_p(1)$. For $\beta > 0$, we say that $(X_k)_k$ is (n_k^β) -consistent for x if $|X_k - x| = O_p(n_k^{-\beta})$.*

From Chebyshev's inequality, we immediately get:

Lemma 3.4 (Rate of convergence: Upper bound). *Let $(X_k)_k$ be a sequence of unbiased estimators for a parameter x . If $\text{Var}[X_k] = O(n_k^{-2\beta})$ for some $\beta > 0$, then $|X_k - x| = O_p(n_k^{-\beta})$.*

Proof. Since $\text{Var}[X_k] = O(n_k^{-2\beta})$, we can choose $M_\delta > 0$ for every $\delta > 0$ such that

$$\sup_k \frac{n_k^{2\beta} \text{Var}[X_k]}{M_\delta^2} < \delta.$$

Applying Chebyshev's inequality, we have

$$\sup_k \mathbb{P} \left[n_k^\beta |X_k - x| \geq M_\delta \right] \leq \sup_k \frac{n_k^{2\beta} \text{Var}[X_k]}{M_\delta^2} < \delta.$$

□

For the other direction:

Lemma 3.5 (Rate of convergence: Lower bound). *Let $(X_k)_k$ be a sequence of unbiased estimators for a parameter x , such that $X_k \sim \mathcal{N}(x, \sigma_k^2)$. For $\beta > 0$, if*

$$\limsup_k \frac{\sigma_k^2}{n_k^{-2\beta}} = +\infty,$$

then for all $M > 0$

$$\limsup_k \mathbb{P} \left[n_k^\beta |X_k - x| > M \right] = 1,$$

that is, $(X_k)_k$ is not (n_k^β) -consistent.

Proof. Note that $\sigma_k^{-1}(X_k - x) \sim \mathcal{N}(0, 1)$. □

Example 3.6 (Star tree sequence: A first phase transition). *Let $\mathcal{T} = (\mathbb{T}_k)_k$ be a sequence of star trees with $n_k \rightarrow +\infty$ and $T_k \rightarrow +\infty$, with parameters α, μ, σ . Let $\hat{\mu}_{\text{ML}}^{(k)}$ be the MLE for μ given α on $\mathbb{T}_k$. Then, from (3.2) in Example 3.2,*

$$\text{Var}[\hat{\mu}_{\text{ML}}^{(k)}] = \gamma \left[e^{-2\alpha T_k} + \frac{1 - e^{-2\alpha T_k}}{n_k} \right] \rightarrow 0,$$

and the MLE (and $\bar{Y}$) is consistent for μ . Furthermore, if

$$\liminf_k \frac{2\alpha T_k}{\log n_k} > 1,$$

then

$$n_k \text{Var}[\hat{\mu}_{\text{ML}}^{(k)}] \leq \gamma [n_k e^{-2\alpha T_k} + 1] = \gamma \exp \left(\log n_k \left(1 - \frac{2\alpha T_k}{\log n_k} \right) \right) + \gamma = O(1)$$

and the MLE is $\sqrt{n_k}$ -consistent by Lemma 3.4. On the other hand, if

$$\liminf_k \frac{2\alpha T_k}{\log n_k} < 1,$$

then

$$n_k \text{Var}[\hat{\mu}_{\text{ML}}^{(k)}] \geq \gamma [n_k e^{-2\alpha T_k}] = \gamma \exp \left(\log n_k \left(1 - \frac{2\alpha T_k}{\log n_k} \right) \right),$$

which goes to $+\infty$ along a subsequence, and the MLE is not $\sqrt{n_k}$ -consistent by Lemma 3.5.

Special cases The tree of life naturally gives rise to two types of tree sequences. If one imagines sampling an increasing number of contemporary species, one obtains a nested sequence, defined as follows.

Definition 3.7 (Nested sequence). *A sequence of trees $(\mathbb{T}_k)_k$ is nested if, for all k , $n_k = k$ and $\mathbb{T}_k$ restricted to $[k-1]$ is identical to $\mathbb{T}_{k-1}$ as an ultrametric.*

Example 3.8 (Caterpillar sequence). *Let $(t_k)_k$ be a sequence of nonnegative numbers such that $\limsup_k t_k = +\infty$. Let $\mathbb{T}_1$ be a one-leaf star with height $T_1 = t_1$. For $k > 1$, let $\mathbb{T}_k$ be the caterpillar-like tree obtained by adding a leaf edge with leaf k to $\mathbb{T}_{k-1}$ at height t_k on the path between 1 and the root of $\mathbb{T}_{k-1}$, if $t_k \leq T_{k-1}$. If instead $t_k > T_{k-1}$, create a new root at height t_k with an edge attached to the root of $\mathbb{T}_{k-1}$ and an edge attached to k .*

If, instead, one is modeling the growth of the tree of life in time, one obtains a growing sequence, defined as follows. Let $\mathbb{T}_0$ be a rooted infinite tree of bounded degree, with branch lengths and no leaves. Think of the branches of $\mathbb{T}_0$ as a continuum of *points* whose distance from the endpoints grows linearly. Then, for $t \geq 0$, we define $\mathcal{B}_t(\mathbb{T}_0)$ as the tree made of the set of points of $\mathbb{T}_0$ at distance at most t from the root.

Definition 3.9 (Growing sequence). *A sequence of trees $(\mathbb{T}_k)_k$ is a growing sequence of trees if there is an infinite tree $\mathbb{T}_0$ as above and an increasing sequence of non-negative reals $(t_k)_k$ such that $\mathbb{T}_k$ is isomorphic to $\mathcal{B}_{t_k}(\mathbb{T}_0)$ as an ultrametric.*

Example 3.10 (Yule sequence). Let $\mathbb{T}_0$ be a tree generated by a pure-birth (Yule) process with rate $\lambda > 0$: starting with one lineage, each current lineage splits independently after an exponential time with mean λ^{-1} (see e.g. [Semple and Steel, 2003](#)). For any (possibly random) sequence of increasing non-negative reals (t_k) with $t_k \rightarrow +\infty$, $\mathcal{B}_{t_k}(\mathbb{T}_0)$ (that is, $\mathbb{T}_0$ run up to time t_k), forms a growing sequence.

Growth Our asymptotic results depend on how fast the tree grows. We use several standard notions of growth, which play an important role in random walks, percolation and ancestral state reconstruction on trees (see e.g. [Peres, 1999](#)). Fix a tree sequence $\mathcal{T} = (\mathbb{T}_k)_k$ with heights $(T_k)_k$ and numbers of tips $(n_k)_k$.

Definition 3.11 (Growth). The lower growth and upper growth of $\mathcal{T}$ are defined respectively as

$$\underline{\Lambda}^g = \liminf_k \frac{\log n_k}{T_k},$$

and

$$\overline{\Lambda}^g = \limsup_k \frac{\log n_k}{T_k}.$$

In case of equality we define the growth $\Lambda^g = \underline{\Lambda}^g = \overline{\Lambda}^g$. (Note that our definition differs slightly from [Peres \(1999\)](#) in that we consider the “exponential rate” of growth.)

That is, for all $\epsilon > 0$, eventually

$$e^{(\underline{\Lambda}^g - \epsilon)T_k} \leq n_k \leq e^{(\overline{\Lambda}^g + \epsilon)T_k},$$

and along appropriately chosen subsequences

$$n_{k_j} \geq e^{(\overline{\Lambda}^g - \epsilon)T_{k_j}},$$

and

$$n_{k'_j} \leq e^{(\underline{\Lambda}^g + \epsilon)T_{k'_j}}.$$

We also need a stronger notion of growth. For a tree $\mathbb{T}$, thinking of the branches of $\mathbb{T}$ as a continuum of *points*, a cutset π is a set of points of $\mathbb{T}$ such that

all paths from the root to a leaf must cross π . Let Π^k be the set of cutsets of $\mathbb{T}_k$.

Definition 3.12 (Branching number). *The branching number of $\mathcal{T}$ is defined as*

$$\Lambda^b = \sup \left\{ \Lambda \geq 0 : \inf_{k, \pi \in \Pi^k} \sum_{x \in \pi} e^{-\Lambda \delta_k(\rho, x)} > 0 \right\},$$

where $\delta_k(\rho, x)$ is the length of the path from the root to x in $\mathbb{T}_k$. (Again, unlike [Peres \(1999\)](#), we consider the exponential rate of branching.)

Because the leaf set $\mathcal{L}_k$ forms a cutset, it holds that

$$\Lambda^b \leq \underline{\Lambda}^g \leq \overline{\Lambda}^g.$$

Unlike the growth, the branching number takes into account aspects of the “shape” of the tree.

Example 3.13 (Star tree sequence, continued). *Consider again the setup of [Example 3.6](#). The infimum*

$$\inf_{\pi \in \Pi^k} \sum_{x \in \pi} e^{-\Lambda \delta(\rho, x)},$$

is achieved by taking $\pi = \mathcal{L}_k$. Hence $\Lambda^b = \underline{\Lambda}^g$. We showed in [Example 3.6](#) that the MLE of μ given α is $\sqrt{n_k}$ -consistent if $\overline{\Lambda}^g < 2\alpha$, but not $\sqrt{n_k}$ -consistent if $\overline{\Lambda}^g > 2\alpha$.

Finally, we will need a notion of uniform growth.

Definition 3.14 (Uniform growth). *Let $\mathcal{T} = (\mathbb{T}_k)_k$ be a tree sequence. For any point x in $\mathbb{T}_k$, let $n_k(x)$ be the number of leaves below x and let $T_k(x)$ be the distance from x to the leaves. Then the uniform growth of $\mathcal{T}$ is defined as*

$$\Lambda^{ug} = \lim_{M \rightarrow +\infty} \sup_{k, x \in \mathbb{T}_k} \frac{\log n_k(x)}{T_k(x) \vee M}.$$

(The purpose of the M in the denominator is to alleviate boundary effects.)

3.2.3 Statement of results

We can now state our main results.

Results concerning the mean μ We first give a characterization of the consistency of the MLE of μ . In words, the MLE sequence is consistent if, in the limit, we can find arbitrarily many descendants, arbitrarily far away from the leaves. In particular this criterion implies that under Assumption 1 the MLE of μ is always consistent on nested and growing sequences. This theorem is proved in Section 3.3.2, along with a related result involving the branching number.

Theorem 3.1 (Consistency of $\hat{\mu}_{\text{ML}}$). *Let $(\mathbb{T}_k)_k$ be a sequence of trees satisfying Assumption 1. Let $\hat{\mu}_{\text{ML}}^{(k)}$ be the corresponding sequence of MLEs of μ given α . Denote by $\tilde{\pi}_t^k$ the cutset of $\mathbb{T}_k$ at time t away from the leaves and let T_k be the height of $\mathbb{T}_k$. Then $(\hat{\mu}_{\text{ML}}^{(k)})_k$ is consistent for μ if and only if for all $s \in (0, +\infty)$*

$$\liminf_k |\tilde{\pi}_s^k| = +\infty.$$

We further obtain bounds on the variance of the MLE. When the upper growth is above 2α , we show that the MLE of μ cannot be $\sqrt{n_k}$ -consistent. If further the branching number is above 2α , we give tight bounds on the convergence rate of the MLE. Roughly we show that, in the latter case, the variance behaves like $n_k^{2\alpha/\wedge^g}$. Or perhaps a more accurate way to put it is that the “effective number of samples” n_k^{eff} is $e^{2\alpha T_k}$, in the sense that $\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] = \Theta((n_k^{\text{eff}})^{-1})$. Example 3.29 shows that these bounds cannot be improved in general.

Theorem 3.2 (Convergence rate of $\hat{\mu}_{\text{ML}}$: Supercritical regime). *Let $(\mathbb{T}_k)_k$ be a tree sequence. If $\overline{\Lambda}^g > 2\alpha$, then for all $\epsilon > 0$ there is a subsequence $(k_j)_j$ along which*

$$\text{Var}_{\mathbb{T}_{k_j}}[\hat{\mu}_{\text{ML}}^{(k_j)}] \geq \gamma n_{k_j}^{-2\alpha/(\overline{\Lambda}^g - \epsilon)}. \quad (3.3)$$

In particular $(\hat{\mu}_{\text{ML}}^{(k)})_k$ is not $\sqrt{n_k}$ -consistent. If, further,

1. $\Lambda^b > 2\alpha$: then

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] = \Theta(e^{-2\alpha T_k}).$$

Moreover in terms of n_k , for all $\epsilon > 0$, there are constants $0 < C', C < +\infty$ such that

$$C' n_k^{-2\alpha/(\underline{\Lambda}^g - \epsilon)} \leq \text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] \leq C n_k^{-2\alpha/(\overline{\Lambda}^g + \epsilon)},$$

and, in addition to (3.3),

$$\exists \text{ subsequence } (k'_j)_j, \text{ s.t. } \text{Var}_{\mathbb{T}_{k'_j}}[\hat{\mu}_{\text{ML}}^{(k'_j)}] \leq \gamma n_{k'_j}^{-2\alpha/(\underline{\Lambda}^g + \epsilon)}.$$

2. $\Lambda^b < 2\alpha$: then, for all $\epsilon > 0$, there are constants $0 < C', C < +\infty$ such that

$$C' n_k^{-2\alpha/(\underline{\Lambda}^g - \epsilon)} \leq \text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] \leq C n_k^{-(\Lambda^b - \epsilon)/(\overline{\Lambda}^g + \epsilon)},$$

where the lower bound above holds provided $\underline{\Lambda}^g > 0$, and

$$\exists \text{ subsequence } (k'_j)_j, \text{ s.t. } \text{Var}_{\mathbb{T}_{k'_j}}[\hat{\mu}_{\text{ML}}^{(k'_j)}] \leq \gamma n_{k'_j}^{-(\Lambda^b - \epsilon)/(\underline{\Lambda}^g + \epsilon)}.$$

In the other direction, the picture is somewhat murkier. See Example 3.29. However, under extra regularity conditions, $\sqrt{n_k}$ -consistency can be established. In words, the growth of the tree must be sufficiently homogeneous.

Theorem 3.3 (Convergence rate of $\hat{\mu}_{\text{ML}}^{(k)}$: Subcritical regime). *Let $(\mathbb{T}_k)_k$ be a tree sequence with $\overline{\Lambda}^g < 2\alpha$. Then*

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] = \Omega(n_k^{-1}).$$

Further if:

1. [Equality of growth and branching number] $\Lambda^b = \overline{\Lambda}^g > 0$ then for all $\epsilon > 0$

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] = O(n_k^{-(1-\epsilon)}).$$

2. [Bounded uniform growth] $\Lambda^{\text{ug}} < 2\alpha$ then

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] = O(n_k^{-1}).$$

Theorems 3.2 and 3.3 are proved in Section 3.3.3. All our results on the estimation of μ leverage a useful characterization of the variance of linear estimators in terms of electrical networks. An analogous characterization is used in ancestral state reconstruction Peres (1999). Note that our results are not as clean as those obtained for ancestral state reconstruction. As Example 3.29 shows, estimation of μ is somewhat sensitive to the “homogeneity” of the growth.

In Section 3.3.4 we apply our results to the special cases of trees with bounded branch lengths and the Yule model. Finally, in Section 3.3.5, we show that our assumption that α is known is inconsequential provided a good estimate of α is available. Such an estimate is discussed next.

Results concerning the parameters α and γ Our main result for α and γ is a $\sqrt{n_k}$ -consistent estimator.

Theorem 3.4 (Estimating α and γ : $\sqrt{n_k}$ -consistency). *Let $\mathcal{T} = (\mathbb{T}_k)_k$ be a sequence of ultrametric trees satisfying Assumptions 1 and 2 (stated in Section 3.4.1). Then there is an estimator $(\hat{\alpha}_k, \hat{\gamma}_k)_k$ of (α, γ) such that $|\hat{\alpha}_k - \alpha| = O_p(n_k^{-1/2})$ and $|\hat{\gamma}_k - \gamma| = O_p(n_k^{-1/2})$.*

The proof, found in Section 3.4.2, is based on the common notion of contrasts. Assumption 2 ensures the existence of an appropriate set of such contrasts. The key point is that this extra assumption can be satisfied no matter what the growth and branching number are, indicating that the estimation of α and γ is unaffected by the growth of the tree unlike μ . Intuitively, μ is a more “global” parameter.

In Section 3.4.3 we apply this result to the special cases of trees with bounded branch lengths and the Yule model.

3.3 Estimating μ

In this section, we study several estimators of μ . In particular we derive the convergence rate of the MLE under conditions on the growth of the tree. For ease of presentation, we assume through most of this section that α is known. However, in Section 3.3.5 we discuss the sensitivity of the MLE to estimation errors on α .

3.3.1 Bounding the variance of the MLE

Fix an ultrametric species tree $\mathbb{T}$ with leaf set $\mathcal{L}$, number of tips $n = |\mathcal{L}|$, and root ρ . We also fix $\alpha > 0$.

A formula for the variance Let $\theta = (\theta_\ell)_{\ell \in \mathcal{L}}$, with $\theta' \mathbf{1} = 1$ and $\theta_\ell \in [0, 1]$ for all ℓ , and recall that

$$Y_\theta = \sum_{\ell \in \mathcal{L}} \theta_\ell Y_\ell$$

is an unbiased estimator of μ . By defining, for each branch b ,

$$\theta_b = \sum_{\ell \in \mathcal{L}} \mathbb{1}_{b \in p(\rho, \ell)} \theta_\ell, \quad (3.4)$$

where $p(\rho, \ell)$ is the path from ρ to ℓ , we naturally associate to the coefficients θ a flow on the edges of $\mathbb{T}$, defined as follows.

Definition 2 (Flow). A flow η is a mapping from the set of edges to the set of positive numbers such that, for every edge b , we have

$$\eta_b = \sum_{b' \in O_b} \eta_{b'}$$

where O_b is the set of outgoing edges stemming from b (with the edges oriented away from the root). Define $\|\eta\| = \sum_{b \in O_\rho} \eta_b$. We say that η is a unit flow if $\|\eta\| = 1$. We

extend η to vertices v in $\mathbb{T}$ by defining η_v as the flow on the edge entering v . Similarly, for a point x in $\mathbb{T}$, we let η_x be the flow on the corresponding edge or vertex.

For every edge b of $\mathbb{T}$, we set

$$R_b = (1 - e^{-2\alpha|b|})e^{2\alpha\delta(\rho,b)}$$

where $|b|$ is the length of b and $\delta(\rho, b)$ is the length of the path from the root to b (inclusive).

Lemma 3.15 (Variance formula). *For any unit flow θ from ρ to $\mathcal{L}$, we have*

$$\text{Var}[Y_\theta] = \gamma e^{-2\alpha T} \left(1 + \sum_{b \in E} R_b \theta_b^2 \right) \quad (3.5)$$

where E is the set of edges.

Proof. The proof follows from a computation of [Evans et al. \(2000\)](#). For every node u of the tree, by a telescoping argument,

$$e^{2\alpha\delta(\rho,u)} - 1 = \sum_{b \in p(\rho,u)} R_b \quad (3.6)$$

where $\delta(\rho, u)$ is the distance from ρ to u , and $p(\rho, u)$ is the path from ρ to u . Denote by $v \wedge w$ the most recent common ancestor of v and w . Then

$$\begin{aligned} \text{Var}[Y_\theta] &= \gamma \sum_{v,w \in \mathcal{L}} \theta_v \theta_w \frac{e^{-2\alpha T}}{e^{-2\alpha\delta(\rho, v \wedge w)}} \\ &= \gamma e^{-2\alpha T} \sum_{v,w \in \mathcal{L}} \theta_v \theta_w \left(1 + \sum_{b \in p(\rho, v \wedge w)} R_b \right) \\ &= \gamma e^{-2\alpha T} \left(1 + \sum_{b \in E} R_b \sum_{v,w \in \mathcal{L}} \mathbb{1}_{b \in p(\rho, v \wedge w)} \theta_v \theta_w \right) \end{aligned}$$

$$\begin{aligned}
&= \gamma e^{-2\alpha T} \left[1 + \sum_{b \in E} R_b \left(\sum_{v \in \mathcal{L}} \mathbb{1}_{b \in p(\rho, v)} \theta_v \right) \left(\sum_{w \in \mathcal{L}} \mathbb{1}_{b \in p(\rho, w)} \theta_w \right) \right] \\
&= \gamma e^{-2\alpha T} \left(1 + \sum_{b \in E} R_b \theta_b^2 \right),
\end{aligned}$$

where the second line follows from (3.6), the fourth line follows from $\mathbb{1}_{b \in p(\rho, v \wedge w)} = \mathbb{1}_{b \in p(\rho, v)} \mathbb{1}_{b \in p(\rho, w)}$, and the last line follows from (3.4). $\square$

Remark 3.16. Note that (3.5), which holds for a general $(\theta_\ell)_{\ell \in \mathcal{L}}$ extended to branches by (3.4), implies that it suffices to consider non-negative θ_ℓ 's when minimizing $\text{Var}[Y_\theta]$ under $\|\theta\| = 1$. Indeed assume $(\theta_\ell)_{\ell \in \mathcal{L}}$ contains negative values and consider the non-negative flow

$$\theta' = \frac{\theta_+}{\|\theta_+\|},$$

where θ_+ indicates the positive part component-wise. Because $\|\theta_+\| > 1$, we have $\theta'_\ell < |\theta_\ell|$ for all leaves ℓ and hence $\theta'_b < |\theta_b|$ for all branches b .

The following example will be useful below.

Example 3.17 (Spherically symmetric trees). Let $\mathbb{T}$ be a spherically symmetric, ultrametric tree, that is, a tree such that all vertices at the same graph distance from the root have the same number of outgoing edges, all of the same length. Let D_h , $h = 0, \dots, H-1$, be the out-degree of vertices at graph distance h (where $h = 0$ and $h = H$ correspond to the root and leaves respectively) and let τ_h be the corresponding branch length. Notice that $\beta_1^2 + \dots + \beta_d^2$, subject to $\beta_1 + \dots + \beta_d = 1$, is minimized at $\beta_1 = \dots = \beta_d = 1/d$. Hence, using Lemma 3.1 and arguing inductively from the leaves in (3.5), we see that $\hat{\mu}_{\text{ML}} = \bar{Y}$ in this case. The mean squared error is, by (3.5),

$$\begin{aligned}
\text{Var}[\hat{\mu}_{\text{ML}}] &= \gamma e^{-2\alpha T} \left[1 + \sum_{h=0}^{H-1} \left(\prod_{h'=0}^h D_{h'} \right) (1 - e^{-2\alpha \tau_h}) e^{2\alpha \sum_{h'=0}^h \tau_{h'}} \prod_{h'=0}^h \frac{1}{D_{h'}^2} \right] \\
&= \gamma e^{-2\alpha T} \left[1 + \sum_{h=0}^{H-1} (1 - e^{-2\alpha \tau_h}) \prod_{h'=0}^h \frac{e^{2\alpha \tau_{h'}}}{D_{h'}} \right]. \tag{3.7}
\end{aligned}$$

Finally, combining Lemmas 3.1 and 3.15, we obtain the key formula from which our results about the MLE of μ are derived.

Proposition 3.18 (Variance of $\hat{\mu}_{\text{ML}}$: Main formula). *Let $\mathbb{T}$ be an ultrametric tree with edge set E , leaf set $\mathcal{L}$, root ρ and height T . Let Θ be the set of unit flows from ρ to $\mathcal{L}$. Then*

$$\text{Var}[\hat{\mu}_{\text{ML}}] = \inf_{\theta \in \Theta} \gamma e^{-2\alpha T} \left(1 + \sum_{b \in E} R_b \theta_b^2 \right).$$

As detailed in Peres (1999), a species tree can be interpreted as an electrical network with resistance R_b on edge b . The minimum $\mathcal{R}_{\mathbb{T}}$ of $\sum_{b \in E} R_b \theta_b^2$ over unit flows (corresponding to the MLE) is known as the *effective resistance* of $\mathbb{T}$, which can be interpreted in terms of a random walk on the tree. See Peres (1999) for details.

For $0 \leq t \leq T$, let π_t be the set of points at distance t from the root (that is, the cutset corresponding to time t away from the root). Noting that

$$R_b = 2\alpha \int_{\delta(\rho,b)-|b|}^{\delta(\rho,b)} e^{2\alpha s} ds,$$

we get the following convenient formula:

Lemma 3.19 (Variance formula: Integral form). *For any unit flow θ from ρ to $\mathcal{L}$, we have*

$$\text{Var}[Y_{\theta}] = \gamma e^{-2\alpha T} \left[1 + 2\alpha \int_0^T e^{2\alpha s} \left(\sum_{x \in \pi_s} \theta_x^2 \right) ds \right].$$

As a first important application of Proposition 3.18 and Lemma 3.19, we show that the variance of the MLE of μ can be controlled by the branching number. The result is characterized by a transition at $\Lambda^b = 2\alpha$, similarly to Example 3.13.

Proposition 3.20 (Variance of $\hat{\mu}_{\text{ML}}$: Link to the branching number). *Let $\mathcal{T} = (\mathbb{T}_k)_k$ be a tree sequence with branching number $\Lambda^b > 0$. Then, for all $\Lambda < \Lambda^b$, there is $\mathcal{I}$ such*

that

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] \leq \begin{cases} \gamma \left(1 + \frac{2\alpha}{\mathcal{J}_\Lambda(2\alpha - \Lambda)}\right) e^{-\Lambda T_k}, & \text{if } \Lambda < 2\alpha, \\ \gamma \left(1 + \frac{2\alpha T_k}{\mathcal{J}_\Lambda}\right) e^{-2\alpha T_k}, & \text{if } \Lambda = 2\alpha, \\ \gamma \left(1 + \frac{2\alpha}{\mathcal{J}_\Lambda(\Lambda - 2\alpha)}\right) e^{-2\alpha T_k}, & \text{if } \Lambda > 2\alpha. \end{cases}$$

Proof. For $\Lambda < \Lambda^b$, let

$$\mathcal{J}_\Lambda = \inf_{k, \pi \in \Pi^k} \sum_{x \in \pi} e^{-\Lambda \delta_k(\rho, x)} > 0.$$

By the max-flow min-cut theorem (see e.g. [Lawler, 1976](#)), there is a flow $\eta^{(k)}$ on $\mathbb{T}_k$ with

$$\|\eta^{(k)}\| \geq \mathcal{J}_\Lambda \quad (3.8)$$

and

$$\eta_x^{(k)} \leq e^{-\Lambda \delta_k(\rho, x)}, \quad (3.9)$$

for all points x in $\mathbb{T}_k$. Normalize $\eta^{(k)}$ as $\theta^{(k)} = \eta^{(k)} / \|\eta^{(k)}\|$. By [Proposition 3.18](#) and [Lemma 3.19](#), for $\Lambda \neq 2\alpha$,

$$\begin{aligned} \text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] &\leq \gamma e^{-2\alpha T_k} \left[1 + 2\alpha \int_0^{T_k} e^{2\alpha s} \left(\sum_{x \in \pi_s^k} (\theta_x^{(k)})^2 \right) ds \right] \\ &\leq \gamma e^{-2\alpha T_k} \left[1 + 2\alpha \int_0^{T_k} e^{2\alpha s} \left(\sum_{x \in \pi_s^k} \theta_x^{(k)} \frac{e^{-\Lambda \delta_k(\rho, x)}}{\mathcal{J}_\Lambda} \right) ds \right] \\ &\leq \gamma e^{-2\alpha T_k} \left[1 + \frac{2\alpha}{\mathcal{J}_\Lambda} \int_0^{T_k} e^{(2\alpha - \Lambda)s} ds \right] \\ &= \gamma e^{-2\alpha T_k} \left[1 + \frac{2\alpha}{\mathcal{J}_\Lambda(2\alpha - \Lambda)} (e^{(2\alpha - \Lambda)T_k} - 1) \right] \\ &= \gamma \left[e^{-2\alpha T_k} + \frac{2\alpha}{\mathcal{J}_\Lambda(2\alpha - \Lambda)} (e^{-\Lambda T_k} - e^{-2\alpha T_k}) \right]. \end{aligned}$$

where the second line follows from (3.8) and (3.9), and the third line follows from the fact that $\delta_k(\rho, x) = s$ for $x \in \pi_s^k$ by definition and that $\sum_{x \in \pi_s^k} \theta_x^{(k)} = 1$.

Similarly if $\Lambda = 2\alpha$

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] \leq \gamma \left[e^{-2\alpha T_k} + \frac{2\alpha e^{-2\alpha T_k} T_k}{\mathcal{I}_\Lambda} \right].$$

□

Removing bottlenecks Examining (3.5), one sees that a natural bound on $\text{Var}[Y_\theta]$ is obtained by “splitting an edge” in $\mathbb{T}$.

Definition 3.21 (Edge splitting). *Let $\mathbb{T}$ be an ultrametric tree with edge set E . Let $b_0 = (x_0, y_0)$ be a branch in $\mathbb{T}$ (where x_0 is closer to the root) and let $b_i = (y_0, y_i)$, $i = 1, \dots, D$, be the outgoing edges at y_0 . The operation of splitting branch b_0 to obtain a new tree $\mathbb{T}'$ with edge set E' is defined as follows: remove $b_0, b_1, \dots, b_D$ from $\mathbb{T}$; add D new edges $b'_i = (x_0, y_i)$ of length $|b_0| + |b_i|$, $i = 1, \dots, D$ (see Figure 3.1). We call merging the opposite operation of undoing the above splitting.*

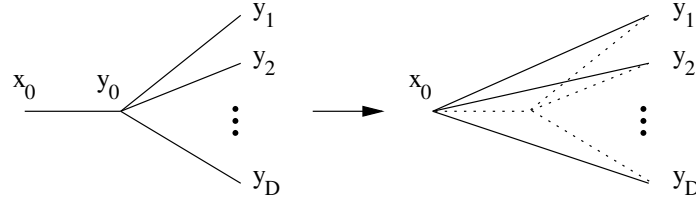


Figure 3.1: Edge splitting procedure.

Note that the number of tips in $\mathbb{T}$ and $\mathbb{T}'$ above are the same, and therefore we can use the same estimator Y_θ on both of them.

Lemma 3.22 (Splitting an edge). *Let $\mathbb{T}$ be an ultrametric tree, let b_0 be a branch in $\mathbb{T}$, and let $\mathbb{T}'$ be obtained from $\mathbb{T}$ by splitting b_0 . Then for any nonnegative $\theta = (\theta_\ell)_{\ell \in \mathcal{L}}$*

$$\text{Var}_{\mathbb{T}'}[Y_\theta] \leq \text{Var}_{\mathbb{T}}[Y_\theta].$$

Proof. We use the notation of Definition 3.21. Denote by $(\theta_b)_{b \in E}$ and $(\theta'_b)_{b \in E'}$ the flows associated to θ by (3.4) on $\mathbb{T}$ and $\mathbb{T}'$ respectively. For any branch b , except $b_0, b_1, \dots, b_D$ and $b'_1, \dots, b'_D$, we have $\theta_b = \theta'_b$, as the descendant leaves of b on $\mathbb{T}$ and $\mathbb{T}'$ are the same. Think of $b'_i = (x_0, y_i)$, $i = 1, \dots, D$, as being made of two consecutive edges $b''_i = (x_0, y'_i)$ and $b'''_i = (y'_i, y_i)$ with $|b''_i| = |b_0|$ and $|b'''_i| = |b_i|$ (and note, for sanity check, that $R_{b'_i} = R_{b''_i} + R_{b'''_i}$). Then, $\theta_{b_i} = \theta_{b'''_i}$ and $R_{b_i} = R_{b'''_i}$, and by (3.5)

$$\begin{aligned} \text{Var}_{\mathbb{T}}[Y_\theta] - \text{Var}_{\mathbb{T}'}[Y_\theta] &= R_{b_0} \theta_{b_0}^2 - \sum_{i=1}^D R_{b''_i} \theta_{b''_i}^2 \\ &= R_{b_0} \left(\sum_{i=1}^D \theta_{b''_i} \right)^2 - R_{b_0} \sum_{i=1}^D \theta_{b''_i}^2 \\ &\geq 0, \end{aligned}$$

where we used that $R_{b_0} = R_{b''_i}$ and the nonnegativity of the $\theta_{b''_i}$'s. $\square$

Comparing $\mathbb{T}$ to a star we then get:

Proposition 3.23 (Lower bound on the variance of $\hat{\mu}_{\text{ML}}$). *Let $\mathbb{T}$ be an ultrametric tree with n tips and height T . Then*

$$\text{Var}_{\mathbb{T}}[\hat{\mu}_{\text{ML}}] \geq \gamma \left(e^{-2\alpha T} + \frac{1 - e^{-2\alpha T}}{n} \right).$$

Proof. Split all edges in $\mathbb{T}$ by repeatedly applying Lemma 3.22 until a star tree with n leaves and height T is obtained. The result then follows from (3.2). $\square$

Proceeding in reverse:

Proposition 3.24 (Upper bound on the variance of $\hat{\mu}_{\text{ML}}$). *Let $\mathbb{T}$ be an ultrametric tree with height T . Recall that π_t be the set of points at distance t from the root. Then*

$$\text{Var}_{\mathbb{T}}[\hat{\mu}_{\text{ML}}] \leq \inf_{0 \leq t \leq T} \gamma \left(e^{-2\alpha(T-t)} + \frac{1 - e^{-2\alpha(T-t)}}{|\pi_t|} \right).$$

Proof. Let $0 \leq t \leq T$. For all points x in π_t , choose one descendant leaf ℓ_x of x and define θ as

$$\theta_\ell = \begin{cases} \frac{1}{|\pi_t|}, & \text{if } \ell = \ell_x \text{ for some } x, \\ 0, & \text{otherwise.} \end{cases}$$

Divide all branches crossing π_t into two branches meeting at π_t . Then merge all branches above π_t (that is, closer to the root) by repeatedly applying Lemma 3.22. By (3.5), removing all branches b with $\theta_b = 0$ does not affect the variance, and from Example 3.17 with $H = 2$, $D_0 = 1$, $D_1 = |\pi_t|$, $\tau_0 = t$, and $\tau_1 = T - t$, we get

$$\begin{aligned} \text{Var}_{\mathbb{T}}[\hat{\mu}_{\text{ML}}] &\leq \gamma e^{-2\alpha T} \left[1 + (1 - e^{-2\alpha t})e^{2\alpha t} + (1 - e^{-2\alpha(T-t)})e^{2\alpha t} \frac{e^{2\alpha(T-t)}}{|\pi_t|} \right] \\ &\leq \gamma \left[e^{-2\alpha(T-t)} + \frac{1 - e^{-2\alpha(T-t)}}{|\pi_t|} \right]. \end{aligned}$$

□

The two estimators $\hat{\mu}_{\text{ML}}$ vs. $\bar{Y}$ As an application of the previous corollary, we provide an example where $\hat{\mu}_{\text{ML}}$ performs significantly better than $\bar{Y}$. Roughly, the example shows that $\bar{Y}$ can perform poorly on asymmetric trees.

Example 3.25. Consider a caterpillar sequence $(\mathbb{T}_k)_k$, as defined in Example 3.8, with $t_{2m+1} = m$ and $t_{2m} = 1$ for all m , as shown in Figure 3.2. Let π_t^k be the time t cutset of $\mathbb{T}_k$ and let T_k be the height of $\mathbb{T}_k$. Note that $T_{2m+1} = T_{2m+2} = m$ and $|\pi_{m-1}^{2m+1}| = |\pi_{m-1}^{2m+2}| = m$. Therefore, by Proposition 3.24,

$$\max \{ \text{Var}_{\mathbb{T}_{2m+1}}[\hat{\mu}_{\text{ML}}], \text{Var}_{\mathbb{T}_{2m+2}}[\hat{\mu}_{\text{ML}}] \} \leq \gamma \left[e^{-2\alpha(m-1)} + \frac{1}{m} \right] \rightarrow 0,$$

as $m \rightarrow +\infty$, and hence $\hat{\mu}_{\text{ML}}$ is consistent. On the other hand, note that $\text{Cov}[Y_i, Y_j] \geq 0$ for all pairs of leaves i, j in $\mathbb{T}_k$. Therefore,

$$\text{Var}_{\mathbb{T}_{2m}}[\bar{Y}] = \frac{1}{4m^2} \text{Var} \left[\sum_{\ell=1}^{2m} Y_\ell \right] \geq \frac{1}{4m^2} \text{Var} \left[\sum_{i=1}^m Y_{2i} \right] = \frac{1}{4m^2} \sum_{i,j=1}^m \text{Cov}[Y_{2i}, Y_{2j}]$$

$$\geq \frac{1}{4m^2} m^2 \gamma e^{-2\alpha} = \frac{\gamma e^{-2\alpha}}{4}.$$

So, $\bar{Y}$ is not consistent.

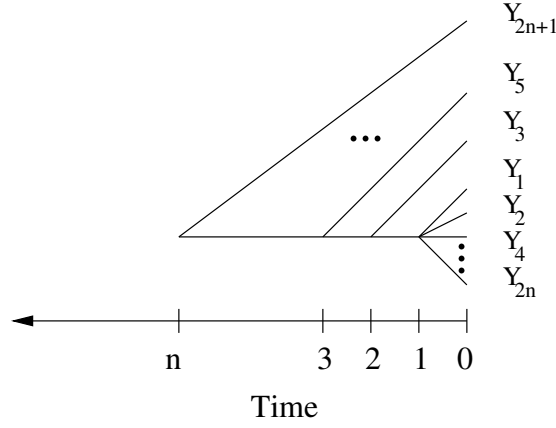


Figure 3.2: Example where the MLE $\hat{\mu}$ is consistent while $\bar{Y}$ is not.

3.3.2 Criterion for consistency of the MLE

In the previous subsection, we gave an example where $\bar{Y}$ is not consistent for μ , but $\hat{\mu}_{\text{ML}}$ is. Here we give a general criterion under which $\hat{\mu}_{\text{ML}}$ is consistent.

Theorem 3.1 (Consistency of $\hat{\mu}_{\text{ML}}$). *Let $(\mathbb{T}_k)_k$ be a sequence of trees satisfying Assumption 1. Let $\hat{\mu}_{\text{ML}}^{(k)}$ be the corresponding sequence of MLEs of μ given α . Denote by $\tilde{\pi}_t^k$ the cutset of $\mathbb{T}_k$ at time t away from the leaves and let T_k be the height of $\mathbb{T}_k$. Then $(\hat{\mu}_{\text{ML}}^{(k)})_k$ is consistent for μ if and only if for all $s \in (0, +\infty)$*

$$\liminf_k |\tilde{\pi}_s^k| = +\infty. \quad (3.10)$$

Proof. Assume (3.10) holds. From Proposition 3.24, for all s ,

$$\limsup_k \text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] \leq \limsup_k \gamma \left[e^{-2\alpha s} + \frac{1 - e^{-2\alpha s}}{|\tilde{\pi}_s^k|} \right] \leq \gamma e^{-2\alpha s}.$$

Taking s to $+\infty$ gives consistency. On the other hand, assume by contradiction that $(\hat{\mu}_{\text{ML}}^{(k)})_k$ is consistent but that

$$\liminf_k |\tilde{\pi}_s^k| < +\infty,$$

for some $s \in (0, +\infty)$. Let $(k_j)_j$ be the corresponding subsequence and L , the limit above. Divide all branches in $\mathbb{T}_{k_j}$ crossing $\tilde{\pi}_s^{k_j}$ into two branches meeting at $\tilde{\pi}_s^{k_j}$. Split edges in $\mathbb{T}_{k_j}$ above $\tilde{\pi}_s^{k_j}$ (closer to the root) repeatedly until the tree above $\tilde{\pi}_s^{k_j}$ forms a star. Let $\mathbb{T}'$ be the resulting tree, let $b'_1, \dots, b'_D$ be the branches emanating from the root, where $D \leq L$ by assumption, and let $\tilde{\pi}'$ be the cutset at time s from the leaves. For the unit flow θ' corresponding to the MLE on $\mathbb{T}'$, by Lemma 3.22 and counting only those edges above $\tilde{\pi}'$ in $\mathbb{T}'$ in (3.5), we have

$$\begin{aligned} \text{Var}_{\mathbb{T}_{k_j}}[\hat{\mu}_{\text{ML}}^{(k_j)}] &\geq \gamma e^{-2\alpha T_{k_j}} \left[1 + \left(1 - e^{-2\alpha(T_{k_j}-s)} \right) e^{2\alpha(T_{k_j}-s)} \sum_{i=1}^D (\theta'_{b'_i})^2 \right] \\ &\geq \gamma e^{-2\alpha T_{k_j}} \left[1 + \left(1 - e^{-2\alpha(T_{k_j}-s)} \right) \frac{e^{2\alpha(T_{k_j}-s)}}{D} \right] \\ &\geq \gamma \left[e^{-2\alpha T_{k_j}} + \left(1 - e^{-2\alpha(T_{k_j}-s)} \right) \frac{e^{-2\alpha s}}{L} \right], \end{aligned}$$

where we used the fact that $\beta_1^2 + \dots + \beta_D^2$, subject to $\beta_1 + \dots + \beta_D = 1$, is minimized at $\beta_1 = \dots = \beta_D = 1/D$. Since $T_{k_j} \rightarrow +\infty$ under Assumption 1,

$$\limsup_k \text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] \geq \gamma \frac{e^{-2\alpha s}}{L} > 0,$$

and we get a contradiction. □

Corollary 3.5 (Consistency: Nested sequence). *Let $(\mathbb{T}_k)_k$ be a nested sequence satisfying Assumption 1. Then the MLE for μ is consistent on $(\mathbb{T}_k)_k$.*

Proof. Let k_j be the subsequence such that $T_{k_{j+1}} > T_{k_j}$ for every j and $T_i = T_{k_j}$ for all $i = k_j + 1, \dots, k_{j+1} - 1$. Then, for all $s \in (0, +\infty)$, as k goes to $+\infty$ π_s^k eventually

contains all leaves k_j such that $T_{k_j} \geq s$. Since $T_k \rightarrow +\infty$ by Assumption 1, the result follows. $\square$

For nested sequences, it was shown in Ho and Ané (2013a) that Assumption 1 is necessary, as on bounded-height tree sequences the MLE of μ is not consistent.

Corollary 3.6 (Consistency: Growing sequence). *Let $(\mathbb{T}_k)_k$ be a growing sequence satisfying Assumption 1. Then the MLE for μ is consistent on $(\mathbb{T}_k)_k$.*

Proof. Fix $s \in (0, +\infty)$. For $L = 1, 2, \dots$, let k'_L be the smallest k such that $n_k \geq L$ and let k''_L be the smallest $k > k'_L$ such that $T_k \geq T_{k'_L} + s$. Then, for all $k \geq k''_L$, $|\pi_s^k| \geq L$. Letting L go to $+\infty$ gives the result. $\square$

More generally, Assumption 1 does not suffice, as the following example shows.

Example 3.26 (Consistency: Counter-example). *Using the notation of Example 3.17, let $(\mathbb{T}_{2m})_m$ be a sequence of spherically symmetric trees with $H = 2$, degrees $D_0^{(2m)} = 2$, $D_1^{(2m)} = m$, $\tau_0^{(2m)} = m - 1$, and $\tau_1^{(2m)} = 1$. Although $n_{2m} = 2m \rightarrow +\infty$ and $T_{2m} = m \rightarrow +\infty$, that is, Assumption 1, we have by (3.7)*

$$\begin{aligned} \text{Var}_{\mathbb{T}_{2m}}[\hat{\mu}_{\text{ML}}^{(2m)}] &= \gamma e^{-2\alpha m} \left[1 + (1 - e^{-2\alpha(m-1)}) \frac{e^{2\alpha(m-1)}}{2} + (1 - e^{-2\alpha}) \frac{e^{2\alpha m}}{2m} \right] \\ &\geq \gamma \frac{e^{-2\alpha}}{2} > 0, \end{aligned}$$

and the MLE is not consistent. Taking $s = 1$ in (3.10) explains why.

However, in general, the branching number does provide a simple, sufficient condition.

Proposition 3.27 (Consistency: Branching number condition). *Let $\mathcal{T} = (\mathbb{T}_k)_k$ be a tree sequence satisfying Assumption 1 with branching number Λ^b . Then $\Lambda^b > 0$ suffices for the consistency of the MLE of μ . Note that this condition is independent of α .*

Proof. By Proposition 3.20, taking $0 < \Lambda < \min\{2\alpha, \Lambda^b\}$,

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] \leq \gamma \left[1 + \frac{2\alpha}{\mathcal{J}_\Lambda(2\alpha - \Lambda)} \right] e^{-\Lambda T_k} \rightarrow 0,$$

as $k \rightarrow +\infty$. □

The condition in Proposition 3.27 is not necessary, as the following example shows.

Example 3.28 (Caterpillar sequence: Consistency). *Using the notation of Example 3.8, let $\mathcal{T} = (\mathbb{T}_k)_k$ be a caterpillar sequence with $t_k = \omega(\log k)$. The MLE of μ is consistent on $\mathcal{T}$ by Corollary 3.5. However let $\Lambda > 0$. For $k = 1, 2, \dots$*

$$\inf_{\pi \in \Pi^k} \sum_{x \in \pi} e^{-\Lambda \delta_k(\rho, x)} \leq k e^{-\Lambda \omega(\log k)} \rightarrow 0,$$

as $k \rightarrow +\infty$, where we used that $T_k \geq t_k$. Therefore $\Lambda^b = \overline{\Lambda}^g = 0$.

3.3.3 Phase transition on the rate of convergence of the MLE

We provided necessary and sufficient conditions for the consistency of the MLE of μ . Moreover we showed that, under Assumption 1, the MLE is consistent for nested and growing sequences. Here we provide bounds on the rate of convergence of the MLE. In particular we give conditions for $\sqrt{n_k}$ -consistency. We show that the latter undergoes a phase transition, generalizing Example 3.13.

Loss of $\sqrt{n_k}$ -consistency When the upper growth is above 2α , we show that the MLE of μ cannot be $\sqrt{n_k}$ -consistent. If further the branching number is above 2α , we give tight bounds on the convergence rate of the MLE. Roughly we show that, in the latter case, the variance behaves like $n_k^{2\alpha/\Lambda^g}$. Or perhaps a more accurate way to put it is that the “effective number of samples” n_k^{eff} is $e^{2\alpha T_k}$, in the sense that $\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] = \Theta((n_k^{\text{eff}})^{-1})$.

Theorem 3.2 (Convergence rate of $\hat{\mu}_{\text{ML}}$: Supercritical regime). *Let $(\mathbb{T}_k)_k$ be a tree sequence. If $\overline{\Lambda}^g > 2\alpha$, then for all $\epsilon > 0$ there is a subsequence $(k_j)_j$ along which*

$$\text{Var}_{\mathbb{T}_{k_j}} [\hat{\mu}_{\text{ML}}^{(k_j)}] \geq \gamma n_{k_j}^{-2\alpha/(\overline{\Lambda}^g - \epsilon)}. \quad (3.11)$$

In particular $(\hat{\mu}_{\text{ML}}^{(k)})_k$ is not $\sqrt{n_k}$ -consistent. If, further,

1. $\Lambda^b > 2\alpha$: then

$$\text{Var}_{\mathbb{T}_k} [\hat{\mu}_{\text{ML}}^{(k)}] = \Theta(e^{-2\alpha T_k}).$$

Moreover in terms of n_k , for all $\epsilon > 0$, there are constants $0 < C', C < +\infty$ such that

$$C' n_k^{-2\alpha/(\underline{\Lambda}^g - \epsilon)} \leq \text{Var}_{\mathbb{T}_k} [\hat{\mu}_{\text{ML}}^{(k)}] \leq C n_k^{-2\alpha/(\overline{\Lambda}^g + \epsilon)}, \quad (3.12)$$

and, in addition to (3.11),

$$\exists \text{ subsequence } (k'_j)_j, \text{ s.t. } \text{Var}_{\mathbb{T}_{k'_j}} [\hat{\mu}_{\text{ML}}^{(k'_j)}] \leq \gamma n_{k'_j}^{-2\alpha/(\underline{\Lambda}^g + \epsilon)}. \quad (3.13)$$

2. $\Lambda^b < 2\alpha$: then, for all $\epsilon > 0$, there are constants $0 < C', C < +\infty$ such that

$$C' n_k^{-2\alpha/(\underline{\Lambda}^g - \epsilon)} \leq \text{Var}_{\mathbb{T}_k} [\hat{\mu}_{\text{ML}}^{(k)}] \leq C n_k^{-(\Lambda^b - \epsilon)/(\overline{\Lambda}^g + \epsilon)}, \quad (3.14)$$

where the lower bound in (3.14) holds provided $\underline{\Lambda}^g > 0$, and

$$\exists \text{ subsequence } (k'_j)_j, \text{ s.t. } \text{Var}_{\mathbb{T}_{k'_j}} [\hat{\mu}_{\text{ML}}^{(k'_j)}] \leq \gamma n_{k'_j}^{-(\Lambda^b - \epsilon)/(\underline{\Lambda}^g + \epsilon)}. \quad (3.15)$$

Proof. Assume $\overline{\Lambda}^g > 2\alpha$. As remarked after Definition 3.11, for all $\epsilon > 0$, eventually

$$\exp((\underline{\Lambda}^g - \epsilon)T_k) \leq n_k \leq \exp((\overline{\Lambda}^g + \epsilon)T_k), \quad (3.16)$$

that is,

$$n_k^{-2\alpha/(\underline{\Lambda}^g - \epsilon)} \leq e^{-2\alpha T_k} \leq n_k^{-2\alpha/(\overline{\Lambda}^g + \epsilon)}. \quad (3.17)$$

Moreover for all $\epsilon > 0$ there are subsequences $(k_j)_j$ and $(k'_j)_j$ such that

$$n_{k_j} \geq \exp\left((\bar{\Lambda}^g - \epsilon)T_{k_j}\right) \text{ and } n_{k'_j} \leq \exp\left((\underline{\Lambda}^g + \epsilon)T_{k'_j}\right). \quad (3.18)$$

By Proposition 3.23,

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] \geq \gamma \left[e^{-2\alpha T_k} + \frac{1 - e^{-2\alpha T_k}}{n_k} \right] \geq \gamma e^{-2\alpha T_k}. \quad (3.19)$$

Then (3.11) follows from (3.18) and (3.19). Hence $(\hat{\mu}_{\text{ML}}^{(k)})_k$ is not $\sqrt{n_k}$ -consistent by Lemma 3.5.

Assume $\Lambda^b > 2\alpha$. Let $2\alpha < \Lambda < \Lambda^b$. By Proposition 3.20

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] \leq \gamma \left[1 + \frac{2\alpha}{J_\Lambda(\Lambda - 2\alpha)} \right] e^{-2\alpha T_k}. \quad (3.20)$$

Note that $\bar{\Lambda}^g \geq \underline{\Lambda}^g \geq \Lambda^b > 2\alpha$ and hence, by (3.19) and (3.20),

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] = \Theta(e^{-2\alpha T_k}).$$

Combining the last equation with (3.16) gives the result in terms of n_k .

Assume instead that $\Lambda^b < 2\alpha$. Let $\Lambda < \Lambda^b$. By Proposition 3.20

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] \leq \gamma \left[1 + \frac{2\alpha}{J_\Lambda(2\alpha - \Lambda)} \right] e^{-\Lambda T_k}. \quad (3.21)$$

The rest of the argument is similar to the previous case. $\square$

The following example shows that, when $\Lambda^b < 2\alpha$, our upper bound on the variance may not be achieved, but cannot be improved in general.

Example 3.29 (Two-level tree). Let $(\mathbb{T}_k)_k$ be a sequence of spherically symmetric trees, as defined in Example 3.17, with $H = 2$, $D_0^{(k)} = e^{\Lambda_0 \tau_0^{(0)}}$, $D_1^{(k)} = e^{\Lambda_1 \tau_1^{(k)}}$, for some

$\Lambda_0 < \Lambda_1$ and $\frac{\tau_0^{(k)}}{\tau_0^{(k)} + \tau_1^{(k)}} = \sigma$, with $0 < \sigma < 1$. Then, by (3.7),

$$\begin{aligned} \text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] &= \gamma e^{-2\alpha(\tau_0^{(k)} + \tau_1^{(k)})} \left[1 + \sum_{h=0,1} (1 - e^{-2\alpha\tau_h^{(k)}}) \prod_{h'=0}^h \frac{e^{2\alpha\tau_{h'}^{(k)}}}{D_{h'}^{(k)}} \right] \\ &= \gamma [e^{-2\alpha(\tau_0^{(k)} + \tau_1^{(k)})} + (1 - e^{-2\alpha\tau_0^{(k)}}) e^{-(\Lambda_0\tau_0^{(k)} + 2\alpha\tau_1^{(k)})} \\ &\quad + (1 - e^{-2\alpha\tau_1^{(k)}}) e^{-(\Lambda_0\tau_0^{(k)} + \Lambda_1\tau_1^{(k)})}]. \end{aligned} \quad (3.22)$$

Note that

$$\frac{\log n_k}{T_k} = \frac{\Lambda_0\tau_0^{(k)} + \Lambda_1\tau_1^{(k)}}{\tau_0^{(k)} + \tau_1^{(k)}} = \Lambda_0\sigma + \Lambda_1(1 - \sigma) = \Lambda^g.$$

To compute the branching number, it suffices to consider cutsets with m_0 middle vertices and the $D_1^{(k)}(D_0^{(k)} - m_0)$ tips below the rest of the middle vertices. Then

$$\mathcal{J}_k \equiv \inf_{\pi \in \Pi^k} \sum_{x \in \pi} e^{-\Lambda \delta_k(\rho, x)} = \begin{cases} D_0^{(k)} e^{-\Lambda \tau_0^{(k)}}, & \text{if } D_1^{(k)} > e^{\Lambda \tau_1^{(k)}} \\ n_k e^{-\Lambda T_k}, & \text{otherwise.} \end{cases}$$

Hence if $\Lambda \geq \Lambda_1 > \Lambda^g$ we are in the second case and

$$n_k e^{-\Lambda T_k} = e^{-(\Lambda - \Lambda^g)T_k} \rightarrow 0,$$

as $k \rightarrow +\infty$. If $\Lambda < \Lambda_1$ we are in the first case and

$$D_0^{(k)} e^{-\Lambda \tau_0^{(k)}} = e^{-(\Lambda - \Lambda_0)\tau_0^{(k)}},$$

so that $\Lambda^b = \Lambda_0$. If $\Lambda^b = \Lambda_0 \geq 2\alpha$, the dominant term in the variance is $\gamma e^{-2\alpha(\tau_0^{(k)} + \tau_1^{(k)})} = \gamma n_k^{-2\alpha/\Lambda^g}$, as predicted by Theorem 3.2. If instead $\Lambda^b = \Lambda_0 < 2\alpha$, there are two cases. If $\Lambda_1 \leq 2\alpha$, the dominant term in the variance is

$$\gamma e^{-(\Lambda_0\tau_0^{(k)} + \Lambda_1\tau_1^{(k)})} = \gamma e^{-\Lambda^g T_k} = \gamma n_k^{-1},$$

and we have $\sqrt{n_k}$ -consistency. If instead $\Lambda_1 > 2\alpha$, the dominant term in the variance is

$$\gamma e^{-(\Lambda_0 \tau_0^{(k)} + 2\alpha \tau_1^{(k)})} = \gamma e^{-(\Lambda^b \sigma + 2\alpha(1-\sigma))T_k} = \gamma n_k^{-(2\alpha/\Lambda^g)(1-\sigma) - (\Lambda^b/\Lambda^g)\sigma}.$$

Therefore, depending on the value of σ , we can get the full range of exponent values between (3.12) and (3.14). Note also that by taking Λ_1 large enough and σ close enough to 1 it is possible to have $\Lambda^g < 2\alpha$, yet not $\sqrt{n_k}$ -consistency. (Below, we will consider imposing the extra condition $\Lambda^b = \Lambda^g$. Then we must have $\Lambda_0 = \Lambda_1$ and this case cannot arise.)

Conditions for $\sqrt{n_k}$ -consistency The previous example also shows that when the upper growth is below 2α , the picture is somewhat murky. (In fact if $\bar{\Lambda}^g = 0$ we may not have consistency, as Example 3.26 shows.) The issue in Example 3.29 is the inhomogeneous growth rate. However, under extra regularity conditions, $\sqrt{n_k}$ -consistency can be established. We give examples in the next section.

Theorem 3.3 (Convergence rate of $\hat{\mu}_{ML}^{(k)}$: Subcritical regime). *Let $(\mathbb{T}_k)_k$ be a tree sequence with $\bar{\Lambda}^g < 2\alpha$. Then*

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{ML}^{(k)}] = O(n_k^{-1}).$$

Further if:

1. [Equality of growth and branching number] $\Lambda^b = \bar{\Lambda}^g > 0$ then for all $\epsilon > 0$

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{ML}^{(k)}] = O(n_k^{-(1-\epsilon)}).$$

2. [Bounded uniform growth] $\Lambda^{ug} < 2\alpha$ then

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{ML}^{(k)}] = O(n_k^{-1}).$$

Proof. One direction follows immediately from Proposition 3.23 which implies

$$\text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] \geq \gamma \frac{1 - e^{-2\alpha T_k}}{n_k} = \Omega(n_k^{-1}).$$

We prove the other direction separately in each case.

Assume first that $0 < \Lambda^b = \bar{\Lambda}^g < 2\alpha$. For $\epsilon > 0$ (small), choose Λ such that

$$\bar{\Lambda}^g - \epsilon < \Lambda < \bar{\Lambda}^g = \Lambda^b < 2\alpha.$$

By Proposition 3.20, eventually

$$\begin{aligned} \text{Var}_{\mathbb{T}_k}[\hat{\mu}_{\text{ML}}^{(k)}] &\leq \gamma \left[1 + \frac{2\alpha}{\mathcal{J}_\Lambda(2\alpha - \Lambda)} \right] e^{-\Lambda T_k} \\ &\leq \gamma \left[1 + \frac{2\alpha}{\mathcal{J}_\Lambda(2\alpha - \Lambda)} \right] n_k^{-(\bar{\Lambda}^g - \epsilon)/(\bar{\Lambda}^g + \epsilon)}. \end{aligned}$$

Assume instead that $\Lambda^{ug} < 2\alpha$. We show that $\bar{Y}$ (and hence the MLE by Proposition 3.18) achieves $\sqrt{n_k}$ -consistency in this case. Let θ be the corresponding flow on $\mathbb{T}_k$. By Lemma 3.19, letting $\Lambda^{ug} < \Lambda < 2\alpha$, for k large enough

$$\begin{aligned} \text{Var}_{\mathbb{T}_k}[\bar{Y}] &= \gamma e^{-2\alpha T_k} \left[1 + 2\alpha \int_0^{T_k} e^{2\alpha s} \left(\sum_{x \in \pi_s^k} \left(\frac{n_k(x)}{n_k} \right)^2 \right) ds \right] \\ &\leq \gamma e^{-2\alpha T_k} \left[1 + 2\alpha \int_0^{T_k} e^{2\alpha s} \left(\sum_{x \in \pi_s^k} \left(\frac{n_k(x)}{n_k} \right) \frac{e^{\Lambda[(T_k-s)+M]}}{n_k} \right) ds \right] \\ &\leq \gamma e^{-2\alpha T_k} \left[1 + e^{\Lambda M} \frac{2\alpha}{n_k(2\alpha - \Lambda)} e^{\Lambda T_k} (e^{(2\alpha - \Lambda)T_k} - 1) \right] \\ &= \gamma \left[e^{-2\alpha T_k} + e^{\Lambda M} \frac{2\alpha}{n_k(2\alpha - \Lambda)} (1 - e^{-(2\alpha - \Lambda)T_k}) \right]. \end{aligned}$$

The result follows from the fact that $e^{\Lambda[T_k+M]} \geq n_k$. □

3.3.4 Special cases

Bounded branch lengths. We first consider binary, ultrametric species trees with bounded edge lengths, where binary means the out-degree of every vertex is 2 except for the leaves.

Corollary 3.7 (Special case: Bounded edge lengths). *Let $(\mathbb{T}_k)_k$ be a sequence of binary, ultrametric trees with edge sets E_k and*

$$f = \inf_{k, b \in E_k} |b|, \quad g = \sup_{k, b \in E_k} |b|,$$

satisfying Assumption 1. Then the following hold.

1. *If $f > 0$, then $\hat{\mu}_{\text{ML}}^{(k)}$ is consistent.*
2. *If $f > \frac{\log 2}{2\alpha}$, then $\hat{\mu}_{\text{ML}}^{(k)}$ is $\sqrt{n_k}$ -consistent.*
3. *If $g < \frac{\log 2}{2\alpha}$, then $\hat{\mu}_{\text{ML}}^{(k)}$ is not $\sqrt{n_k}$ -consistent.*

Proof. By Theorem 3.1, to prove consistency, it suffices to show that

$$\liminf_k |\tilde{\pi}_s^k| = +\infty.$$

But, if $f > 0$, for a point x in $\mathbb{T}_k$

$$n_k(x) \leq 2^{T_k(x)/f},$$

so that

$$|\tilde{\pi}_s^k| \geq \frac{n_k}{2^{s/f}} \rightarrow +\infty.$$

If further $f \geq \frac{\log 2}{2\alpha(1-\epsilon)}$ for some $\epsilon > 0$,

$$n_k(x) \leq 2^{T_k(x)/f} \leq \exp\left(\frac{(\log 2)T_k(x)}{f}\right) \leq \exp(2\alpha(1-\epsilon)T_k(x))$$

so that

$$\Lambda^{\text{ug}} < 2\alpha,$$

and the second result follows from Theorem 3.3.

If instead $g \leq \frac{\log 2}{2\alpha(1+\epsilon)}$ for some $\epsilon > 0$,

$$n_k \geq 2^{T_k/g} \geq \exp\left(\frac{(\log 2)T_k}{g}\right) \geq \exp(2\alpha(1+\epsilon)T_k)$$

so that

$$\overline{\Lambda}^g > 2\alpha,$$

and the second result follows from Theorem 3.2. $\square$

Yule model We also specialize the results to the Yule model.

Corollary 3.8 (Special case: Yule model). *Let $(T_k)_k$ be a Yule sequence, as defined in Example 3.10, with rate $0 < \lambda < +\infty$. Then, with probability 1 (on the generation of T_0),*

1. $(\hat{\mu}_{\text{ML}}^{(k)})_k$ is consistent.
2. If $\lambda < 2\alpha$, $(\hat{\mu}_{\text{ML}}^{(k)})_k$ is $\sqrt{n_k}$ -consistent.
3. If $\lambda > 2\alpha$, $(\hat{\mu}_{\text{ML}}^{(k)})_k$ is not $\sqrt{n_k}$ -consistent and for all $\epsilon > 0$ there is $0 < C', C < +\infty$ such that

$$C'n_k^{-2\alpha\lambda^{-1}-\epsilon} \leq \text{Var}_{T_k}[\hat{\mu}_{\text{ML}}^{(k)}] \leq Cn_k^{-2\alpha\lambda^{-1}+\epsilon}.$$

Proof. By Theorems 3.1, 3.2, and 3.3, it suffices to prove that $\Lambda^b = \overline{\Lambda}^g = \lambda$ with probability 1.

A Galton-Watson (GW) branching process is a discrete-time non-negative integer-valued population process defined as follows: at each time step, each individual in the population has an independent number of offsprings, according to a distribution F , that form the population at the next time. In Peres (1999), it

is shown that a GW tree where F has mean m has branching number and upper growth equal to $\log m$.

To compute the branching number of an infinite Yule tree $\mathbb{T}_0$, we use a comparison to a GW tree. Fix $\epsilon > 0$. Let F be the distribution of the number of lineages in $\mathbb{T}_0$ at time ϵ . By standard branching process results ([Athreya and Ney, 2004](#)), $m = e^{\lambda\epsilon}$. By the memoryless property of the exponential, the number of lineages $|\pi_{N\epsilon}|$ in the Yule tree at time $N\epsilon$ is identically distributed to the population size Z_N of a GW tree with offspring distribution F at time N . Then

$$\frac{\log |\pi_s|}{s} \leq \frac{\log Z_{\lceil s/\epsilon \rceil}}{s} = \frac{\lceil s/\epsilon \rceil}{s} \cdot \frac{\log Z_{\lceil s/\epsilon \rceil}}{\lceil s/\epsilon \rceil},$$

which implies that

$$\overline{\Lambda}^g \leq \frac{1}{\epsilon} \cdot \log e^{\lambda\epsilon} = \lambda.$$

Similarly, let π be a cutset in $\mathbb{T}_0$ and let π_ϵ be the cutset obtained by rounding up the points in π to the next ϵ -multiple closer to the root (removing duplicates). Let $\delta_{\text{GW}}(v)$ be the distance from the root to vertex v in the GW tree. Then

$$\sum_{x \in \pi} e^{-\Lambda \delta_0(\rho, x)} \geq \sum_{y \in \pi_\epsilon} e^{-\Lambda(\delta_{\text{GW}}(y)+1)\epsilon} = e^{-\Lambda\epsilon} \sum_{y \in \pi_\epsilon} e^{-(\epsilon\Lambda)\delta_{\text{GW}}(y)} > 0$$

whenever $\epsilon\Lambda < \log e^{\lambda\epsilon}$, so that

$$\Lambda^b \geq \lambda.$$

□

3.3.5 Sensitivity to estimate of α

So far in this section, we considered the MLE of μ *given* α . Here we look at the sensitivity of the MLE to estimation errors on α . In the next section, we provide conditions under which a $\sqrt{n_k}$ -consistent estimator of α exists. In particular, these conditions are unrelated to the growth or height of the species tree, and apply to the two special cases above. Moreover the estimator of α we derive does

not require the knowledge of μ .

Hence suppose that we have a $\sqrt{n_k}$ -consistent estimator $\hat{\alpha}_k$ of α . Let $\widehat{\text{Var}}_{T_k}$ denote the variance under the parameter $\alpha = \hat{\alpha}_k$ (with μ and γ unchanged) and let $\hat{\theta}_k$ be the corresponding weights of the MLE of μ , that is, the choice of weights assuming that $\alpha = \hat{\alpha}_k$ and minimizing $\widehat{\text{Var}}_{T_k}[Y_{\theta}]$.

For all k and under the true α , $Y_{\hat{\theta}_k}$ is an unbiased estimator of μ . Moreover, because $\hat{\alpha}_k = \alpha + o(1)$ and so on, the bounds in Theorems 3.2 and 3.3 apply to $\widehat{\text{Var}}_{T_k}[Y_{\hat{\theta}_k}]$ as well (for k large enough). The quantity of interest is $\text{Var}_{T_k}[Y_{\hat{\theta}_k}]$. By Lemma 3.15,

$$\begin{aligned} \text{Var}_{T_k}[Y_{\hat{\theta}_k}] &= \gamma e^{-2\alpha T_k} + \gamma \sum_{b \in E_k} (1 - e^{-2\alpha|b|}) e^{2\alpha(\delta_k(\rho, b) - T_k)} (\hat{\theta}_k)_b^2 \\ &= (1 + O(T_k n_k^{-1/2})) \left[\gamma e^{-2\hat{\alpha}_k T_k} + \gamma \sum_{b \in E_k} (1 - e^{-2\hat{\alpha}_k|b|}) e^{2\hat{\alpha}_k(\delta_k(\rho, b) - T_k)} (\hat{\theta}_k)_b^2 \right] \\ &= (1 + O(T_k n_k^{-1/2})) \widehat{\text{Var}}_{T_k}[Y_{\hat{\theta}_k}], \end{aligned}$$

provided $T_k n_k^{-1/2} = o(1)$. Hence, for instance if $\underline{\Delta}^g > 0$, $T_k = O(\log n_k)$ and we get that $\text{Var}_{T_k}[Y_{\hat{\theta}_k}]$ satisfies the bounds in Theorems 3.2 and 3.3.

3.4 Convergence rate of a new estimator for α and γ

In this section, we provide a novel estimator for (α, γ) . Under natural assumptions on the species tree, we show that this estimator is $\sqrt{n_k}$ -consistent. Moreover this estimator does not require the knowledge of μ . Interestingly, in contrast to what we showed for μ , the conditions for $\sqrt{n_k}$ -consistency in this case do not involve the growth—or even the height—of the species tree. This is in line with the results in [Ho and Ané \(2013a\)](#), who found that μ requires an unbounded tree height to be microergodic, whereas α and γ do not.

Note, however, that the MLE of α and γ are not simple linear estimators, which makes them harder to study here. In particular, unlike in the case of μ ,

we do not provide lower bounds on their rate of convergence.

We illustrate our estimator in two special cases.

3.4.1 Contrast-based estimator

We first describe the estimator. The analysis of its convergence rate is performed in Section 3.4.2.

Contrasts Our estimator relies on an appropriately chosen set of contrasts, that is, differences between pairs of leaf states (see e.g. [Felsenstein, 2004](#)). More specifically, we choose contrasts associated with internal nodes, as follows. Let $\mathbb{T}$ be an ultrametric species tree with leaves $\mathcal{L}$ and internal vertices $\mathcal{I}$. For two leaves ℓ and ℓ' , we let $\ell \wedge \ell'$ be their most recent common ancestor. Assume that all internal vertices of $\mathbb{T}$ have out-degree at least 2. Let $i \in \mathcal{I}$ be an internal vertex of $\mathbb{T}$, and let $\ell_1^i \neq \ell_2^i$ be two leaves such that $\ell_1^i \wedge \ell_2^i = i$. Let P_i be the path connecting ℓ_1^i and ℓ_2^i . We define the corresponding contrast $\mathcal{C}_i = Y_{\ell_1^i} - Y_{\ell_2^i}$. Let $T(i)$ be the height of i from the leaves. We say that $T(i)$ is the height of $\mathcal{C}_i$.

Lemma 3.30 (Contrasts: Distribution ([Ho and Ané, 2013a](#))). *Let $i_1, \dots, i_m$ be a collection of internal nodes of $\mathbb{T}$. Let $\mathcal{C}_{i_1}, \dots, \mathcal{C}_{i_m}$ be an arbitrary set of associated contrasts. Assume that the corresponding paths $P_{i_1}, \dots, P_{i_m}$ are pairwise non-intersecting, that is, none of the pairs of paths share a vertex. Then $\mathcal{C}_{i_1}, \dots, \mathcal{C}_{i_m}$ are mutually independent, multivariate normal with $\mathcal{C}_i \sim \mathcal{N}(0, 2\gamma(1 - e^{-2\alpha T(i)}))$.*

Proof. Indeed, expanding the covariance, we get for $j \neq j'$

$$\gamma^{-1} \text{Cov}[\mathcal{C}_j, \mathcal{C}_{j'}] = e^{-\alpha d_{\ell_1^j \ell_1^{j'}}} - e^{-\alpha d_{\ell_1^j \ell_2^{j'}}} - e^{-\alpha d_{\ell_2^j \ell_1^{j'}}} + e^{-\alpha d_{\ell_2^j \ell_2^{j'}}} = 0,$$

since, by assumption, $\ell_\iota^j \wedge \ell_{\iota'}^{j'}$ is the same vertex for all $\iota, \iota' = 1, 2$. □

The following lemma will be useful in identifying an appropriate collection of contrasts.

Lemma 3.31 (Contrasts: A large collection (Ho and Ané, 2013a)). *Let $\mathbb{T}$ be an ultrametric tree and let $\mathcal{I}_{(a,b)}$ be the set of internal nodes of $\mathbb{T}$ whose height from the leaves lies in (a, b) . For every $a < b$, we can select a set of independent contrasts $\mathcal{C}$, associated with internal nodes in $\mathcal{I}_{(a,b)}$, such that*

$$|\mathcal{C}| \geq n(a, b)/2,$$

where $n(a, b) = |\mathcal{I}_{(a,b)}|$. In particular, the heights of the contrasts in $\mathcal{C}$ lie in (a, b) and their corresponding paths are pairwise non-intersecting.

Proof. Start with the lowest vertex i in $\mathcal{I}_{(a,b)}$ and choose a pair of vertices ℓ_1^i and ℓ_2^i such that $\ell_1^i \wedge \ell_2^i = i$. Remove i and its descendants as well as the edge immediately above i (and fuse consecutive edges separated by degree-2 vertices). As a result, the number of internal vertices in (a, b) decreases by at most 2. Repeat until no vertex is left in $\mathcal{I}_{(a,b)}$. $\square$

The estimator For a sequence of trees $\mathcal{T} = (\mathbb{T}_k)_k$, let $\mathcal{L}_k$ be the leaf set of $\mathbb{T}_k$; $\mathcal{I}_k$, the set of its internal vertices; $n_k = |\mathcal{L}_k|$ and $n_k(a, b) = |\mathcal{I}_k(a, b)|$; and $T_k(i)$, the height of i , for each $i \in \mathcal{I}_k$. The idea behind our estimator is to set up a system of equations that characterize α and γ uniquely. Our construction relies on the following condition. We illustrate this condition on two special cases below.

Assumption 2 (Linear-sized bands). *Assume that there are constants $0 < \beta < 1$ and $0 < c_1 < c'_1 < c_2 < c'_2 < \infty$ such that $n_k(c_\iota, c'_\iota) \geq \beta n_k$, $\iota = 1, 2$, for all k large enough.*

We set up our equations as follows. Let $m_k = \lfloor \beta n_k / 2 \rfloor$. Under Assumption 2, by Lemma 3.31, for each k we can choose *two* collections of independent contrasts $(\mathcal{C}_{i_r}^k)_{r=1}^{m_k}$ and $(\mathcal{C}_{j_r}^k)_{r=1}^{m_k}$ with corresponding heights $T_k(i_r) \in (c_1, c'_1)$ and $T_k(j_r) \in (c_2, c'_2)$ for every $r = 1, 2, \dots, m_k$. (Note that the two collections are

not independent.) For $r = 1, \dots, m$, let

$$\hat{a}_k = \frac{1}{m_k} \sum_{r=1}^{m_k} (\mathcal{C}_{i_r}^{(k)})^2, \quad \hat{b}_k = \frac{1}{m_k} \sum_{r=1}^{m_k} (\mathcal{C}_{j_r}^{(k)})^2,$$

and note that

$$\begin{aligned} a_k &\equiv \mathbb{E}[\hat{a}_k] = 2\gamma \left(1 - \frac{1}{m_k} \sum_{r=1}^{m_k} e^{-2\alpha T_k(i_r)} \right) \equiv 2\gamma h_k^1(\alpha), \\ b_k &\equiv \mathbb{E}[\hat{b}_k] = 2\gamma \left(1 - \frac{1}{m_k} \sum_{r=1}^{m_k} e^{-2\alpha T_k(j_r)} \right) \equiv 2\gamma h_k^2(\alpha). \end{aligned}$$

Notice that, under Assumption 2, $a_k \in [2\gamma(1 - e^{-2\alpha c_2}), 2\gamma(1 - e^{-2\alpha c_1})] \equiv [\underline{a}_\alpha, \bar{a}_\alpha]$ and $b_k \in [2\gamma(1 - e^{-2\alpha c_4}), 2\gamma(1 - e^{-2\alpha c_3})] \equiv [\underline{b}_\alpha, \bar{b}_\alpha]$. As shown below,

$$H_k(\alpha) = \frac{a_k}{b_k} = \frac{h_k^1(\alpha)}{h_k^2(\alpha)}$$

is invertible in α on $(0, +\infty)$. Hence a natural estimator of (α, γ) is obtained by setting

$$\hat{\alpha}_k = H_k^{-1} \left(\frac{\hat{a}_k}{\hat{b}_k} \right), \tag{3.23}$$

$$\hat{\gamma}_k = \frac{\hat{a}_k}{2h_k^1(\hat{\alpha}_k)}. \tag{3.24}$$

We will show in the proof of invertibility below that H_k is actually strictly increasing, and therefore relatively straightforward to invert numerically. It remains to prove invertibility.

Lemma 3.32 (Invertibility of the system). *Under Assumption 2, $H_k(\alpha)$ is strictly positive, differentiable, and invertible on $(0, +\infty)$.*

Proof. We have that

$$\begin{aligned}
\frac{\partial \log H_k(\alpha)}{\partial \alpha} &= \frac{\sum_{r=1}^{m_k} 2T_k(i_r) e^{-2\alpha T_k(i_r)}}{\sum_{r=1}^{m_k} (1 - e^{-2\alpha T_k(i_r)})} - \frac{\sum_{r=1}^{m_k} 2T_k(j_r) e^{-2\alpha T_k(j_r)}}{\sum_{r=1}^{m_k} (1 - e^{-2\alpha T_k(j_r)})} \\
&= \frac{\sum_{r=1}^{m_k} \sum_{r'=1}^{m_k} 2T_k(i_r) e^{-2\alpha T_k(i_r)} (1 - e^{-2\alpha T_k(j_{r'})})}{\sum_{r=1}^{m_k} (1 - e^{-2\alpha T_k(i_r)}) \sum_{r=1}^{m_k} (1 - e^{-2\alpha T_k(j_r)})} \\
&\quad - \frac{\sum_{r=1}^{m_k} \sum_{r'=1}^{m_k} 2T_k(j_{r'}) e^{-2\alpha T_k(j_{r'})} (1 - e^{-2\alpha T_k(i_r)})}{\sum_{r=1}^{m_k} (1 - e^{-2\alpha T_k(i_r)}) \sum_{r=1}^{m_k} (1 - e^{-2\alpha T_k(j_r)})} \quad (3.25)
\end{aligned}$$

Note that the function $\frac{x e^{-x}}{1 - e^{-x}}$ is strictly decreasing on $(0, \infty)$ because its derivative is $\frac{e^{-x}(1 - x - e^{-x})}{(1 - e^{-x})^2} < 0$ on $(0, +\infty)$. Therefore

$$\frac{2T_k(i_r) e^{-2\alpha T_k(i_r)}}{1 - e^{-2\alpha T_k(i_r)}} \geq \frac{2c'_1 e^{-2\alpha c'_1}}{1 - e^{-2\alpha c'_1}} > \frac{2c_2 e^{-2\alpha c_2}}{1 - e^{-2\alpha c_2}} \geq \frac{2T_k(j_{r'}) e^{-2\alpha T_k(j_{r'})}}{1 - e^{-2\alpha T_k(j_{r'})}},$$

that is,

$$\begin{aligned}
&2T_k(i_r) e^{-2\alpha T_k(i_r)} (1 - e^{-2\alpha T_k(j_{r'})}) \\
&\quad - 2T_k(j_{r'}) e^{-2\alpha T_k(j_{r'})} (1 - e^{-2\alpha T_k(i_r)}) > 0, \quad (3.26)
\end{aligned}$$

for every r, r' , so that each (r, r') -term in (3.25) is strictly positive. Hence, we can deduce that $\partial \log H_k(\alpha) / \partial \alpha > 0$, that is, $\log H_k$ (and hence H_k itself) is strictly increasing on $(0, +\infty)$ and continuous, and therefore invertible. $\square$

Note that we cannot use the law of large numbers to derive consistency (despite the independence of the contrasts) because a_k/b_k is a bounded, but not necessarily convergent, sequence and H_k^{-1} is continuous, but depends on k . Instead we argue directly about $\sqrt{n_k}$ -consistency below.

3.4.2 Rate of convergence

Our main result in this section is the $\sqrt{n_k}$ -consistency of our estimator for (α, γ) .

Theorem 3.4 (Estimating α and γ : $\sqrt{n_k}$ -consistency). *Let $\mathcal{T} = (\mathbb{T}_k)_k$ be a sequence*

of ultrametric trees satisfying Assumptions 1 and 2 and let $(\hat{\alpha}_k, \hat{\gamma}_k)_k$ be the estimator defined in (3.23) and (3.24). Then $|\hat{\alpha}_k - \alpha| = O_p(n_k^{-1/2})$ and $|\hat{\gamma}_k - \gamma| = O_p(n_k^{-1/2})$.

Proof. Note that $\mathbb{E}[\hat{a}_k] = a_k$ and

$$\text{Var}[\hat{a}_k] = \frac{8\gamma^2}{m_k^2} \sum_{r=1}^{m_k} (1 - e^{-2\alpha T_k(i_r)})^2 \leq \frac{8\gamma^2}{m_k} (1 - e^{-2\alpha c_1})^2 = O(m_k^{-1}) = O(n_k^{-1}),$$

where we used that $([2\gamma(1 - e^{-2\alpha T_k(i_r)})]^{-1/2} \mathcal{C}_{i_r}^k)^2$ is χ_1^2 -distributed and, therefore, has variance 2. Hence, by Lemma 3.4, $|\hat{a}_k - a_k| = O_p(n_k^{-1/2})$. Similarly, $|\hat{b}_k - b_k| = O_p(n_k^{-1/2})$. Our claim that $|\hat{\alpha}_k - \alpha_k| = O_p(n_k^{-1/2})$ then follows from the following straightforward lemmas.

Lemma 3.33. *Let x, y be two positive numbers such that $0 < x_* \leq x \leq x^* < \infty$ and $0 < y_* \leq y \leq y^* < \infty$. Assume that $|x - x'| \leq \epsilon$ and $|y - y'| \leq \epsilon$ with $\epsilon < y_*/2$. Then*

$$\left| \frac{x'}{y'} - \frac{x}{y} \right| < \frac{4(x^* + y^*)}{y_*^2} \epsilon.$$

Proof. We have

$$\left| \frac{x'}{y'} - \frac{x}{y} \right| = \left| \frac{y(x' - x) + x(y - y')}{yy'} \right| \leq \frac{y^*|x' - x|}{y_*(y_*/2)} + \frac{x^*|y - y'|}{y_*(y_*/2)}.$$

□

Lemma 3.34. *If $0 < z_* \leq z \leq z^* < \infty$, $|z' - z| \leq \epsilon$ and $\epsilon < z_*/2$, then there is a constant $\Delta(z_*, z^*)$ depending on c_1, c'_1, c_2, c'_2 such that for all k*

$$\sup_{t \in [0,1]} |(H_k^{-1})'(tz' + (1-t)z)| \leq \Delta(z_*, z^*).$$

Proof. We use the proof of Lemma 3.32. Let $\zeta_\alpha = \zeta_\alpha(c_1, c'_1, c_2, c'_2) > 0$ be the smallest possible difference in (3.26) for a fixed α . Let α_*, α^* be defined as

$$\frac{1}{2}z_* = \frac{\bar{a}_{\alpha_*}}{\underline{b}_{\alpha_*}}, \quad \frac{3}{2}z^* = \frac{\underline{a}_{\alpha^*}}{\bar{b}_{\alpha^*}}.$$

Then $[\alpha_*, \alpha^*] \supseteq H_k^{-1}([\frac{1}{2}z_*, \frac{3}{2}z^*])$ for all k . Note that

$$\begin{aligned} \sup_{t \in [0,1]} |(H_k^{-1})'(tz' + (1-t)z)| &\leq \sup_{z \in [\frac{1}{2}z_*, \frac{3}{2}z^*]} |(H_k^{-1})'(z)| \\ &= \sup_{z \in [\frac{1}{2}z_*, \frac{3}{2}z^*]} \left(\frac{\partial H_k}{\partial \alpha}(H_k^{-1}(z)) \right)^{-1} \\ &= \sup_{z \in [\frac{1}{2}z_*, \frac{3}{2}z^*]} \left(\left[H_k \frac{\partial \log H_k}{\partial \alpha} \right] (H_k^{-1}(z)) \right)^{-1} \\ &\leq \sup_{\alpha \in [\alpha_*, \alpha^*]} \frac{\underline{b}_\alpha}{\bar{a}_\alpha} \cdot \frac{(1 - e^{-2\alpha c'_1})(1 - e^{-2\alpha c'_2})}{\zeta_\alpha} \\ &\equiv \Delta(z_*, z^*). \end{aligned}$$

□

We finish the proof of Theorem 3.4. Fix $\delta > 0$ (small) and pick M_δ such that $\mathbb{P}[\hat{a}_k - a_k \geq M_\delta n_k^{-1/2}] < \delta/2$ and similarly for $\hat{b}_k$. Then, by Lemma 3.33 and Assumption 1, for k large enough

$$\begin{aligned} &\mathbb{P} \left[\left| \frac{\hat{a}_k}{\hat{b}_k} - \frac{a_k}{b_k} \right| \geq \frac{4(\bar{a}_\alpha + \bar{b}_\alpha)}{\underline{b}_\alpha^2} M_\delta n_k^{-1/2} \right] \\ &\leq \mathbb{P} \left[\left| \frac{\hat{a}_k}{\hat{b}_k} - \frac{a_k}{b_k} \right| \geq \frac{4(\bar{a}_\alpha + \bar{b}_\alpha)}{\underline{b}_\alpha^2} M_\delta n_k^{-1/2}, |\hat{a}_k - a_k| \leq M_\delta n_k^{-1/2}, |\hat{b}_k - b_k| \leq M_\delta n_k^{-1/2} \right] \\ &\quad + \mathbb{P} [|\hat{a}_k - a_k| \geq M_\delta n_k^{-1/2}] + \mathbb{P} [|\hat{b}_k - b_k| \geq M_\delta n_k^{-1/2}] \\ &\leq 0 + \frac{\delta}{2} + \frac{\delta}{2} = \delta, \end{aligned}$$

so that

$$\left| \frac{\hat{a}_k}{\hat{b}_k} - \frac{a_k}{b_k} \right| = O_p(n_k^{-1/2}).$$

Secondly, using Rolle's theorem, we have

$$|\hat{\alpha}_k - \alpha| \leq \sup_{t \in [0,1]} \left| (H_k^{-1})' \left(t \frac{\hat{a}_k}{\hat{b}_k} + (1-t) \frac{a_k}{b_k} \right) \right| \cdot \left| \frac{\hat{a}_k}{\hat{b}_k} - \frac{a_k}{b_k} \right|.$$

Let M_δ be such that

$$\mathbb{P} \left[\left| \frac{\hat{a}_k}{\hat{b}_k} - \frac{a_k}{b_k} \right| \geq M_\delta n_k^{-1/2} \right] < \delta.$$

Fix $\epsilon' > 0$ and let

$$z_* = \frac{a_{\alpha-\epsilon'}}{\bar{b}_{\alpha-\epsilon'}}, \quad z^* = \frac{\bar{a}_{\alpha+\epsilon'}}{\underline{b}_{\alpha+\epsilon'}}.$$

Then, by Lemma 3.34, letting

$$\mathcal{H}_k = \left\{ \sup_{t \in [0,1]} \left| (H_k^{-1})' \left(t \frac{\hat{a}_k}{\hat{b}_k} + (1-t) \frac{a_k}{b_k} \right) \right| \cdot \left| \frac{\hat{a}_k}{\hat{b}_k} - \frac{a_k}{b_k} \right| \geq \Delta^{-1}(z_*, z^*) M_\delta n_k^{-1/2} \right\},$$

we have for k large enough

$$\begin{aligned} & \mathbb{P} \left[|\hat{\alpha}_k - \alpha| \geq \Delta^{-1}(z_*, z^*) M_\delta n_k^{-1/2} \right] \\ & \leq \mathbb{P}[\mathcal{H}_k] \\ & \leq \mathbb{P} \left[\mathcal{H}_k, \left| \frac{\hat{a}_k}{\hat{b}_k} - \frac{a_k}{b_k} \right| < M_\delta n_k^{-1/2} \right] + \mathbb{P} \left[\left| \frac{\hat{a}_k}{\hat{b}_k} - \frac{a_k}{b_k} \right| \geq M_\delta n_k^{-1/2} \right] \\ & \leq 0 + \delta = \delta. \end{aligned}$$

That implies

$$|\hat{\alpha}_k - \alpha| = O_p(n_k^{-1/2}).$$

The argument for $\hat{\gamma}_k$ is similar. To deal with the denominator, note that

$$\mathbb{P} \left[|2h_k^1(\hat{\alpha}_k) - 2h_k^1(\alpha)| \geq M_\delta n_k^{-1/2} \right]$$

$$\begin{aligned}
&\leq \mathbb{P} \left[2 \left| \frac{1}{m_k} \sum_{r=1}^{m_k} (e^{-2\hat{\alpha}_k T_k(i_r)} - e^{-2\alpha T_k(i_r)}) \right| \geq M_\delta n_k^{-1/2} \right] \\
&\leq \mathbb{P} \left[2 \frac{1}{m_k} \sum_{r=1}^{m_k} |e^{-2\hat{\alpha}_k T_k(i_r)} - e^{-2\alpha T_k(i_r)}| \geq M_\delta n_k^{-1/2} \right] \\
&\leq \mathbb{P} \left[2 \frac{1}{m_k} \sum_{r=1}^{m_k} 2T_k(i_r) |\hat{\alpha}_k - \alpha| \geq M_\delta n_k^{-1/2} \right] \\
&\leq \mathbb{P} \left[4c_1 |\hat{\alpha}_k - \alpha| \geq M_\delta n_k^{-1/2} \right],
\end{aligned}$$

where we used that $|e^{-x} - e^{-y}| \leq |x - y|$ for $x, y \geq 0$. Then use Lemma 3.33 as above. $\square$

Remark 3.35. *The previous result can be further generalized to sublinear-sized collections of contrasts whose height grows with k (with a different rate of convergence). We leave out the details.*

3.4.3 Special cases

We show that Assumption 2 is satisfied in two natural settings.

Bounded edge lengths Let $(\mathbb{T}_k)_k$ be a sequence of binary, ultrametric trees with edge sets E_k and

$$f = \inf_{k, b \in E_k} |b|, \quad g = \sup_{k, b \in E_k} |b|,$$

satisfying Assumption 1.

Corollary 3.9 (Special case: Bounded edge lengths). *If $0 < f \leq g < +\infty$ then Assumption 2 is satisfied, and hence $|\hat{\alpha}_k - \alpha| = O_p(n_k^{-1/2})$ and $|\hat{\gamma}_k - \gamma| = O_p(n_k^{-1/2})$.*

Proof. By Theorem 3.4 it suffices to show that Assumption 2 is satisfied. We claim that

$$n_k(f, g) \geq n_k/2 \tag{3.27}$$

and

$$n_k(2g, 3g) \geq n_k/2^{2g/f}$$

so that we can take $c_1 = f$, $c'_1 = g$, $c_2 = 2g$, and $c'_2 = 3g$. Indeed, looking backwards in time, by the definition of f and g the n_k lineages originating from the leaves cannot coalesce in the $(0, f)$ time interval, but must coalesce at least once in the (f, g) time interval. Then (3.27) follows from our assumption that $\mathbb{T}_k$ is binary. Similarly the largest possible number of descendant leaves of a point at height $2g$ is $2^{2g/f}$. Hence the number of lineages at time $2g$ is at least $n_k/2^{2g/f}$ and all such lineages must coalesce at least once in the next g -length time interval, looking backwards. $\square$

Yule model We also apply the results to the Yule model. For simplicity, we take a special sequence of times (although this assumption is not crucial). Let $\mathbb{T}_0$ be an infinite Yule tree with rate $0 < \lambda < +\infty$. For $k \geq 1$, let t_k be the first time at which $\mathbb{T}_0$ has $k + 1$ lineages. Then $n_k = k$ for all k and $t_k \rightarrow +\infty$ so that Assumption 1 is satisfied.

Corollary 3.10 (Special case: Yule model). *Let $(\mathbb{T}_k)_k$ be a Yule sequence as above. Then Assumption 2 is satisfied asymptotically, and hence $|\hat{\alpha}_k - \alpha| = O_p(n_k^{-1/2})$ and $|\hat{\gamma}_k - \gamma| = O_p(n_k^{-1/2})$.*

Remark 3.36. *Note that, unlike the case of Corollary 3.8, we do not prove that the result holds with probability 1 on the choice of $\mathbb{T}_0$. That is, the convergence in probability above involves stochasticity over both $\mathbb{T}_0$ and the overlaid OU process.*

Proof. Let $\tau_i = t_i - t_{i-1}$ be the amount of time during which $\mathbb{T}_0$ has i lineages (with $t_0 = 0$). Then $(\tau_i)_i$ are independent exponential random variables with parameters $(1/(i\lambda))_i$. Let $T_i^j = \sum_{r=i+1}^j \tau_r$. Note that

$$\mathbb{E}[T_i^j] = \sum_{r=i+1}^j \mathbb{E}[\tau_r] \equiv \lambda^{-1} \sum_{r=i+1}^j \frac{1}{r},$$

with

$$\log \left(\frac{j}{i+1} \right) = \int_{i+1}^j \frac{1}{x} dx \leq \sum_{r=i+1}^j \frac{1}{r} \leq \int_i^j \frac{1}{x} dx \leq \log \left(\frac{j}{i} \right).$$

Similarly,

$$\text{Var}[T_i^j] = \sum_{r=i+1}^j \text{Var}[\tau_r] = \lambda^{-2} \sum_{r=i+1}^j \frac{1}{r^2}, \quad (3.28)$$

with

$$\sum_{r=i+1}^j \frac{1}{r^2} \leq \int_i^{+\infty} \frac{1}{x^2} dx = \frac{1}{i}.$$

By Chebyshev's inequality, for all $0 < \sigma < 1$,

$$\mathbb{P} \left[|T_{\lfloor \sigma k \rfloor}^k - \mathbb{E} [T_{\lfloor \sigma k \rfloor}^k] | \geq \epsilon \right] = O(k^{-1}),$$

where we used (3.28). Let $0 < \sigma'_2 < \sigma_2 < \sigma'_1 < \sigma_1 < 1$. From the previous equation, we get for $\iota = 1, 2$

$$\mathbb{P} \left[T_{\lfloor \sigma_\iota k \rfloor}^k \leq \lambda^{-1} \log \left(\frac{k}{\lfloor \sigma_\iota k \rfloor + 1} \right) - \epsilon \right] = O(k^{-1}),$$

and

$$\mathbb{P} \left[T_{\lfloor \sigma'_\iota k \rfloor}^k \geq \lambda^{-1} \log \left(\frac{k}{\lfloor \sigma'_\iota k \rfloor} \right) + \epsilon \right] = O(k^{-1}).$$

Take

$$\alpha_\iota = \lambda^{-1} \log \left(\frac{1}{\sigma_\iota} \right), \quad \alpha'_\iota = \lambda^{-1} \log \left(\frac{1}{\sigma'_\iota} \right) \quad \text{and} \quad \epsilon < \alpha_1 \wedge \frac{1}{2} [\alpha_2 - \alpha'_1].$$

Then Assumption 2 is satisfied asymptotically with

$$c_\iota = \alpha_\iota - \epsilon, \quad c'_\iota = \alpha'_\iota + \epsilon \quad \text{and} \quad \beta = [\sigma_1 - \sigma'_1] \wedge [\sigma_2 - \sigma'_2],$$

because then

$$\mathbb{P}\left[\mathbf{c}_1 < T_{[\sigma_1 k]}^k < T_{[\sigma'_1 k]}^k < \mathbf{c}'_1 < \mathbf{c}_2 < T_{[\sigma_2 k]}^k < T_{[\sigma'_2 k]}^k < \mathbf{c}'_2\right] \geq 1 - O(k^{-1}).$$

□

4 PRACTICAL INFERENCE DIFFICULTIES AND CHALLENGES WITH MODEL SELECTION

Abstract For the study of macroevolution, phenotypic data are analyzed across species on a dated phylogeny using phylogenetic comparative methods. In this context, the Ornstein-Uhlenbeck (OU) process is now being used extensively to model selectively-driven trait evolution, whereby a trait is attracted to a selection optimum μ with selection pressure α . We report here theoretical properties of the maximum likelihood (ML) estimators for these parameters, including their non-uniqueness and inaccuracy, and show that theoretical expectations indeed apply to real trees. We provide necessary conditions for ML estimators to be well-defined and practical implications for model parametrization. We then show how these limitations carry over to difficulties in detecting shifts in selection regimes along a phylogeny. When the phylogenetic placement of these shifts is unknown, we identify a “large p - small n ” problem where traditional model selection criteria fail and favor overly complex scenarios. Instead, we propose a modified criterion that is better adapted to change-point models. The challenges we identify here are inherent to trait evolution models on phylogenetic trees when observations are limited to present-day taxa, and require the addition of fossil taxa to be alleviated.

Funding This work was supported in part by the National Science Foundation (DMS 1106483).

4.1 Introduction

Analysis of trait evolution is at the heart of evolutionary biology. Much effort on reconstructing phylogenies is motivated by the subsequent analysis of traits, to study how traits correlate to each other, to ecological covariates or to study how rates of trait evolution vary across clades ([Pennell and Harmon, 2013](#)).

The present work focuses on the statistical power that phylogenies provide for studying continuous trait evolution, when traits evolve under natural selection as modeled by an Ornstein-Uhlenbeck (OU) process ([Hansen, 1997](#); [Hunt et al., 2008](#)). Under the OU process, the trait undergoes random drift with variance accumulation rate σ^2 , but it is also attracted to a “selection optimum” μ with selection strength α . In the particular case when $\alpha = 0$, the action of natural selection disappears and the process reduces to random drift only, as modeled by a Brownian motion (BM). The OU model has been used to detect the presence of natural selection, as opposed to neutral evolution. For instance, [Scales et al. \(2009\)](#) showed evidence for natural selection acting on the fiber-type composition of locomotor muscles of lizards, with a common selective pressure across their clade. [Harmon et al. \(2010\)](#) showed broad support for the OU model of natural selection across many clades. The OU model has also been used to account for a partial but unknown degree of phylogenetic effect when testing for the correlation between traits. Indeed, the presence of natural selection is expected to reduce the level of phylogenetic correlation, and the selection strength parameter α can be used to adaptively estimate the level of phylogenetic correlation when testing the relationships between various traits (e.g. [Smith et al. \(2008\)](#); [Lavin et al. \(2008\)](#)), see also [Freckleton and Jetz \(2009\)](#) for separating the phylogenetic correlation from the spatial correlation, and [Ives and Garland \(2010\)](#) for an analogue of the OU correlation model on binary data).

The OU process is also ideal to model changes in selection regimes, such as changes in the selection optimum μ between different parts of the tree (as pioneered by [Butler and King \(2004\)](#), see also [Scales et al. \(2009\)](#) for instance), or changes in the selection strength α or the variance rate σ^2 ([Beaulieu et al., 2012](#)). Changes in the selection optimum μ have been modeled elegantly by [Hansen et al. \(2008\)](#); [Bartoszek et al. \(2012\)](#), under the assumption that μ is driven by a continuous predictor whose evolution is modeled with a BM or OU process.

While maximum likelihood (ML) is widely used with OU models, very few studies have focused on the statistical properties of these methods on trees. Some recent work showed that statistical behaviors expected under standard models

with independent units should not be taken for granted. A first basic, expected property of statistical inference is the convergence of parameter estimates to the true parameter values as more and more data are collected. This property was shown to break down on phylogenies for some parameters (Ané, 2008). Other properties were also investigated, such as the expected increase of power from larger and larger trees (Ané, 2008; Ho and Ané, 2013a), and the (un)expected properties of standard model selection tools such as AIC or BIC (Ané, 2008; Boettiger et al., 2012). These limitations are reviewed in Section 4.2. The rest of this paper focuses on statistical issues with the OU model for trait evolution. Section 4.3 discusses the necessary conditions for ML estimators to be unique and practical implications for model parametrization. In Section 4.4, we use simulations to illustrate the conditions when ML fails to accurately estimate the OU model parameters. Pitfalls of studying changes in selection regimes along a phylogeny when searching among a very large number of models are discussed in Section 4.5. Our simulations show that the power to detect phylogenetic changes in the selection optimum depends little on the tree size. Moreover, when the phylogenetic position of the changes are unknown, AIC and BIC fail to select the correct model in favor of overly complex models with many changes. We introduce here a phylogenetic adaptation of the modified BIC suggested by Zhang and Siegmund (2007). We show that this new criterion avoids overfitting and performs much better than both AIC and BIC.

In what follows, we consider OU models with a single selection strength α and a single variance accumulation (drift) parameter σ^2 over the whole tree. We also restrict our attention to the case of a known phylogeny with known branch lengths, to focus on the challenges brought about by the trait evolution model.

4.2 Limitations due to phylogenetic dependence

Limitations in inferring ancestral states have been reported very early in empirical studies because large standard errors were observed (e.g. Schluter et al., 1997; Garland and Ives, 2000). Confidence intervals for ancestral states can be so wide

as to span the range of observed present-day values, so much so that “[...] it might be best to accept our limitations and not even try to estimate ancestral states from comparative data.” (Martins, 2000). These limitations due to phylogenetic dependence have recently been explained theoretically. Ané (2008) showed that under BM and some general conditions, the ancestral state is not estimated consistently as the tree grows indefinitely. More specifically, the variance of the maximum likelihood estimator (MLE) of the ancestral state cannot be lower than $\sigma^2 t/k$ where k is the number of daughters of the root node and t is the length of the shortest branch stemming from the root. This result proves that the ancestral state reconstruction accuracy is very limited, unless evolution proceeded slowly or the ancestral node of interest consists of a large polytomy, or unless fossil taxa can be included in the analysis. The accuracy loss due to phylogenetic dependence can be substantial. Using the phylogeny from Bininda-Emonds et al. (2007) with 4507 species for instance (fig. 4.4, left) and under BM evolution, the reconstruction at the base of the mammal tree has the same accuracy as that obtained from only 5 independent taxa, if such existed, from a star tree of the same height. With marsupials excluded (4249 species), trait reconstruction for the non-marsupial mammal ancestor has the same accuracy as would be obtained from about 19 independent species. Interestingly, this number of equivalent independent observations was shown to be lower than the adjusted degree of freedom defined by Paradis and Claude (2002) as the tree length to tree height ratio, when the tree is ultrametric.

Under the OU model, Ho and Ané (2013a) proved a similar issue with the ancestral state and the selection optimum μ . If the tree is ultrametric and the height of the tree is bounded, μ cannot be estimated consistently as the tree grows indefinitely, regardless of the estimation method. They also showed theoretically that the presence of fossil taxa, making the tree non ultrametric, is crucial to increase precision in μ .

While contemporary species bear limited information on ancestral states and optimal values, rates of evolution and correlation among traits are typically much easier to estimate. Under BM evolution, independent contrasts (Felsenstein, 1985)

have been used to detect significant correlations in many studies, too numerous to count. Many studies also successfully detected shifts in the rate of trait evolution σ^2 (e.g. O'Meara et al., 2006; Davis et al., 2007; Eastman et al., 2011; Venditti et al., 2011). Indeed, the precision in estimated rates and regression coefficients is known to increase with the square-root of the number of taxa, unlike ancestral state estimates (Ané, 2008).

Inference on the level of phylogenetic correlation from contemporary taxa is also challenging. One common way to measure phylogenetic correlation is through a λ parameter (Freckleton et al., 2002), where $\lambda = 0$ corresponds to no phylogenetic correlation and $\lambda = 1$ corresponds to the BM model. Using simulations, Boettiger et al. (2012) showed clearly that λ estimates can be very uncertain, and typically have a lot less precision than estimates of the variance rate σ^2 .

Another common way to measure phylogenetic correlation is through the OU model, with $\alpha = \infty$ corresponding to no phylogenetic correlation and $\alpha = 0$ to the BM model. Here again, Ho and Ané (2013a) showed situations when α is not consistently estimable (while the variance rate σ^2 might be), and Boettiger et al. (2012) demonstrated with simulations that the power to detect the presence of selection pressure ($\alpha > 0$) can be disappointingly low even on very large trees.

Traditional methods for model selection are also challenged by phylogenetic dependence. The standard BIC was shown to be inappropriate under BM evolution (Ané, 2008). Its tendency to select overly simple models was explained by its penalty increasing with the number of present-day species instead of a number of equivalent independent observations. Both AIC and BIC were shown in simulations to result in a strong bias for over-parametrized models when applied to OU models with multiple selection regimes (Boettiger et al., 2012), casting doubts on the appropriateness of these criteria for phylogenetic comparative methods. Given these limitations when selection parameters are constant along the whole tree, precision and model selection issues are expected to be even more pronounced under complex OU models with multiple selection regimes in the tree (Butler and King, 2004; Hansen et al., 2008; Beaulieu et al., 2012; Bartoszek

et al., 2012).

4.3 (Un)identifiability of selection regimes

A desired property of any model is that the likelihood reaches its maximum at a unique point over the range of parameter values. Unfortunately, we show that it is not always the case for OU models on trees, when the maximum likelihood is reached on a ridge making the MLE undefined and some parameters unidentifiable. Consider first the simple OU model on an ultrametric tree with a single selection optimum μ . Given trait y_0 at the root, the observation Y_i for individual i is normally distributed with mean $y_0 e^{-\alpha T} + \mu(1 - e^{-\alpha T})$, where T is the height of the tree and α the selection strength. We can write this as

$$Y = X\beta + e \quad (4.1)$$

where $\beta = (y_0, \mu)$, the residual variation e has covariance derived from the OU process and X is a design matrix with 2 columns, $e^{-\alpha T}\mathbf{1}$ and $(1 - e^{-\alpha T})\mathbf{1}$, where $\mathbf{1}$ is a vector of ones. In general, X needs to be of full rank for the MLE to be unique and for all parameters in β to be identifiable. In our example, the two columns of X are parallel (each one is a rescaled version of the other), so X is not of full rank and y_0 and μ cannot be identified apart from each other. This is true regardless of α . Surprisingly, this lack of identifiability has gone unnoticed for a long time, even though it can occur frequently with OU tree models. One consequence is that the MLE of β is not unique, regardless of the size of the tree. To illustrate the problem we simulated data under an OU process on the 4507-species ultrametric mammal tree from [Bininda-Emonds et al. \(2007\)](#) with ancestral state $y_0 = 3$, selection optimum $\mu = 2$, selection strength $\alpha = 1$ and variance rate $\sigma^2 = 2$. The tree was rescaled to have height $T = 1$ so that $\alpha = 1$ corresponds to an intermediate level of phylogenetic correlation. For our simulated data, the MLEs were $\hat{\sigma}^2 = 2.07$ and $\hat{\alpha} = 0.805$. However, every point (y_0, μ) lying on the line $y_0 e^{-\hat{\alpha} T} + \mu(1 - e^{-\hat{\alpha} T}) = 1.998$ maximized the

likelihood (fig. 4.1). This line formed a ridge of the likelihood surface. Hence, any good optimization procedure should report some convergence issue and any one reported MLE value on this line is obviously not necessarily a good estimate for the true value of (y_0, μ) .

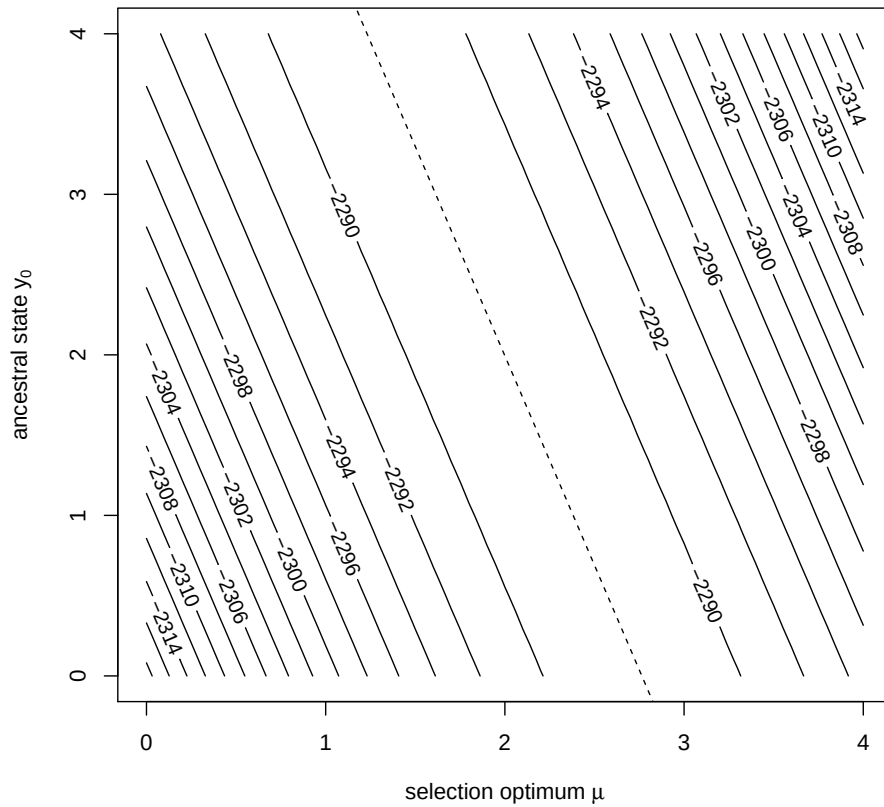


Figure 4.1: Log likelihood surface with respect to (y_0, μ) from data simulated on a 4507-taxon tree, with other parameters fixed to their ML values. The dash line $y_0 e^{-\hat{\alpha}T} + \mu(1 - e^{-\hat{\alpha}T}) = 1.998$ is where the likelihood achieves its maximum.

4.3.1 Identifiability of selection optima

More generally, non-identifiability occurs whenever $\mathbf{X}$ is not of full rank, in which case the likelihood surface has a hyperplane ridge and the MLE is undefined. The unidentifiability of the ancestral state and selection optima can occur in models where μ takes different values $(\mu_1, \mu_2, \dots, \mu_m)$ along different parts of an ultrametric tree, even when the phylogenetic location of the different regimes is known. It is the case when the part of the tree under the influence of μ_ℓ forms a connected subtree for every ℓ (fig. 4.2 left, proof in the appendix B).

Note that connected subtrees do not need to form clades. The condition of connected subtrees is equivalent to there being a minimal number $m - 1$ of changes of selection regimes along the tree. This is the case, for instance, if there is only one shift in the selection optimum, dividing the tree into 2 connected subtrees, each with its own selection regime. In this common situation, the ancestral state and the two selection optima are not separately identifiable.

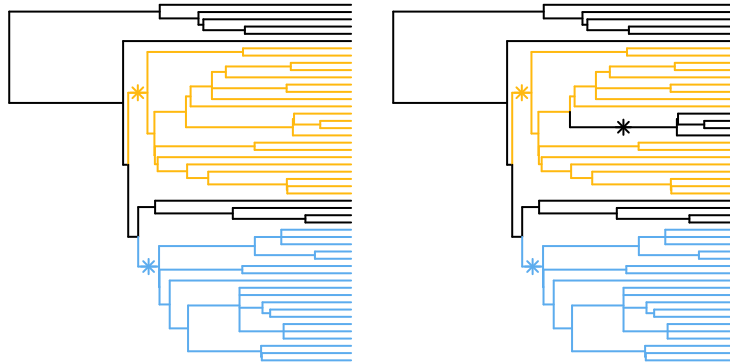


Figure 4.2: Edges are “painted” according to their selection regime, with one optimum μ_ℓ for each color. Unidentifiability case (left): every selection regime forms a connected component. Identifiability case (right): one regime (black) covers two disconnected parts in the tree.

On the contrary, OU models with multiple regimes are identifiable when selection regimes are not perfectly correlated with the tree, with at least one regime covering two disconnected subtrees (fig. 4.2 right). However, we need to

emphasize that identifiability only guarantees the uniqueness of the ML estimator. It does not guarantee its performance. In fact, the next section discusses situations when the MLE has poor precision even on large trees.

If we relax the assumption of known phylogenetic location(s) for the putative shifts in selection regimes (i.e. tree “painting”), then some of these location parameters also become unidentifiable, like the precise timing of a shift along a branch or the number of shifts along a single branch (see appendix B). Also, alternative scenarios with separate shifts along adjacent edges might not be distinguishable. In particular, 2 shifts along 2 sister edges have the same signature of expected values at extant taxa as one shift on either of the 2 sister edges and the other shift earlier along the parent edge. This is kept in mind for model selection in section 4.5.

4.3.2 Diagnostics and reparametrization

A lack of identifiability can be diagnosed by a lack of convergence during ML optimization, as algorithms may search the ridge of the likelihood surface forever. This might explain some convergence failures and the lack of reasonable estimates for some μ_i ’s reported by Butler and King (2004) or Beaulieu et al. (2012). Luckily, in studies using several selective regimes, hypotheses typically have at least one regime form a disconnected subtree —see for instance the study of fiber-type composition of iliofibularis muscle in lizards (Figure 1 in Scales et al., 2009). These models are thus identifiable with unique ML parameter estimates. Another source of non-identifiability is when $\alpha = 0$, because μ has no influence on the trait when selection is not acting. This can cause an almost flat ridge in the likelihood surface for very low estimates of α (Butler and King, 2004). Note also that a ridge in the likelihood plagues any Bayesian method, with the prior distribution having complete influence over which values get supported along the ridge.

To fix this lack of identifiability, we can re-parametrize the model to use a new design matrix of full rank. For example, with a single regime we can re-express the model by using $(\beta_0, \alpha, \sigma^2)$ instead of $(y_0, \mu, \alpha, \sigma^2)$ where $\beta_0 = y_0 e^{-\alpha T} + \mu(1 - e^{-\alpha T})$.

With this new parametrization, $Y = \beta_0 \mathbf{1} + \mathbf{e}$ but now β_0 is identifiable with a unique MLE.

To illustrate, we re-analyzed flower diameter of 25 Euphorbiaceae species under an OU model with ancestral optimum size μ_0 and a shift to a larger optimum flower size μ_1 at the base of the 3-species Rafflesiaceae clade (see Figure 1 in [Davis et al., 2007](#)). As noted before, y_0 , μ_0 and μ_1 are not identifiable separately. We re-parametrize the model using $(\beta_0, \beta_1, \alpha, \sigma^2)$ where $\beta_0 = y_0 e^{-\alpha T} + \mu_0(1 - e^{-\alpha T})$ is the expected flower size in non-Rafflesiaceae and $\beta_1 = (1 - e^{-\alpha t})(\mu_1 - \mu_0)$ is the difference in expected flower diameter between Rafflesiaceae and non-Rafflesiaceae. Here T is the age of the tree and t the age of the shift, which we do not need to estimate with our reparametrization. The model can be written $Y = \beta_0 \mathbf{1} + \beta_1 X_1 + \mathbf{e}$ where X_1 is a column of zeros and ones, with ones corresponding to Rafflesiaceae. We rescaled branch lengths in the phylogeny to a tree height of 1 and used `phylolm` ([Ho and Ané, 2013b](#)) on the log-transformed diameters (as in [Davis et al., 2007](#)). We obtained $\hat{\beta}_0 = 0.88$ (that is, $\exp(0.88) = 2.41$ mm), $\hat{\beta}_1 = 4.40$ (that is, an $\exp(4.4) = 81.2$ -fold size increase in extant taxa), $\hat{\sigma} = 2.38$ and $\hat{\alpha} = 11.6$, which indicates strong selection pressure.

Another approach is to assume that the ancestral state y_0 is random according to some prior distribution. If the OU process is homogeneous with a single selection optimum, the stationary distribution of that process is a natural choice for the prior distribution of y_0 : Gaussian with mean μ and variance $\sigma^2/(2\alpha)$. However, the stationary distribution has no clear definition when there are several regimes. To use this approach for the flower diameter data above, we assumed that y_0 followed a normal distribution with mean μ_0 and variance $\sigma^2/(2\alpha)$. By doing this, we drop y_0 from the set of parameters but keep the selection optima μ_0 , μ_1 . We also need to keep the shift age t in the model, which we assumed to be 0.605, at the base of Rafflesiaceae. We obtained the same $\hat{\sigma}^2$, $\hat{\alpha}$ as before, $\hat{\mu}_1 = 0.88$ and $\hat{\mu}_2 = 5.28$ (that is, a 196-fold increase in optimal flower size).

Another key ingredient to restore identifiability is to add fossil data, whenever possible. Fossil taxa can provide very influential information on both y_0 and μ (or μ_t 's) because they are at a shorter distance T from the root and under a lesser

influence of μ than contemporary tips.

4.4 (Non)-Microergodicity of selection parameters

Unlike in traditional regression models with independent residuals, identifiability does not guarantee that the MLE converges to the true parameter when the sample size increases indefinitely. A requirement stronger than identifiability is needed instead, historically called *microergodicity* (Stein, 1999). To understand this concept, we first consider two simple examples.

Example 1: Assume that $(Y_n)_{n=1}^{\infty}$ are independent binary observations with unknown mean $p = P\{Y_i = 1\}$, and $P\{Y_i = 0\} = 1 - p$. In this case the sample mean $\bar{Y}_n = (1/n) \sum_{i=1}^n Y_i$ is a “good” estimator of p , in the sense that it converges to p as the sample size n increases, from the law of large numbers. Here we keep gaining more precision by collecting more data.

Example 2: Consider binary observations $(Y_n)_{n=1}^{\infty}$ with mean $p = P\{Y_i = 1\}$. Assume that Y_2 is independent of Y_1 but that $Y_{2k-1} = Y_1$ and $Y_{2k} = Y_2$ for $k \geq 1$. Here, observing the entire sequence does not provide any extra information than observing Y_1 and Y_2 only, due to the extreme correlation between sampling units. Collecting more data does not increase information and obviously, there does not exist any estimator $f(Y_1, \dots, Y_n)$ that would converge to the true value of p .

To conceptualize the difference between these two examples and to understand the amount of information carried by comparative data, we use the concepts of orthogonal distributions and of microergodicity.

Definition: Two distributions P_1 and P_2 are *orthogonal* if there exists an event A such that $P_1(A) = 1$ and $P_2(A) = 0$.

Intuitively, this means that we can use data and tell with certainty whether these data came from P_1 or P_2 . If the event A is observed, then there is certainty that the model associated with P_2 is wrong. But if we observe that A does not occur, then the model associated with P_1 is certainly wrong. With example 1 we might want to compare the models where $p = p_1 = 0.5$ versus $p = p_2 = 0.8$,

and call P_i the distribution of the entire sequence $(Y_n)_{n=1}^{\infty}$ when $p = p_i$ ($i = 1$ or 2). Consider the event $A = \{\bar{Y}_n \text{ converges to } 0.5\}$. By the law of large numbers $P_1(A) = 1$ and $P_2(A) = 0$ so P_1 and P_2 are orthogonal. This reflects the fact that we can identify whether $p = 0.5$ versus $p = 0.8$ with certainty from the entire sequence $(Y_n)_{n=1}^{\infty}$ in example 1.

Definition: Let $(Y_n)_{n \geq 1}$ be an infinite sequence of observations and let P_{θ} be the joint distribution of $(Y_n)_{n \geq 1}$ under some model and parameters collected in a vector θ . A function $f(\theta)$ of θ is said to be *microergodic* if for every θ_1, θ_2 such that $f(\theta_1) \neq f(\theta_2)$, then P_{θ_1} and P_{θ_2} are orthogonal.

We think of $(Y_n)_{n \geq 1}$ as trait values obtained from infinitely many species as a best-case scenario, as if we were able to increase taxon sampling indefinitely. With the definition above, a function of parameters $f(\theta)$ is microergodic if the full data set $(Y_n)_{n \geq 1}$ contains enough information to tell the value of $f(\theta)$ with certainty. Unless θ is microergodic, there is no good estimator for it (Zhang, 2004), even from an infinite sample. Therefore microergodicity is necessary for a precise estimation of model parameters. In example 1, the mean p is microergodic, as the argument above used to distinguish $p = 0.5$ from 0.8 can be repeated to distinguish any two values of p . In example 2, p is not microergodic. The dependence between observations maintains uncertainty about the value of p , even from infinitely many Y_n observations.

Recently, Ho and Ané (2013a) investigated microergodicity for the one-regime OU model on an ultrametric tree and with a random ancestral state at the root (to avoid the unidentifiability issue mentioned earlier). While this simple model is unlikely to be adequate for most real traits, it is still considered when comparing hypotheses and more complex models are expected to require even more data. Ho and Ané (2013a) proved that if the height of the tree is bounded, as when sampling from a group of interest, then the selection optimum μ is not microergodic. Therefore, information from contemporary taxa is not sufficient to estimate μ exactly, even if infinitely many taxa were observable. On the other hand, σ^2 was shown to be microergodic in the case when the tree has

many “young” internal nodes. Unfortunately, an additional assumption to ensure enough variation in internal node ages, was required for the selection strength α and the stationary variance $\gamma = \sigma^2/(2\alpha)$ to be microergodic. The microergodicity of α is necessary for a good estimator $\hat{\alpha}$ to exist, but is it sufficient? This question remains open. In the particular case of a symmetric tree, we can prove that the restricted MLE of α converges to the true α with more data, if the microergodicity condition is met (using tools in [Ho and Ané, 2013a](#), appendix B).

To illustrate these theoretical results, we simulated data on several very large phylogenies from across the tree of life: a 9993-species phylogeny of birds ([Jetz et al., 2012](#)), a 4507-species mammal tree ([Bininda-Emonds et al., 2007](#)), a 839-taxon tree on Fabales ([Simon et al., 2009](#)) and a 140-species phylogeny of ants ([Moreau et al., 2006](#)). We also used a 400-language phylogeny ([Gray et al., 2009](#)). For simplicity, we present here the results on the largest tree only (9993 taxa). The results on the other phylogenies are similar (see Supplementary Material).

Twenty sequences of nested phylogenies from 50 to 9993 taxa were created by randomly selecting subsets of taxa from 20 bootstrap trees from [Jetz et al. \(2012\)](#), conditional on the root being the only common ancestor of the selected taxa to guarantee that all trees have the same ancestral species and same height. Trees were rescaled by a common factor to have height 1. Data were simulated from a one-regime OU model with $\mu = 0$, stationary variance $\gamma = 1$, selection strength $\alpha = 0.1, 1$ or 10 , and variance rate $\sigma^2 = 2\alpha\gamma$. This model is very close to a Brownian motion when $\alpha = 0.1$ and very close to phylogenetic independence when $\alpha = 10$. Figure 4.3 shows the MLEs $\hat{\mu}$, $\hat{\gamma}$, $\hat{\alpha}$ and $\hat{\sigma}^2$, which were obtained using `phylolm` ([Ho and Ané, 2013b](#)). When selection is not strong ($\alpha = 0.1$ or 1), the accuracy of $\hat{\mu}$ does not improve with more taxa, illustrating the non-microergodicity of μ . Under weak selection ($\alpha = 0.1$) $\hat{\alpha}$ and $\hat{\gamma}$ are strongly biased with few taxa, with selection being overestimated almost all the time with less than 100 taxa. Their bias improves with more taxa but their precision does not: there is as much uncertainty about α and γ from 9993 taxa as from 50. This illustrates that α and γ are not microergodic. On the other hand, the variance rate σ^2 is estimated precisely, regardless of the selection strength. Under

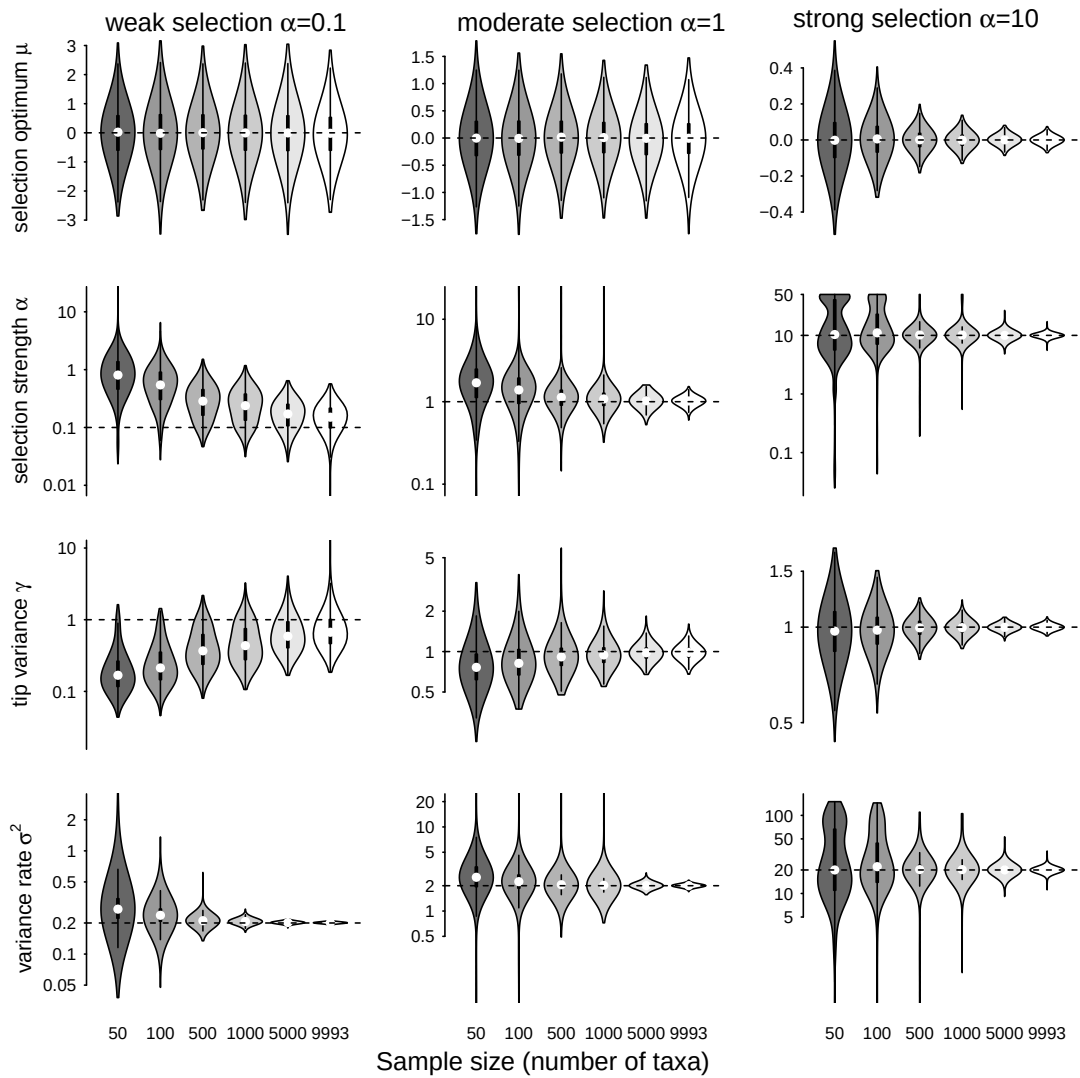


Figure 4.3: Boxplot and violin plots showing the distribution of the MLE of μ , γ , α , and $\sigma^2 = 2\alpha\gamma$ on subtrees from the 9993-species bird phylogeny (Jetz et al., 2012) with 2000 simulations at each sample size. Each column corresponds to one set of true parameter values: $\mu = 0$; $\gamma = 1$; $\alpha = 0.1$ (left), 1 (middle), 10 (right); and $\sigma^2 = 2\alpha\gamma$.

strong selection ($\alpha = 10$), phylogenetic inertia is very low and all parameters are estimated as usual: with little or no bias and with increasing precision from more taxa. These simulations illustrate the main difficulty of detecting selection. When it is moderate, phylogenetic correlation between taxa greatly reduces the actual amount of information on selection parameters, especially on the target of selection. With these limitations in mind for a single selection regime, we now turn to models with multiple regimes.

4.5 Power to detect shifts in the selection optimum

Detecting changes in selection regime is important to study the drivers of selection pressure. For example, [Brawand et al. \(2011\)](#) detected evolutionary shifts in gene expression in testes for a large number of mammalian genes, by comparing a single-regime OU model to a two-regime OU model in which the optimal gene expression level μ undergoes a shift on a specific lineage (see also [Rohlf et al., 2014](#)). In this section, we use simulations to study the power to detect such shifts. We first consider the case when the phylogenetic location of the shift is known, such as when it is hypothesized to be driven by an environmental change or change in another trait that can be mapped onto the phylogeny (e.g. [Butler and King, 2004](#)). Next we consider the case when the number and phylogenetic position of the shift(s) need to be estimated.

Known location: We simulated data from an OU model with two selection regimes on the 4507-species mammal tree from [Bininda-Emonds et al. \(2007\)](#). The shift in selection optimum was placed at the base of Euarchontoglires (Figure 4.4, left), so that each regime applied to about half the taxa. Twenty sequences of nested trees from 10 to 4507 taxa were created by randomly selecting subsets of taxa, under the constraint that the number of sampled Euarchontoglires was half the total number of sampled taxa. Along each tree, 100 data sets were simulated according to the OU model with ancestral state and ancestral optimum $y_0 = \mu_0 = 0$, $\gamma = 1$, $\alpha = 1$, and a Euarchontogline optimum shift

to $\mu_1 = 1.079, 2.157$ or 4.314 . This model corresponds to expected values in extant taxa of $\beta_0 = 0$ for non-Euarchontoglires and of $\beta_1 = 0.5, 1$ or 2 for Euarchontoglires. The parametrization with β values was used for inference, to

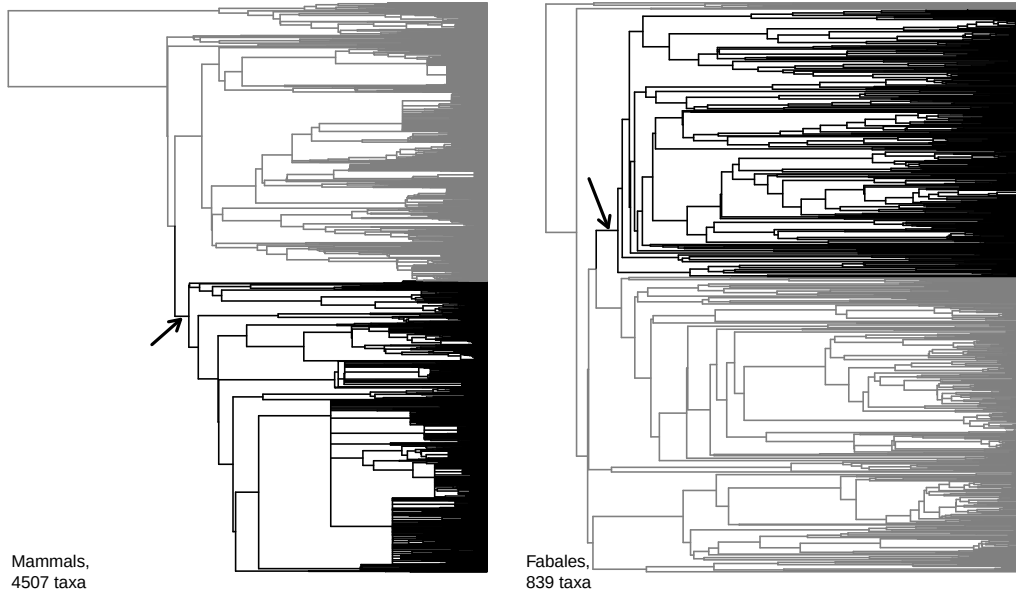


Figure 4.4: Position of change in the selection optimum, used in simulations along a tree for mammals (Bininda-Emonds et al., 2007) and Fabales (Simon et al., 2009).

avoid the non-identifiability issue described earlier. Testing the shift in selection regime is then equivalent to testing whether $\beta_1 = 0$ or not. Figure 4.5 (left) shows that the estimation error in β_1 decreased from 10 to 50 taxa, but did not change much from 50 to 4507 taxa. This phenomenon reflects the non-microergodicity of both selection optima, and suggests that the power to detect the shift ($\beta_1 \neq 0$) does not increase significantly as the tree grows. The same simulation was done on the 839-taxon tree on Fabales (Simon et al., 2009) with similar results (see Supplementary Material).

From each simulated data set, we rejected the null hypothesis of no shift if $|\hat{\beta}_1|/\text{SE}(\hat{\beta}_1) > 1.96$ where the standard error of $\hat{\beta}_1$ for each tree size was estimated

as the standard deviation of $\hat{\beta}_1$ over the 2000 simulations. Fig. 4.5 (right) shows that the power depends mostly on the effect size and very little on the tree size: when the signal is weak (effect size $\beta_1/\sqrt{\gamma} = 0.5$) the shift is detected only 39.8% of the time even from 4507 taxa. On the other hand, the shift is detected 99% of the time from only 50 taxa when the signal is strong (effect size $\beta_1/\sqrt{\gamma} = 2$). In the example above on flower diameter evolution from Davis et al. (2007), the shift at the base of Rafflesiaceae was statistically significant ($SE(\hat{\beta}_1) = 0.316$, p-value 10^{-12}). This is likely *not* because of a large sample size: there were 22 and 3 species in each regime. Instead, the shift was detectable because of a very large effect size; $\beta_1/\sqrt{\gamma}$ was estimated to be 8.92.

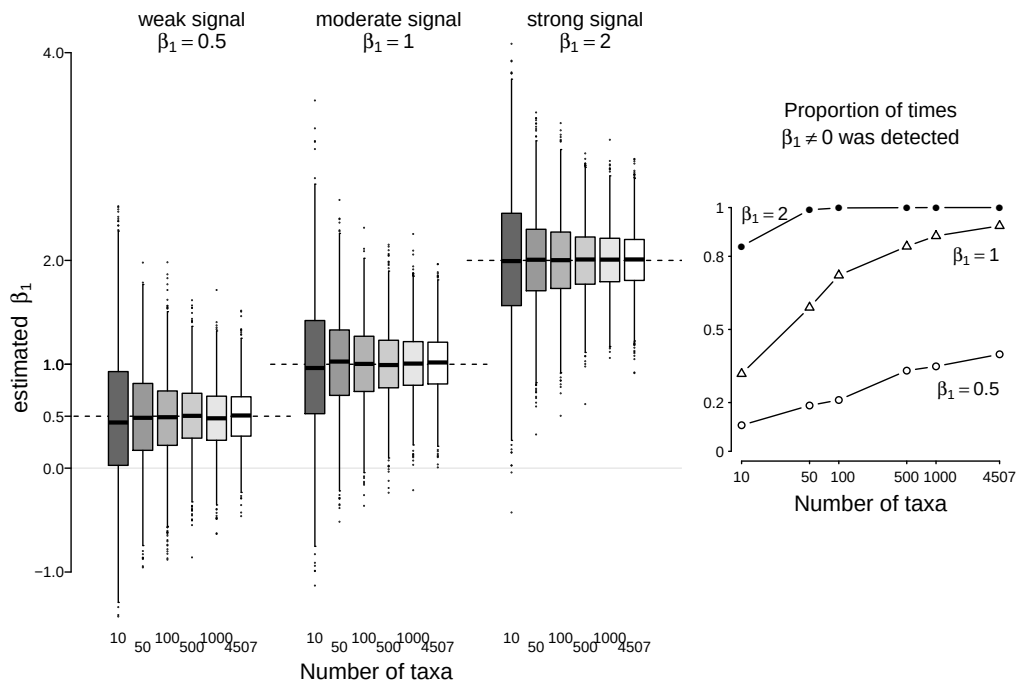


Figure 4.5: Left: boxplots showing the distribution of the MLE of the shift, $\hat{\beta}_1$, from data simulated on the 4507-species mammal tree in fig. 4.4 (left). Right: proportion of times the no-shift hypothesis $H_0 : \beta_1 = 0$ was rejected.

To further illustrate that the power to detect a selection optimum shift depends little of tree size, we used the approach introduced by Boettiger et al. (2012) to

detect the shift. This method uses a bootstrap-based likelihood-ratio test and simultaneously estimates the power of the test. We simulated data from the OU model on the 839-taxon tree on Fabales (Simon et al., 2009, Figure 4.4 right) with one shift in μ at the base of a clade containing roughly half the total number of taxa. Twenty sequences of nested trees from 10 to 839 taxa were created by sampling taxa randomly, but requiring an equal number of taxa in each regime when $n < 839$. Along each tree, one data set D was simulated under the OU model as before with $\gamma = 1$, $\alpha = 1$, $y_0 = \mu_0 = 0$ and a shift varying from $\mu_1 = 0.873, 1.747$ to 3.494 , corresponding to $\beta_0 = 0$ and $\beta_1 = 0.5, 1$ or 2 . Each dataset D was analyzed as follows (Boettiger et al., 2012). First, model parameters were estimated under a single regime (model 1). Bootstrap data sets $D_1^*, D_2^*, \dots, D_{1000}^*$ were then simulated independently with these estimated parameters under model 1. For each D_i^* , the log likelihood ratio $\delta = 2(\log L_2 - \log L_1)$ was computed to compare model 1 to the two-regime OU model (model 2) with the shift placed at its correct location. Second, D was used again to estimate parameters under model 2 and parametric bootstrap replicates were simulated under the two-regime model to obtain a sample of δ values under that model. The area of the overlap between the two distributions of δ was then used to measure the power to detect the shift: the smaller this overlap, the greater the power (see fig. 4.6 insets for examples of this overlap). As figure 4.6 shows, the power to detect the shift increases very little (i.e. the overlap decreases very little) with the number of taxa. Instead, the overlap and hence the power are most influenced by the true magnitude of the shift.

Unknown location: When the number and phylogenetic placement of shifts are unknown, a model selection procedure is needed. The most complex model includes a shift on every branch b in the tree:

$$Y = \beta_0 \mathbf{1} + \sum_b \beta_b X_b + e$$

where e has phylogenetic covariance from the OU process and each X_b is a

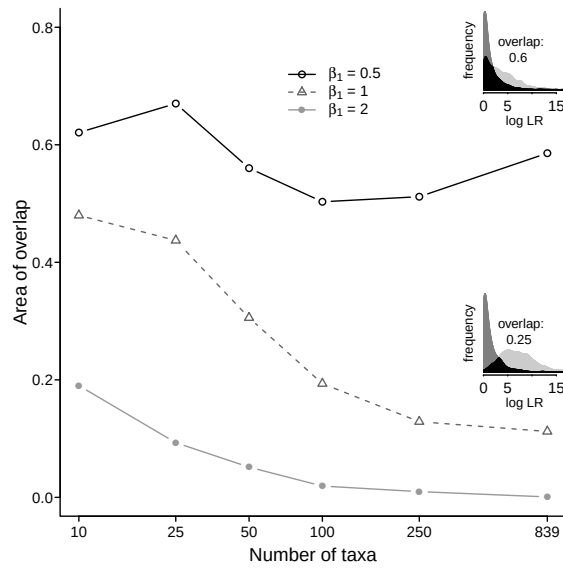


Figure 4.6: Average overlap between two distributions of the log likelihood ratio (log LR) statistic δ : under the 1-regime and two-regime models. Data were simulated under two regimes on the 839-taxon tree in fig. 4.4 (right). Inset: example distribution of δ under 1 regime (dark gray) or 2 regimes (light gray) and overlap (black).

column of zeros and ones, with ones corresponding to descendants of branch b . The expected change β_b in extant taxa is linked to the shift in selection optimum $\Delta\mu$ through $\beta_b = (1 - e^{-\alpha t_b})\Delta\mu$ where t_b is the shift's age, due to phylogenetic inertia. We seek to identify on which branch there is a shift, that is, for which branch $\beta_b \neq 0$. To do so, we consider here a traditional stepwise selection method. The simplest one-regime model is first evaluated. At each step, a list of candidate models is made by modifying the current model: either adding a shift on a branch, moving an existing shift to a neighbor branch, or dropping a shift. The procedure stops if the current model is better than every candidate model based on a criterion like AIC (Akaike, 1974) or BIC (Schwarz, 1978). Otherwise the best model in the list is selected and used as the new current model for the next step. For identifiability purposes again, we consider a maximum of one shift per

branch and our procedure disregards models with “extinct” regimes, i.e. regimes applying to internal branches only. Models with shifts located on sister branches are considered, with the caveat that their exact location among the 2 sister edges and their parent edge is not identifiable.



Figure 4.7: Left: 140-species phylogeny of ants (Moreau et al., 2006) used for simulations, with true shifts in the selection optimum μ located on branches b_1 , b_2 and b_3 . AIC selected 10 shifts when allowed no more than 10, along edges marked in bold. Edge numbers indicate the order in which estimated shifts were added by AIC. Right: simulated normalized trait data.

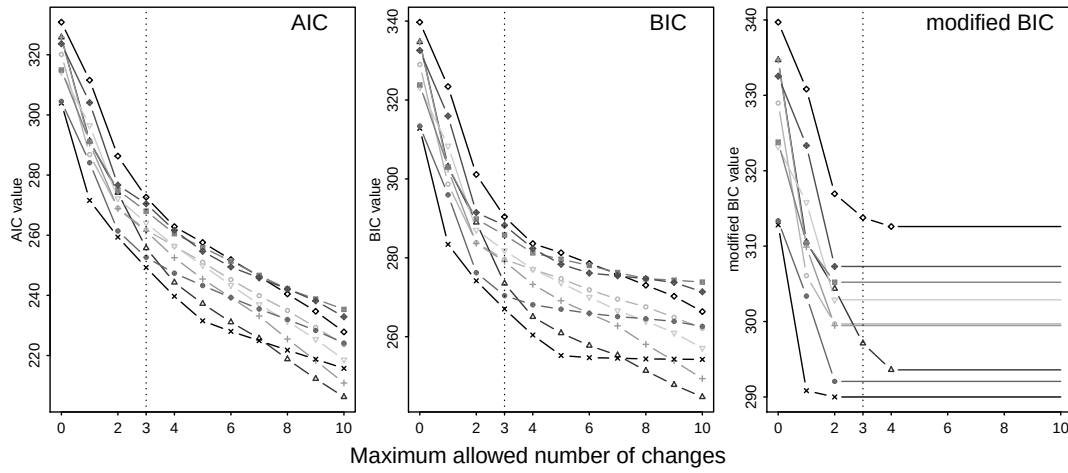


Figure 4.8: AIC (left), BIC (middle) and modified BIC (right) values of the best OU model with a maximum allowed number m of shifts in the selection optimum. Each line corresponds to 1 data set, simulated under OU evolution with 3 true shifts. Separate points are used for strictly decreasing values.

We simulated 10 data sets along a 140-taxon tree under the OU model with 3 shifts, located on branches labeled b_1 , b_2 and b_3 (fig. 4.7, left). We used a moderate selection strength $\alpha = 1$ after rescaling the tree to a height of 1, $\gamma = 1$, and simulated shifts in selection optima leading to $\beta_{b_1} = 3$, $\beta_{b_2} = 2$ and $\beta_{b_3} = 1$. The first shift is expected to be easy to detect but the other two are not, because their magnitude is less and because they are located on sister edges. As discussed before, due to a lack of identifiability, these two shifts are expected to be detected on either b_2 , b_3 and/or on their parent edge. The stepwise model selection procedure was used with the constraint that the number of shifts in any proposed model may not exceed a given number m , with m from 0 to 10. Figure 4.7 shows one simulation, for which AIC selected 10 shifts when allowed no more than $m = 10$. The first 3 corresponded to the true shifts (with some location error). With one exception, all other estimated shifts corresponded to a single taxon or a pair of sister taxa, misinterpreting one or two extreme value(s) for a shift in selection regime. Figure 4.8 shows a steady decrease in AIC and BIC values with m , leading to the estimation of the maximum allowed of 10 shifts in all cases,

and a strong overestimation of regime shifts. With the allowed maximum m increased to 40, AIC selected 40 shifts for all 10 data sets. BIC also selected 40 shifts for 9 data sets and 28 shifts for the remaining data set.

We conjecture that the failure of AIC and BIC may be caused by two phenomena related to the very large number of candidate models on large trees. First, there are more models to choose from than available data points, a problem coined “large p , small n ” in the statistics literature. Indeed, with n taxa on a rooted tree there are $2n - 2$ models with one shift, one for each of the $2n - 2$ branches. So with a single shift and a single extra parameter, the number of potential models is already larger than the number of data points, providing ample opportunity for overfitting. The second issue with heterogeneous tree models is combinatorial: the number of models explodes with the number of shifts. On the 140-taxon tree used above, there are $2n - 2 = 278$ models with a single shift, but 38,503 models with 2 shifts (from choosing 2 of the 278 edges), and $6.4 * 10^{17}$ models with 10 shifts. This explosion gives an advantage to 10-shift models over 2-shift models, and to 2-shift models over 1-shift models when using AIC or BIC, which are not meant to protect against multiple testing. This phenomenon occurs in change-point models for time series, for which the failure of AIC and BIC has already been recognized (e.g. [Zhang and Siegmund, 2007, 2012](#); [Rigaill et al., 2012](#)). In particular, [Zhang and Siegmund \(2007\)](#) introduced a modified BIC with the following penalty:

$$3 \log n + (2m - 1) \log n + \sum_{i=0}^m \log n_i$$

where m is the number of shifts, n is the sample size, n_0 is the number of observations which are not influenced by any shift, and n_i is the number of observations influenced by shift i ($n_0 + n_1 + \dots + n_m = n$). The first term $3 \log n$ penalizes β_0 , γ , and α as in the usual BIC. The second term $(2m - 1) \log n$ penalizes each additional change twice as much as the BIC penalty. The reason for doing so is because the model actually adds two more parameters for each

change: its value and its position on the tree. The last term penalizes each regime based on the number of observations influenced by that regime. Notice that this term is less than $(m + 1) \log n$. So, the penalty of the modified BIC is between $(2m + 2) \log n$ to $(3m + 3) \log n$, which is roughly 2 to 3 times greater than the BIC penalty as m gets larger.

We applied this modified BIC in our stepwise selection procedure with the same 10 data sets as before, naively ignoring the independence assumption in [Zhang and Siegmund \(2007\)](#). This modified BIC had much better performance than AIC and BIC in our simulations (fig. 4.8, right), selecting between 2 and 4 shifts only. In all these 10 cases, the true shift on branch b_1 was detected first. The second detected shift was either on the parent branch of b_2 and b_3 (6 times) or on b_2 (4 times) and overfitting at or near external edges was avoided. We repeated the modified BIC stepwise selection on 100 new simulated data sets. The average number of detected shifts was 2.55 (standard deviation 0.99), demonstrating that overfitting is not an issue for modified BIC here. Further improvements would likely be achieved by adjusting for phylogenetic correlation in the BIC penalty, where n is replaced by an effective sample size (see [Ané, 2008](#), for the BM model). The modified BIC could also be used in a variety of problems, such as to detect changes in rate or in selection strength. However, more extensive simulations are needed to quantify the performance of modified BIC penalties for these problems.

Fully Bayesian methods are not immune to this large p - small n issue if the prior distribution is not carefully thought to counteract the combinatorial explosion of models. [Venditti et al. \(2011\)](#) used a Bayesian reversible-jump framework to detect changes in the evolutionary rate of body size. They detected evidence of shifts in about one-third of all branches (1494 branches) in their 3185 mammal phylogeny. Their prior distribution for the number of changes was not discussed, so it is possible that the very large estimated number of shifts was driven by the combinatorial explosion of models and equal weights to all models a priori.

[Eastman et al. \(2011\)](#) proposed a truncated Poisson prior for the number of

shifts, which places a high prior weight on models with few shifts and a much lower weight a priori on each individual model with many shifts. [Rabosky \(2014\)](#) used a similar approach to infer differential rates of species diversification or shifts in trait evolutionary rate, using a compound Poisson prior process for the location and number of shifts. Again, this choice controls the combinatorial explosion of models by giving a Poisson prior probability to the entire set of models with a given number of shifts. Simulations showed that overfitting was not an issue.

The modified BIC used here is one attempt to solve the large p - small n problem. Many methods have been developed to solve this problem for independent data, such as LASSO ([Tibshirani, 1996](#); [Zhao and Yu, 2006](#)) or least angle regression ([Efron et al., 2004](#)). As pointed out for the modified BIC, more work should be done to adjust these methods for phylogenetic correlation and adapt them to change-point detection problems on evolutionary trees.

4.6 Conclusion

We discussed here difficulties for studying trait evolution from present-day species: unidentifiability of ancestral states, non-microergodicity of selection parameters for OU models, and limited power to detect changes in selection regimes from present-day species. These issues are intrinsic to the nature of comparative data with observations at the tips of the tree only. Our study does not mean that OU models are limited and should be abandoned. If trait evolution truly followed an OU process, then analyses using an OU model are certainly most appropriate. But investigators should be aware from the onset that increasing data collection from more present-day taxa may not provide the power needed to discriminate between hypotheses. These difficulties are not tied to maximum likelihood: non-microergodic parameters are bound to be estimated with imprecision, regardless of the estimation method. When possible, a better strategy to increase precision and power is to obtain fossil data. As mentioned in [Section 4.3](#), fossil data restore the identifiability of both the ancestral state and the selection optimum, which can

then be estimated by maximum likelihood, and with increasing precision in some situations (see [Ho and Ané, 2013a](#)). The benefit of combining fossil with living taxa has been recognized empirically and is receiving increased recognition ([Slater and Harmon, 2013](#)), with interest to test more and more complex evolutionary processes ([Slater, 2013](#)).

Finally, we illustrated the breakdown of traditional model selection criteria (AIC and BIC) to detect changes in the trait evolution process and we identified a “large p , small n ” issue: the number of candidate models is larger than the number of taxa, and explodes with the number of shifts. We introduced a modified criterion borrowed from change-point models for time series, which does not overestimate the number of shifts like AIC and BIC do. While fully Bayesian methods have been proposed for related heterogeneous trait evolution models on trees, more work is needed to discover appropriate model selection tools in a frequentist framework.

5 A LINEAR-TIME ALGORITHM FOR TRAIT EVOLUTION

Abstract We developed a linear-time algorithm applicable to a large class of trait evolution models, for efficient likelihood calculations and parameter inference on very large trees. Our algorithm solves the traditional computational burden associated with two key terms, namely the determinant of the phylogenetic covariance matrix $\mathbf{V}$ and quadratic products involving the inverse of $\mathbf{V}$. Applications include Gaussian models such as Brownian motion (BM) derived models like Pagel’s lambda, kappa, delta and the early-burst model; Ornstein-Uhlenbeck models to account for natural selection with possibly varying selection parameters along the tree; as well as non-Gaussian models such as phylogenetic logistic regression, phylogenetic Poisson regression and phylogenetic generalized linear mixed models. Outside of phylogenetic regression, our algorithm also applies to phylogenetic principal component analysis, phylogenetic discriminant analysis or phylogenetic prediction. The computational gain opens up new avenues for complex models or extensive resampling procedures on very large trees. We identify the class of models that our algorithm can handle as all models whose covariance matrix has a 3-point structure. We further show that this structure uniquely identifies a rooted tree whose branch lengths parametrize the trait covariance matrix, which acts as a similarity matrix. The new algorithm is implemented in the R package `phyloIm`, including functions for phylogenetic linear regression and phylogenetic logistic regression.

Funding This work was funded in part by NSF award DMS-1106483 and DEB-0949121.

5.1 Introduction

Statistical methods are well developed now to account for phylogenetic correlation in the analysis of quantitative traits across species. Most of these methods use

models derived from a Brownian motion (BM) or an Ornstein-Ühlenbeck (OU) process along the tree, where the tree is informed by molecular data. These models are now used on increasingly large data sets from tree-of-life projects with large numbers of taxa (e.g. [Smith and Donoghue, 2008](#); [Cooper and Purvis, 2010](#); [Jetz et al., 2012](#)) or on a large number of traits such as from expression profiles (e.g. [Brawand et al., 2011](#)). The traditional likelihood calculations use generic methods and do not raise computational issues for a few traits and a dozen taxa. However, these methods are challenged by large phylogenetic covariance matrices on large numbers of taxa, causing very long computing times at best, and possibly failure or inaccurate results ([Hadfield and Nakagawa, 2010](#)). For instance, [Venditti et al. \(2011\)](#) studied body mass across 3185 mammal species. Using generic methods, one single analysis of their data took us 1 hour 20 minutes (using the BM model in `fitContinuous` from the R package `geiger` v1.3-1). To determine lineages where body mass evolution experienced a change in its evolutionary rate, their complex Bayesian model required millions of likelihood calculations. Therefore, traditional likelihood calculations were prohibitively slow and a fast algorithm was absolutely needed for their study. Recently, [Freckleton \(2012\)](#) showed how to calculate the likelihood of BM models in linear time based on an algorithm proposed by [Felsenstein \(1973\)](#), which [Venditti et al. \(2011\)](#) used (personal communication). Independently, [FitzJohn \(2012\)](#) recently proposed a fast Gaussian elimination method to calculate the likelihood of BM and OU models, philosophically similar to the pruning algorithm for molecular sequence data. However, the applicability of these methods to non-Gaussian traits is not clear and remains open ([Freckleton, 2012](#)).

We present here a linear-time algorithm that applies to a broader class of phylogenetic trait models than the fast algorithms cited above. We separately calculate two computationally difficult terms involving the phylogenetic covariance matrix. These are necessary for many phylogenetic methods, making our algorithm applicable to a number of non-Gaussian models for which these two terms are used separately and in various ways. Our algorithm can be used to speed up parameter estimation and likelihood calculation of trait data under

modified BM models like Pagel’s lambda, kappa, delta and the early-burst model (as can also be done with the methods of [Felsenstein, 1973](#); [Freckleton, 2012](#)). These models can be used to quantify the level of phylogenetic signal in single traits, or they can quantify and account for this phylogenetic signal while testing the correlation between several traits. Our algorithm can also be used to fit OU models with possibly varying selection parameters along the tree ([Butler and King, 2004](#); [Beaulieu et al., 2012](#)), as does [FitzJohn \(2012\)](#). Additionally, our algorithm applies to several non-Gaussian models like phylogenetic logistic regression, phylogenetic Poisson regression, phylogenetic generalized linear model ([Paradis and Claude, 2002](#); [Ives and Garland, 2010](#)), and phylogenetic generalized linear mixed models ([Ives and Helmus, 2011](#)), thus allowing these methods to be used on very large data sets for the first time. Other applications include phylogenetic principal component analysis (PCA, [Revell, 2009](#)), phylogenetic discriminant analysis ([Motani and Schmitz, 2011](#)) and phylogenetic prediction ([Garland and Ives, 2000](#)).

Our method is implemented in the R package `phylolm`, with functions for phylogenetic linear regression and phylogenetic logistic regression. Our algorithm requires that the underlying covariance matrix $\mathbf{V}$ for which inversion is to be avoided belong to a class of generalized 3-point structured matrices. In what follows, we first describe the terms responsible for the computational bottleneck, our algorithm for calculating them efficiently, and computation time comparisons. We then describe the class of 3-point structured matrices which our algorithm applies to, with its connection to a rooted tree structure. Finally, we show how our algorithm applies to a number of tools for trait evolution, including methods for which no other fast algorithm was available before.

5.2 Algorithm

5.2.1 Calculating the likelihood for Gaussian traits

To motivate the method, we start by examining the likelihood for Gaussian trait models as it is used for both maximum likelihood and Bayesian inference. In this context, we explain what terms form the computational bottleneck. Consider a single trait, with data Y at n tips of a tree and an evolution model under which Y has a multivariate normal distribution with mean μ_θ and phylogenetic covariance matrix V_θ , for some covariance parameter θ . This covariance takes a specific form depending on the underlying trait evolution model. The simplest example is the BM model for the error term, which corresponds to $\theta = \sigma^2$ and $V_\theta(i, j) = \sigma^2 t_{ij}$, where t_{ij} is the tree length from the root to the last common ancestor of tips i and j . In a regression framework, the mean value is $\mu_\theta = X\beta$ where the matrix X contains predictors in its columns including a column vector of ones, $\mathbf{1}$. If there are no predictors then $\mu_\theta = \beta_1 \mathbf{1}$ is constant and the intercept β_1 typically represents the ancestral trait value at the root. The likelihood $f(Y|\theta)$ is then given by

$$-2 \log f(Y|\theta) = n \log(2\pi) + \log |V_\theta| + (Y - X\beta)' V_\theta^{-1} (Y - X\beta).$$

Given θ , the estimated regression coefficients are $\hat{\beta} = (X'V_\theta^{-1}X)^{-1}X'V_\theta^{-1}Y$. If there are no predictors, then this is simply the estimated mean (or ancestral state at the root) and simplifies to $\hat{\beta}_1 = (\mathbf{1}'V_\theta^{-1}\mathbf{1})^{-1}\mathbf{1}'V_\theta^{-1}Y$.

Traditional calculations use generic methods for calculating the determinant $|V_\theta|$ and the matrix inverse V_θ^{-1} , such as the Cholesky or QR decomposition, or singular value decomposition (Lawson and Hanson, 1995). However, the full matrix inverse V_θ^{-1} is not needed. Our approach consists in calculating the following key terms only: $|V_\theta|$ and the quadratic products $Y'V_\theta^{-1}Y$ and $X'V_\theta^{-1}Y$, without actually inverting V_θ . The general algorithm to calculate these quadratic products can then also be used to obtain the precision matrix $X'V_\theta^{-1}X$ whose inverse gives the variances of the β regression coefficients. These provide the basis to get standard errors, p-values and test hypotheses about coefficients. Note

that alternative fast algorithms ([Felsenstein, 1973](#); [Freckleton, 2012](#); [FitzJohn, 2012](#)) focus on the likelihood only, not on the standard errors of coefficients as we do here. The gain in computation efficiency is all the more important when the phylogenetic signal parameters in θ are unknown, requiring repeated iterations to estimate θ by maximum likelihood or Markov Chain Monte Carlo (MCMC).

[Hadfield and Nakagawa \(2010\)](#) already discussed the computational burden of inverting the covariance matrix. They report attempts to calculate $\mathbf{V}^{-1}$ for the 4510 mammal species supertree from [Bininda-Emonds et al. \(2007\)](#) that either failed completely or took up to one month of computing time. They discussed similarities between the phylogenetic linear mixed model and animal models in quantitative genetics, for which efficient algorithms have been developed early on. They noted that an expanded covariance matrix can be inverted analytically if the traits at all internal nodes are considered. Therefore, they argued for algorithms that track the unknown trait values at internal nodes, such as expectation-maximization (EM, [Dempster et al., 1977](#)) or MCMC for Bayesian analyses. Here, our algorithm removes the constraint of inverting a large matrix without relying on an expanded state space or on the convergence of an EM algorithm.

5.2.2 The algorithm

This algorithm is designed to calculate the determinant $|\mathbf{V}|$ and quadratic quantities of the form $\mathbf{Q} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$, for any covariance matrix $\mathbf{V}$ within the class of “3-point structured” matrices. This concept is somewhat technical and postponed to the next section, where we show a connection between a 3-point condition on similarity matrices and a rooted tree structure. $\mathbf{X}$ and $\mathbf{Y}$ must have the same number of rows as $\mathbf{V}$, but can have any number of columns. The algorithm also needs to recursively calculate the following quantities: $p = \mathbf{1}'\mathbf{V}^{-1}\mathbf{1}$, $\hat{\mu}_Y = \mathbf{1}'\mathbf{V}^{-1}\mathbf{Y}/p$ and $\tilde{\mu}_X' = \mathbf{X}'\mathbf{V}^{-1}\mathbf{1}/p$. For the BM phylogenetic regression above, the algorithm needs to be applied to the response trait $\mathbf{Y}$ (one column) and to $\mathbf{X}$ containing one column per predictor. Then $\hat{\mu}_Y$ and $\tilde{\mu}_X$ give the ancestral state estimates of

each trait at the root and p is proportional to the precision of these estimates. To obtain the standard errors of regression coefficients, the algorithm also needs to be applied with $\mathbf{Y}$ replaced by $\mathbf{X}$ to get $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$. The algorithm runs in linear time without matrix inversion using a tree-traversal algorithm, as follows. Based on its 3-point structure, we can write $\mathbf{V}$ as

$$\mathbf{V} = t\mathbf{1}\mathbf{1}' + \mathbf{A} \quad \text{with } \mathbf{A} = \begin{pmatrix} \mathbf{V}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{V}_k \end{pmatrix} \quad (5.1)$$

where $t \geq 0$ and $\mathbf{A}$ is a diagonal matrix with one block for each subtree stemming from the root. Each block $\mathbf{V}_s$ in $\mathbf{A}$ is of the form (5.1) for the corresponding subtree rooted by its parent edge (figure 5.1). The algorithm uses a post-order tree traversal, starting from the external edges and working towards the root.

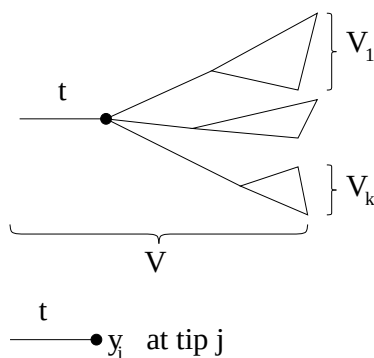


Figure 5.1: Recursion for the tree traversal: at each internal node, $\mathbf{V}$ is of the form $t\mathbf{1}\mathbf{1}' + \mathbf{A}$ with $\mathbf{A}$ block diagonal. Each block $\mathbf{V}_s$ of $\mathbf{A}$ is of this form too.

1. Initialization: for a tree with a single tip, $\mathbf{V} = t > 0$. Set $\log|\mathbf{V}| = \log t$, $p = 1/t$, $\hat{\mu}_Y = \mathbf{y}$, $\tilde{\mu}'_X = \mathbf{x}'$ and $\mathbf{Q} = \mathbf{x}'\mathbf{y}/t$, where $\mathbf{x}$ and $\mathbf{y}$ are the rows of $\mathbf{X}$ and $\mathbf{Y}$ corresponding to the current tip.
2. For a tree with two or more tips, consider (5.1) for $\mathbf{V}$ on this tree. Denote

p_s , $\tilde{\mu}'_{X,s}$, $\hat{\mu}_{Y,s}$ and $\mathbf{Q}_s$ the quantities obtained recursively from subtree s . Define $p_A = \sum_{s=1}^k p_s$ and weights $w_s = p_s/p_A$. Then $p = p_A/(1 + tp_A)$, $\hat{\mu}_Y = \sum_{s=1}^k w_s \hat{\mu}_{Y,s}$, $\tilde{\mu}'_X = \sum_{s=1}^k w_s \tilde{\mu}'_{X,s}$, $\mathbf{Q} = \sum_{s=1}^k \mathbf{Q}_s - \frac{tp_A^2}{1+tp_A} \tilde{\mu}'_X \hat{\mu}_Y$ and $\log |\mathbf{V}| = \sum_{s=1}^k \log |\mathbf{V}_s| + \log(1 + tp_A)$.

3. At the root of the full tree: return $\log |\mathbf{V}|$ and $\mathbf{Q}$ then stop.

For a growing number of taxa n , our algorithm is linear in time because it uses a single tree traversal. It is also linear in memory requirement as it does not seek to calculate the large $\mathbf{V}^{-1}$ matrix. If one actually needed the full inverse matrix, then the Woodbury formula could be used to obtain $\mathbf{V}^{-1}$ in $O(n)$ steps, as detailed in the proof below.

Proof of formulas for the algorithm. We combine the decomposition (5.1) for $\mathbf{V} = \mathbf{A} + t\mathbf{1}\mathbf{1}'$ with the Woodbury formula for the matrix inverse and Sylvester's determinant (Hager, 1989) for matrices of the form $\mathbf{M} = \mathbf{A} + \mathbf{UCV}$:

$$\begin{aligned} \mathbf{M}^{-1} &= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}, \\ \det \mathbf{M} &= \det \mathbf{A} \det \mathbf{C} \det (\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U}). \end{aligned}$$

We use these formulas with the block diagonal matrix $\mathbf{A}$ for which each block $\mathbf{V}_i$ is treated recursively, and obtain:

$$\mathbf{V}^{-1} = \mathbf{A}^{-1} - \frac{t}{1 + tp_A} \mathbf{A}^{-1} \mathbf{1} \mathbf{1}' \mathbf{A}^{-1}$$

where $p_A = \mathbf{1}' \mathbf{A}^{-1} \mathbf{1} = \sum_{i=1}^k \mathbf{1}' \mathbf{V}_i^{-1} \mathbf{1}$, and $\det \mathbf{V} = (1 + tp_A) \det \mathbf{A}$. The recursion for the determinant is immediate then, because $|\mathbf{A}| = \prod_{s=1}^k |\mathbf{V}_s|$. The recursion for the other components uses the facts that $p_A \hat{\mu}'_X = \mathbf{X}' \mathbf{A}^{-1} \mathbf{1}$ and $p_A \hat{\mu}_Y = \mathbf{1}' \mathbf{A}^{-1} \mathbf{Y}$.

□

5.2.3 Link to other fast methods

Our algorithm is based on a tree traversal similar to that used for independent contrasts, which are computed in linear time for each trait separately (Felsenstein, 1973; Freckleton, 2012). These authors use both the unnormalized contrasts and the contrast variances to calculate the trait likelihood. Our algorithm for $\hat{\mu}_Y$ and its precision p parallels the calculation of independent contrasts. The Gaussian elimination method by FitzJohn (2012) is also a pruning algorithm, specifically designed to calculate the likelihood under BM and OU models. Our algorithm calculates key terms separately using a rank-one perturbation of the covariance matrix, for which the 3-point condition is required. The advantage of this method is its application to a broad class of models including non-Gaussian models, and its application to methods using some components separately like phylogenetic PCA or phylogenetic prediction (as detailed later).

5.2.4 Implementation and computing gain

The algorithm was implemented in the R package `phylolm`, with a function to perform phylogenetic regression for quantitative responses under various models for the residual error: BM, BM with trend if the tree is not ultrametric, Pagel's λ , κ and δ models, and OU models with constant selection strength α and variance rate σ^2 . These covariance parameters are estimated jointly with coefficient parameters in the linear model using maximum likelihood. The linear model can combine quantitative and categorical predictors. The package is available at <http://cran.r-project.org/web/packages/phylolm/>.

We quantified the gain in computing time for 10 to 4507 taxa. The tree used for simulation was obtained by randomly sampling the desired number of taxa from the 4507-species ultrametric tree in Bininda-Emonds et al. (2007), rescaled to a total tree height of 1. For various models of phylogenetic trait evolution, 100 data sets were simulated with the following parameters: ancestral state $\mu = 0$, variance rate $\sigma^2 = 1$, optimal value of 1 and $\alpha = 1$ for the OU model, $\lambda = 0.5$, $\kappa = 0.5$, $\delta = 0.5$, or $r = -1$ for the EB model. The parameters were then estimated

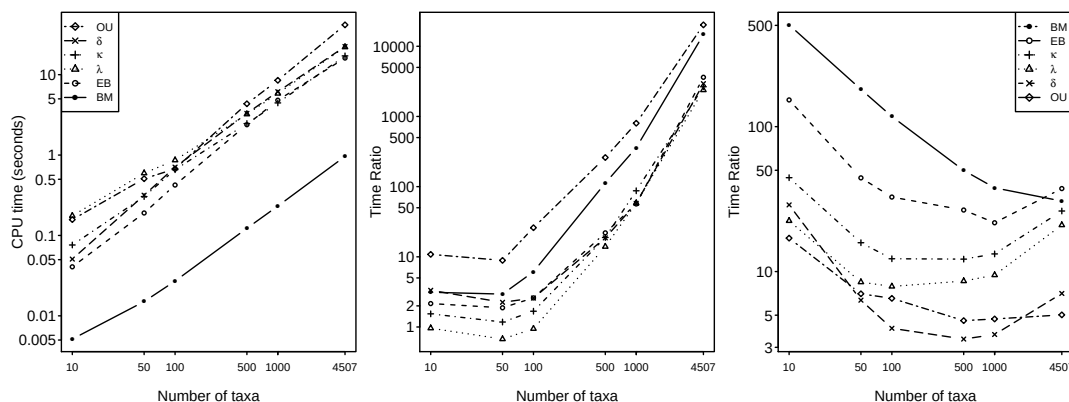


Figure 5.2: Running time for fitting phylogenetic regression models with maximum likelihood under BM, early-burst (EB), Pagel's λ , κ and δ models, and the OU model. Left: average computing time of the proposed algorithm as implemented in `phylolm` showing a linear increase with tree size. Middle: Ratio of average computing time using `fitContinuous` v1.3-1 to `phylolm`'s average computing time. The use of matrix inversion causes a time requirement that explodes with tree size. Right: Ratio of average computing time using `fitContinuous` v1.99-3 to `phylolm`'s average computing time. The use of FitzJohn's pruning algorithm restores a linear computing time with increasing tree size.

on 3.0 GHz processing units using 3 implementations: our fast algorithm in `phylolm`, matrix inversion as used in `fitContinuous` in package `geiger` version 1.3-1, and the newer version of `fitContinuous` in `geiger` version 1.99-3, which uses FitzJohn's fast pruning method (2012). Estimation with `fitContinuous` v1.3-1 was performed on fewer than 100 replicates for 1000 or more taxa, due to very long computing times (3 on 4507 taxa for models other than BM, 10 otherwise). The average computing time using `phylolm` is displayed in figure 5.2 (left), showing a linear increase with the number of taxa. Compared to `phylolm`, `fitContinuous` v1.3-1 is much slower (figure 5.2, center) due to the use of matrix inversion. For instance, our fast algorithm provides over a 100-fold speed increase on 500 taxa under BM, and over a 14,000-fold speed increase for the OU model on 4507 taxa. Freckleton (2012) reported similar computing gains (although for different implementations) with his linear-time method being 300

to 900 times faster than the direct matrix inversion method on a tree with 1000 tips. Thanks to its use of FitzJohn's pruning algorithm, `fitContinuous` v1.99-3 does not suffer the exploding computing burden with increasing tree size (figure 5.2, right). Its computing time was no more than 50 times that of `phylolm` on the 4507-taxon tree. `Phylolm` still showed a computing gain in most situations. Note that `phylolm` incurs an overhead compared to `fitContinuous` for calculating the covariance and standard errors of coefficients for instance. Differences in the optimization procedures used by both functions might also partly explain differences in computing time.

To assess the computing gain for a non-Gaussian evolutionary model, phylogenetic logistic regression was performed on the 3185-taxon tree used by Venditti et al. (2011). A simulation was inspired by Ives and Garland (2010), who explained antipredator behavior (either flee-fight or hide) by the species' body size and group size. We simulated antipredator behavior of 3185 mammals according to their standardized log-transformed body sizes (from Venditti et al., 2011) using the binary trait model and parameters estimated in Ives and Garland (2010): $\alpha = 0.5$ (phylogenetic signal), $b_0 = -0.82$ (intercept) and $b_1 = 0.096$ (body size effect). This simulated data set was analyzed on a 2.2 GHz processing unit using our algorithm implemented in `phyloglm` in our R package `phylolm`, as well as a similar R function using matrix inversion instead. Estimation with `phyloglm` took 11.75 minutes, more than 70 times faster than when using matrix inversion, which took 862.8 minutes. The estimated coefficients were close to the true values ($\hat{b}_0 = -0.80$ and $\hat{b}_1 = 0.07$ with `phyloglm` and $\hat{b}_0 = -0.79$ and $\hat{b}_1 = 0.08$ with matrix inversion). Phylogenetic signal was better estimated by `phyloglm` ($\hat{\alpha} = 0.32$) while matrix inversion led to a convergence failure (due to $\hat{\alpha} = 0.006$ being close to 0). Ives and Garland (2010) recommend running 2000 bootstrap replicates to obtain confidence intervals on the estimated parameters, which becomes possible with the computing gain offered by our algorithm. Note also that further computing gain could be achieved by coding the algorithm in C internally, as our current implementation uses native R code in version 2.0 of `phylolm`.

5.3 The (generalized) 3-point structure

Definition: A matrix V has a 3-point structure if it is symmetric, with non-negative entries ($V_{ij} \geq 0$) and satisfies the following 3-point condition: for any i, j, k (not necessarily distinct), the two smallest of V_{ij} , V_{ik} and V_{jk} are equal.

These conditions are easy to check, by going over each triple i, j, k of indices and comparing the values V_{ij} , V_{ik} , V_{jk} . When $i = k$, the condition implies that $V_{ij} \leq V_{ii}$, so that V can be interpreted as measuring similarities. Note that our 3-point condition differs from the ultrametric condition in [Semple and Steel \(2003\)](#) in that it is defined for similarity maps instead of dissimilarity maps, and is required to hold even when 2 of the 3 indices (i, j, k) are identical. As we show below, the 3-point structure implies a tree structure:

Theorem: V is 3-point structured if and only if it is the covariance matrix of a random variable at the tips of *some* rooted tree under a BM model. (see figure 5.3).

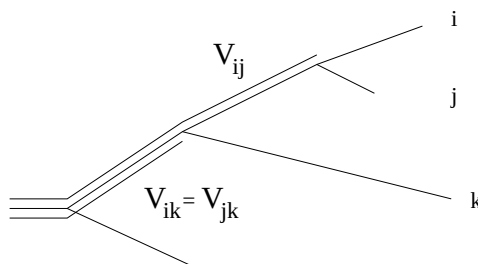


Figure 5.3: A matrix V is 3-point structured if it is symmetric with $V_{ij} \geq 0$ and if it satisfies the 3-point condition: for any i, j and k , the two smallest of V_{ij} , V_{ik} and V_{jk} are equal. Note that this condition can be verified without knowledge of the underlying tree.

Proof. The covariance matrix of a random variable under BM evolution is obviously symmetric with non-negative entries and satisfies the 3-point condition. However, the reverse is far from obvious. Our proof of the reverse shares some common ideas with the proof of the fundamental theorem identifying a

non-rooted tree from the 4-point condition in [Semple and Steel \(2003, theorem 7.2.6\)](#). Consider a symmetric matrix $\mathbf{V}$ satisfying the 3-point condition, with non-negative entries. Note that $\mathbf{V}$ is not assumed to be a covariance matrix. We define a dissimilarity map $\delta_{i,j} = -2V_{i,j}$ if $i \neq j$ and $\delta_{i,i} = 0$ for all i . δ has many negative terms, thus does not immediately identify a tree or positive edge weights. However the 3-point condition for $\mathbf{V}$ ensures that δ is an ultrametric as defined in [Semple and Steel \(2003, p.149\)](#). So there exists an equidistant representation of δ (theorems 7.2.5 and 7.2.8). In other words, there exists a rooted tree and a set of edge weights such that $\delta_{i,j}$ is the sum of edge weights along the path from i to j . Also, all internal edge weights are positive and the distance d_0 from the root to every tip is the same. We extend the tree with a root edge of weight $-d_0$ and add $V_{i,i}$ to the weight of the external edge connected to i . Then all edge weights of the new tree are positive and the covariance matrix of a BM process evolving along this new tree is exactly $\mathbf{V}$. $\square$

Note that our algorithm is readily applicable to matrices of the form $\mathbf{V} = \mathbf{D}_1 \tilde{\mathbf{V}} \mathbf{D}_2$, when $\tilde{\mathbf{V}}$ is 3-point structured and $\mathbf{D}_1, \mathbf{D}_2$ are diagonal matrices with non-zero diagonal elements. Indeed, the quadratic term $\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$ can be obtained with our algorithm using $\tilde{\mathbf{V}}$ and the transformed data $\mathbf{D}_2^{-1}\mathbf{X}$ and $\mathbf{D}_1^{-1}\mathbf{Y}$, which are straightforward to calculate when $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal. The determinant is also easy to calculate from $|\mathbf{V}| = |\mathbf{D}_1| |\tilde{\mathbf{V}}| |\mathbf{D}_2|$. We call such matrices $\mathbf{V}$ *generalized 3-point structured*. Also, $\mathbf{V} + \mathbf{D}$ is again generalized 3-point structured whenever $\mathbf{V}$ is and for every non-negative diagonal matrix $\mathbf{D}$. This fact can be used when $\mathbf{D}$ represents measurement error or variation among individuals of the same species, for instance.

To check whether a matrix $\mathbf{V}$ satisfies the 3-point condition and leads to a particular tree topology T , we need to check that for any 2 tips i and j , $V_{ij} \leq V_{ii}$. If $\mathbf{V}$ is known to be a covariance matrix, then this is automatically satisfied. Next, we also need to check that for any 3 distinct tips, if i and j are more closely related to each other than to k on T , then $V_{ik} = V_{jk} \leq V_{ij}$. The equality $V_{ik} = V_{jk}$ will automatically hold if V_{ab} can generally be expressed as a feature of the

last common ancestor of a and b , because the pairs i, k and j, k share the same common ancestor. In what follows, we use this criterion to check the 3-point condition for models arising from a particular rooted tree topology.

5.3.1 Weak 3-point condition and negative correlations

Our algorithm does not strictly require that all branch lengths (ts) are non-negative on the rooted tree identified from $\mathbf{V}$ by the Theorem. It only requires non-negative branch lengths at the tips, and $1 + tp_A > 0$ at each step of the algorithm. Therefore, our algorithm would apply to a broader class of covariance matrices than to just 3-point structured matrices, with some potentially negative correlations. This class of models would still require the following weak 3-point condition: for each triple i, j, k , at least two of V_{ij} , V_{jk} and V_{ki} are equal.

One example is the PGLS method proposed by [Martins and Hansen \(1997\)](#). In their models the error variance depends on the mean, using the estimated mean μ_a at the most recent common ancestor of tips i and j . For instance, they propose $V_{ij} = \mu_a(1 - \mu_a) \exp(-\alpha t_{ij})$ for a binary response. This formula ensures that $V_{ik} = V_{jk}$ whenever i and j are more closely related to each other than to k , so the weak 3-point condition is satisfied. The shared value $V_{ik} = V_{jk}$ might not necessarily be smaller than V_{ij} , however.

For a dissimilarity matrix, the weak 4-point condition identifies a unique unrooted tree with possibly negative branch lengths ([Semple and Steel, 2003](#)). However, the weak 3-point condition on similarity matrices does not necessarily identify a unique rooted tree. As a counter-example, the similarity matrix below

$$\mathbf{V} = \begin{pmatrix} 1 & a & a & b \\ a & 1 & b & a \\ a & b & 1 & b \\ b & a & b & 1 \end{pmatrix}$$

satisfies the weak 3-point condition for any $a, b \geq 0$. However, if $a \neq b$, we can

see that V does not correspond to any of the 15 possible 4-taxon rooted trees (with possibly negative branch lengths), in the sense that it cannot be written as a matrix of shared path lengths $(t_{ij})_{i,j}$.

5.4 Gaussian trait models with the 3-point structure

5.4.1 Modified Brownian motion models

Evolution models under BM retain their 3-point structure when branch lengths are transformed. Typically, branch lengths are first inferred as divergence times based on molecular data, but can be adaptively stretched or shrunk to better fit the trait data. The goal of these transformations is to use branch lengths as parameters for the phylogenetic correlation between species, as expected from their shared history. Such transformations include Pagel's lambda ([Pagel, 1999](#)) to account for a mixture of BM and independent evolution, which is equivalent to the phylogenetic mixed model ([Lynch, 1991](#)). All internal branches are multiplied by λ in $[0, 1]$ while the tips are maintained at their original distance from the root. Also included is the kappa transformation ([Pagel, 1997](#)) where all branch lengths are raised to the power κ in $[0, 1]$, the delta transformation ([Pagel, 1999](#)) where the distance from the root to all nodes is raised to the power $\delta > 0$, the similar ρ transformation by [Grafen \(1989\)](#), and the acceleration/deceleration (ACDC) model ([Blomberg et al., 2003](#)). Models with BM evolution and a variable rate of variance accumulation also belong to the class of 3-point structured models, including models with a number of different rates $\sigma_1^2 \dots \sigma_k^2$ in different parts of the tree ([O'Meara et al., 2006](#); [Thomas et al., 2006](#); [Revell and Harmon, 2008](#); [Venditti et al., 2011](#); [Eastman et al., 2011](#); [Revell et al., 2012](#)), or models with a continuously varying rate along the tree such as the early-burst (EB) model ([Harmon et al., 2010](#)), which is equivalent to the ACDC transformation. The 3-point structure is maintained under a directional trend (estimable when taxa do not all have the same age), because the covariance matrix under BM evolution depends on the transformed branch lengths only, not the mean value.

5.4.2 Multivariate response

If a multivariate response $\mathbf{Y}$ is considered with m different quantitative traits, the most common model of evolution assumes that the error structure has both a BM-modified phylogenetic covariance $\mathbf{V}$ as considered above, and a between-trait covariance structure $\mathbf{C}$. In other words, the covariance between traits k and l at taxa i and j is taken to be $\text{cov}(Y_{i,k}, Y_{j,l}) = C_{kl}V_{ij}$. The full covariance matrix of the vectorized $Y_{k,i}$ values is then the Kronecker product $\mathbf{C} \otimes \mathbf{V}$. This matrix is not 3-point structured, but its inverse is simply $\mathbf{C}^{-1} \otimes \mathbf{V}^{-1}$ and its determinant is $|\mathbf{C}|^n |\mathbf{V}|^m$. With a small number of traits m , $\mathbf{C}^{-1}$ and $|\mathbf{C}|$ are easy to calculate, and our algorithm can be used for the parts involving $\mathbf{V}^{-1}$ and $|\mathbf{V}|$. More complicated models have been proposed to allow for different phylogenetic signals for different traits. [Freckleton \(2012\)](#) considered Pagel's λ transformation with parameter λ_k for trait k , then using $\text{cov}(Y_{i,k}, Y_{j,l}) = C_{kl}V_{ij}(\sqrt{\lambda_k \lambda_l})$ and independent contrasts to do the computations in linear time. Here again, the covariance matrix of the full response data is not 3-point structured, but our method can still be used for the particular model above, because each matrix $\mathbf{V}(\sqrt{\lambda_k \lambda_l})$ is proportional to the same 3-point structured $\mathbf{V}(1)$ up to its diagonal (in [Appendix C](#)).

5.4.3 Within-species variation

Individual values can be used at the tips instead of species averages as in [Ives et al. \(2007\)](#); [Felsenstein \(2008\)](#) or [Revell and Reynolds \(2012\)](#). This can greatly contribute to increasing tree size and computational burden. Each species is then represented by a star tree of the k individuals from that species. For a univariate response, the branches of this star tree are given length t_{within} to represent and estimate the within-species variance. At this star tree, $p_A = k/t_{\text{within}}$ is the precision of the species average at the root of the star tree, $p = (t + t_{\text{within}}/k)^{-1}$ is the precision at the root of the parent edge, $w_s = 1/k$ equally weights the k individuals and $\hat{\mu}_Y = \bar{y}$ is the usual species average. For a multivariate response with m different traits in $\mathbf{Y}$, our algorithm can be applied with no modification if the within-species phenotypic covariance between traits $\mathbf{C}_w$ is assumed to

be the same as the between-species phylogenetic covariance between traits $\mathbf{C}_b$. In that case, the full covariance matrix for all the Y values stacked as a single vector is $\mathbf{C}_b \otimes \mathbf{V}$ and easy to handle. In most cases however, the assumption of equal within-species and between-species trait correlations is unreasonable. The general model considered by [Ives et al. \(2007\)](#) and [Felsenstein \(2008\)](#) leads to the covariance matrix $\mathbf{C}_b \otimes \mathbf{V} + \mathbf{C}_w \otimes \mathbf{I}$ for the multivariate response $\mathbf{Y}$. This matrix is not 3-point structured, but again our method can be adapted to recursively calculate the key terms involving its inverse and its determinant (Appendix [C](#)).

5.4.4 Ornstein-Uhlenbeck models

The covariance matrix for the univariate OU model also has a 3-point structure provided that the underlying tree is ultrametric. Indeed, it is given by

$$V_{ij} = \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t_{ij}})e^{-\alpha d_{ij}} \quad \text{or} \quad V_{ij} = \frac{\sigma^2}{2\alpha}e^{-\alpha d_{ij}}$$

depending on whether the value at the root is considered to be a fixed parameter or is integrated out as a random variable with the stationary distribution ([Hansen, 1997](#); [Ho and Ané, 2013a](#)). Here t_{ij} is the time of shared ancestry between taxa i and j and d_{ij} is the tree distance between the two taxa. The part involving t_{ij} satisfies the 3-point condition but the part with d_{ij} does not necessarily. It does if the tree is ultrametric because then d_{ij} only depends on the age of the last common ancestor of i and j , as we can write $d_{ij} = 2(T - t_{ij})$ where T is the tree height.

The more general model of adaptive evolution considered by [Butler and King \(2004\)](#) includes variable optima, to allow for different constraints on different lineages. The 3-point structure is retained under this model because changes in the selection optimum do not affect the covariance matrix.

When the tree is not ultrametric, the OU covariance matrix has a generalized 3-point structure. To see this, consider a modified tree where the external branch to taxon i is extended (or shortened) by u_i , in such a way that the modified tree

is ultrametric. The modified covariance matrix $\tilde{\mathbf{V}}$ is then 3-point structured. It is a simple modification of the original covariance matrix: $\mathbf{V} = \mathbf{D}_u \tilde{\mathbf{V}} \mathbf{D}_u$ where $\mathbf{D}_u$ has entries $e^{\alpha u_i}$ on the diagonal and zeros elsewhere. Therefore, the covariance of the OU model on a non-ultrametric tree is indeed generalized 3-point structured. [Beaulieu et al. \(2012\)](#) further extended this adaptive evolution model by allowing different variance rates and/or different selection strengths on different parts of the tree. The covariance matrix under this model also has a generalized 3-point structure. To see this, we consider ‘epochs’ or selection regimes, where epochs are separated by speciation events or changes in the selection regime. [Beaulieu et al. \(2012\)](#) showed that the covariance matrix is then

$$V_{ij} = \exp \left(- \sum_{\gamma=1}^{\kappa(i)} \alpha_{i,\gamma} (s_{i,\gamma} - s_{i,\gamma-1}) - \sum_{\gamma=1}^{\kappa(j)} \alpha_{j,\gamma} (s_{j,\gamma} - s_{j,\gamma-1}) \right) \\ \times \left(\sum_{\gamma=1}^{\kappa(i,j)} \frac{\sigma_{ij,\gamma}^2}{2\alpha_{ij,\gamma}} (e^{2\alpha_{ij,\gamma} s_{ij,\gamma}} - e^{2\alpha_{ij,\gamma} s_{ij,\gamma-1}}) \right)$$

where the s values are transition times between successive epochs going forward in time with $s = 0$ at the root of the tree, $\kappa(i)$ is the number of epochs on the path from the root to taxon i , $\kappa(i,j)$ is the number of epochs on the path from taxon i to taxon j , and the α and σ^2 values are the selection strengths and variance rates during each epoch. If the selection strength is constant then the first term simplifies to $e^{-\alpha t_{ii} - \alpha t_{jj}} = e^{-2\alpha t_{ij} - \alpha d_{ij}}$. If the variance rate is also constant throughout the tree, then the second term simplifies to $\frac{\sigma^2}{2\alpha} (e^{2\alpha t_{ij}} - 1)$, so that the formula above generalizes the one-regime covariance matrix when the state at the root is a fixed parameter of the model. We can rewrite $\mathbf{V} = \mathbf{D} \tilde{\mathbf{V}} \mathbf{D}$ where $\mathbf{D}$ has diagonal element $\exp \left(- \sum_{\gamma=1}^{\kappa(i)} \alpha_{i,\gamma} (s_{i,\gamma} - s_{i,\gamma-1}) \right)$ for taxon i , and

$$\tilde{V}_{ij} = \sum_{\gamma=1}^{\kappa(i,j)} \frac{\sigma_{ij,\gamma}^2}{2\alpha_{ij,\gamma}} (e^{2\alpha_{ij,\gamma} s_{ij,\gamma}} - e^{2\alpha_{ij,\gamma} s_{ij,\gamma-1}}),$$

which only depends on the regimes affecting the root to the last common ancestor of i and j . Then $\tilde{\mathbf{V}}$ is 3-point structured, so $\mathbf{V}$ is generalized 3-point structured and our algorithm can be applied to the OU model with variable selection parameters.

More recently, the multivariate OU model has been used to analyze a response $\mathbf{Y}$ with m traits. For this model, the selection strength α and the variance accumulation rate σ^2 are now $m \times m$ matrices. The covariance between the m traits at tips i and j is then (Bartoszek et al., 2012; King and Butler, 2009)

$$\text{cov}(\mathbf{Y}_{i\cdot}, \mathbf{Y}_{j\cdot}) = e^{-\alpha t_{ii}} \left(\int_0^{t_{ij}} e^{\alpha v} \sigma^2 e^{\alpha v} dv \right) e^{-\alpha t_{jj}} = e^{-\alpha t_{ii}} (e^{\alpha t_{ij}} \gamma e^{\alpha t_{ij}} - \gamma) e^{-\alpha t_{jj}}$$

where the asymptotic covariance γ can be found by solving $\gamma \alpha' + \alpha \gamma = \sigma^2$. If selection acts independently on all traits, i.e. if α is diagonal (which can be obtained by pre-transforming the traits using the eigenvectors of α), then the phylogenetic covariance between taxa is generalized 3-point structured for each pair of traits k and ℓ : $\text{cov}(Y_{i,k}, Y_{j,\ell}) = \sigma_{k\ell}^2 e^{-\alpha_k t_{ii} - \alpha_\ell t_{jj}} (e^{(\alpha_k + \alpha_\ell) t_{ij}} - 1) / (\alpha_k + \alpha_\ell)$. If the traits are under different selection strengths α_k , then different 3-point structured matrices apply to different traits and there is no obvious adaptation of our method to speed up the calculations. Under the special case that $\alpha_k = \alpha$ are all equal, then the same phylogenetic signal applies to all traits. The covariance of the full vectorized $\mathbf{Y}$ is then $\sigma^2 \otimes \mathbf{V}_\alpha$ with the same phylogenetic covariance $\mathbf{V}_\alpha$ as in the univariate case. So again, terms involving its inverse and determinant can be calculated efficiently in that case.

5.5 Applications to non-Gaussian models

Our algorithm can be applied to non-Gaussian traits, as long as the computational bottleneck lies in the calculation of quadratic forms involving the inverse of a generalized 3-point structured covariance matrix, and/or its determinant. It is the case for phylogenetic logistic regression, phylogenetic Poisson regression and phylogenetic generalized linear mixed models, as we explain in this section.

5.5.1 Phylogenetic logistic regression

[Ives and Garland \(2010\)](#) proposed statistical methods for phylogenetic logistic regression when the trait of interest is binary. For instance, they showed that group size influences antipredator behavior in African antelopes, with solitary species being more likely to hide from predators ($y = 0$) than to flee or fight ($y = 1$). Their method accounted for phylogenetic correlation and a possible body size effect. In their model, the binary trait evolves according to a two-step process, leading to both phylogenetic correlation and correlation with predictors. In the first step, the binary trait is assumed to evolve along a phylogenetic tree according to a two-state continuous time Markov process with transition rate α and asymptotic probability $\bar{\mu}$ of being in state 1. In the second step, the predictors $\mathbf{X}$ influence the binary trait values $\mathbf{Y}$ at each tip in such a way that the expectation of $\mathbf{Y}$ satisfies

$$\mu = \frac{\exp(\mathbf{X}\beta)}{1 + \exp(\mathbf{X}\beta)}$$

(details in [Ives and Garland, 2010](#)). Under this model, the covariance of $\mathbf{Y}$ is

$$V_{ij} = \begin{cases} \mu_i(1 - \mu_i) & \text{if } i = j \\ s_i s_j \exp(-\alpha d_{ij}) & \text{if } i \neq j \end{cases}$$

where d_{ij} is the tree distance between tips i and j and

$$s_i = \begin{cases} \mu_i \sqrt{1 - \bar{\mu}} / \sqrt{\bar{\mu}} & \text{if } \mu_i < \bar{\mu} \\ (1 - \mu_i) \sqrt{\bar{\mu}} / \sqrt{1 - \bar{\mu}} & \text{if } \mu_i > \bar{\mu}. \end{cases}$$

This covariance matrix is generalized 3-point structured (using the s_i values to form the diagonal matrix $\mathbf{D}$). The parameters are estimated by an iterative procedure. Given the phylogenetic correlation α , $\hat{\beta}$ is updated using quasi-log-likelihood by solving the following equation

$$\mathbf{U}(\hat{\beta}(\alpha)|\alpha) = (\mathbf{A}\mathbf{X})'\mathbf{V}(\alpha)^{-1}(\mathbf{Y} - \boldsymbol{\mu}) = 0 \quad (5.2)$$

where $\mathbf{A}$ is the diagonal matrix with diagonal terms $\mu_i(1 - \mu_i)$. To reduce bias, the penalized quasi-log-likelihood score adds a term involving the Fisher information matrix $\mathbf{I}(\beta)$:

$$\mathbf{U}_j^*(\hat{\beta}(\alpha)|\alpha) = \mathbf{U}_j(\hat{\beta}(\alpha)|\alpha) + 1/2 \text{ trace}\{\mathbf{I}(\beta)^{-1}[\partial\mathbf{I}(\beta)/\partial\beta_j]\} = 0 \text{ for all } j.$$

where $\mathbf{I}(\beta) = (\mathbf{A}\mathbf{X})'\mathbf{V}(\alpha)^{-1}(\mathbf{A}\mathbf{X})$. Here again, $\mathbf{U}(\hat{\beta}(\alpha)|\alpha)$ and $\mathbf{I}(\beta)$ take the form of a quadratic term that can be calculated efficiently. Based on the new estimated value $\hat{\beta}$ (and thus new estimates of $\bar{\mu}$ and μ), $\hat{\alpha}$ is updated by minimizing

$$SS(\hat{\alpha}(\mu)|\mu) = -\frac{1}{2}(\log|\mathbf{V}(\alpha)| + (\mathbf{Y} - \mu)'\mathbf{V}(\alpha)^{-1}(\mathbf{Y} - \mu)),$$

which contains both the determinant of the covariance matrix and a quadratic term. Finally, the estimated standard error of coefficients $\hat{\beta}$ given $\hat{\alpha}$, $\hat{\sigma}^2(\hat{\beta}|\hat{\alpha}) = \mathbf{I}(\beta)^{-1}$, is again evaluated using a quadratic product. Therefore, all these steps can be carried out efficiently using our algorithm.

5.5.2 Phylogenetic logistic mixed models

Phylogenetic mixed models use a regression formula that involves a random effect with phylogenetic correlation. For example, [Ives and Helmus \(2011\)](#) investigate the presence/absence of n species at m sites to determine whether closely related species are more likely to co-occur at the same sites, and to quantify the phylogenetic signal in species co-occurrence. The observations $\mathbf{Y}$ form a long nm vector of all binary outcomes, for each species at each site. The authors consider the following phylogenetic logistic mixed model with binomial distribution $\mathbb{P}[Y_i = 1] = \mu_i$ and

$$\log \frac{\mu_i}{1 - \mu_i} = \alpha_{\text{spp}[i]} + \mathbf{b}_i + \mathbf{c}_{\text{site}[i]}$$

$$\mathbf{b} \sim \mathcal{N}(0, \mathbf{I}_m \otimes \sigma_{\text{spp}}^2 \Sigma_{\text{spp}}), \quad \mathbf{c} \sim \mathcal{N}(0, \sigma_{\text{site}}^2 \mathbf{I}_m)$$

where $\alpha_{\text{spp}[i]}$ is an unobserved (fixed) species effect and the random effects b_i carry the phylogenetic correlation. The phylogenetic covariance matrix among species Σ_{spp} is derived either from a BM model or an OU model and hence is generalized 3-point structured. The full covariance matrix of the random effects is then

$$\mathbf{C} = \mathbf{I}_m \otimes (\sigma_{\text{spp}}^2 \Sigma_{\text{spp}} + \sigma_{\text{site}}^2 \mathbf{I}_n).$$

To estimate the parameters using approximate maximum likelihood, [Ives and Helmus \(2011\)](#) iteratively update α , $\mathbf{b}$ and $\mathbf{c}$ given the current estimates of σ_{spp}^2 and σ_{site}^2 using penalized quasi-likelihood (equations (B.3) and (B.4) in [Ives and Helmus, 2011](#)). This step requires the calculation of quadratic terms, where the working covariance matrix is $\mathbf{V} = \mathbf{C} + \mathbf{W}^{-1}$ and $\mathbf{W}$ is the $nm \times nm$ diagonal matrix with diagonal elements $\mu_i(1 - \mu_i)$. Our algorithm is applicable to $\mathbf{V}$ because it is block diagonal, and each block has a generalized 3-point structure. Second, the estimates of σ_{spp}^2 and σ_{site}^2 are updated based on the new estimated values of α , $\mathbf{b}$ and $\mathbf{c}$ using maximum likelihood (ML) or restricted maximum likelihood (REML). For this, the log-likelihood function being optimized (eq. (B.5) in [Ives and Helmus, 2011](#)) involves both the determinant of $\mathbf{V}$ and quadratic terms, which can again be calculated in linear time using our algorithm. This likelihood criterion is further used to build confidence intervals for the variances σ_{spp}^2 and σ_{site}^2 , one further step that can benefit from computational efficiency.

5.5.3 Phylogenetic Poisson regression

[Lavergne et al. \(2013\)](#) used phylogenetic Poisson regression to investigate whether the local diversification rate of angiosperm families in the western Mediterranean basin was affected by the family's migration rate across the strait of Gibraltar. To do so, the number $\mathbf{Y}$ of local endemic species per family was explained by the family migration rate $\mathbf{X}$ using the phylogenetic GEE method ([Paradis and Claude, 2002](#)), assuming that the counts $\mathbf{Y}$ follow a Poisson distribution with mean $\mu = \exp(\mathbf{X}\beta)$ and covariance $\mathbf{V} = \phi \mathbf{A}^{1/2} \mathbf{C} \mathbf{A}^{1/2}$ across species. Here ϕ is a dispersion parameter, $\mathbf{A}$ is a diagonal matrix with entries μ on the diagonal,

and $\mathbf{C}$ is a phylogenetic correlation matrix. The coefficients β can be estimated from the generalized estimating equations (similar to equation (5.2) for logistic regression)

$$(\mathbf{AX})'\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) = 0.$$

These equations can be solved through an iterative process in which β is updated at each step t by:

$$\beta_{t+1} = \beta_t - \mathbf{I}(\beta)^{-1}[(\mathbf{AX})'\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu})]$$

where $\mathbf{I}(\beta) = (\mathbf{AX})'\mathbf{V}(\alpha)^{-1}(\mathbf{AX})$. Finally, the standard error of $\hat{\beta}$ can be estimated from $\hat{\sigma}^2(\hat{\beta}) = \mathbf{I}(\beta)^{-1}$. Because the phylogenetic correlation $\mathbf{C}$ is derived from a BM model in [Paradis and Claude \(2002\)](#), $\mathbf{V}$ is generalized 3-point structured and the quadratic terms above can be calculated efficiently by our algorithm.

[Hadfield and Nakagawa \(2010\)](#) described a somewhat different model (in the class of phylogenetic generalized linear mixed models) where the response trait follows a Poisson distribution with mean $\mu = \exp(\mathbf{X}\beta + \mathbf{b} + \mathbf{e})$. Here taxon i has random effects b_i and e_i , where $\mathbf{b}$ is assumed to follow a Brownian motion along the tree and $\mathbf{e}$ is an independent Gaussian noise. This model is equivalent to Pagel's lambda model for the random effects $\mathbf{b} + \mathbf{e}$. [Hadfield \(2010\)](#) uses MCMC for this model in a Bayesian framework. To avoid numerical matrix inversion, the state space is increased to track all random effects at the $n - 2$ non-root internal nodes. Our algorithm could also be used for this model, except that [Hadfield \(2010\)](#) makes further use of the sparsity of $\mathbf{V}^{-1}$ when this matrix is expanded to include internal nodes, but $\mathbf{V}^{-1}$ is no longer sparse when restricted to the tips of the tree. Therefore our algorithm might not bring much computational gain in this MCMC implementation. However, computational benefits would be substantial in a frequentist approach, using approximate ML as in [Ives and Helmus \(2011\)](#). A more general model than that considered by [Hadfield \(2010\)](#) would allow the random effect to follow any evolution trait model, such as OU models or variable rate BM models. For a Poisson-distributed trait $\mathbf{Y}$, the mean

μ_i at taxon i is assumed to satisfy

$$\mu = \exp(\mathbf{X}\beta + \mathbf{b})$$

where $\mathbf{b} \sim \mathcal{N}(0, \mathbf{V})$ and $\mathbf{V}$ is any generalized 3-point structured phylogenetic covariance matrix. Penalized quasi-likelihood estimation and REML estimation can thus be combined iteratively, as done in [Ives and Helmus \(2011\)](#). For a Poisson-distributed response, the working covariance matrix is $\mathbf{V} + \mathbf{W}^{-1}$ with $\mathbf{W}$ diagonal and $W_{ii} = \mu_i$ ([Liang, 2007](#)). This matrix can be handled by the algorithm presented here because it is still generalized 3-point structured.

5.6 Further Applications

The components we derive are the key building blocks for other tools, such as fitting models with REML, reduced major axis (RMA) regression ([Ives et al., 2007](#); [Hansen and Bartoszek, 2012](#)), phylogenetic PCA or phylogenetic discriminant analysis as in [Motani and Schmitz \(2011\)](#). For instance, the PCA approach by [Revell \(2009\)](#) is based on the phylogenetically corrected means $\hat{\mu}_X$ and the evolutionary covariance among traits $\mathbf{C} = (\mathbf{X} - \mathbf{1}'\hat{\mu}_X)' \mathbf{V}^{-1} (\mathbf{X} - \mathbf{1}'\hat{\mu}_X)$. Here $\mathbf{X}$ is an $n \times m$ matrix for multivariate data across n taxa and m traits. If the underlying phylogenetic covariance matrix is generalized 3-point structured, then the algorithm presented here can be used in these applications as well.

Prediction intervals were used in [Garland and Ives \(2000\)](#) to determine if certain species deviate significantly from an allometric prediction. More recently, there has been renewed interest in phylogenetic prediction (e.g. [Cai et al., 2011](#)) to predict traits of interest across wild species to help plant breeding of cultivated species. Let s_0 be a new taxon that can be placed on the current tree. Given the phylogenetic covariance parameter θ , the following prediction of the response trait $Y(s_0)$ at the new taxon is the best linear unbiased prediction ([Goldberger, 1962](#)):

$$\hat{Y}(s_0) = \mathbf{c}'_{\theta} \mathbf{V}_{\theta}^{-1} \mathbf{Y} + (\mathbf{x} - \mathbf{X}' \mathbf{V}_{\theta}^{-1} \mathbf{c}_{\theta})' \hat{\beta}$$

where $\mathbf{x}$ are the known covariate(s) at the new phylogenetic location, $\mathbf{c}_\theta$ is the vector of covariances between the new taxon and the sampled taxa, $\hat{\beta}$ contains the estimated regression coefficients from the sampled taxa with observed response data $\mathbf{Y}$ and covariates $\mathbf{X}$, and $\mathbf{V}_\theta$ is the covariance matrix among the sampled taxa as before. This is also called ‘universal kriging’ in spatial statistics (Cressie, 1993). Prediction intervals can then be obtained based on the prediction variance:

$$\begin{aligned} \mathbb{E}(\hat{Y}(s_0) - Y(s_0))^2 &= \mathbf{V}_\theta(s_0, s_0) - \mathbf{c}_\theta' \mathbf{V}_\theta^{-1} \mathbf{c}_\theta \\ &+ (\mathbf{x} - \mathbf{X}' \mathbf{V}_\theta^{-1} \mathbf{c}_\theta)' (\mathbf{X}' \mathbf{V}_\theta^{-1} \mathbf{X})^{-1} (\mathbf{x} - \mathbf{X}' \mathbf{V}_\theta^{-1} \mathbf{c}_\theta). \end{aligned}$$

All these quantities can be calculated easily with our algorithm for quadratic products, without the need to invert the large matrix $\mathbf{V}_\theta$. These quantities were already used in Garland and Ives (2000) for the case of simple linear regression, where the tree needed to be re-rooted for each new taxonomic position where prediction was desired. Our approach is general to any regression model, fast enough for very large trees, and does not require any re-rooting.

5.7 Discussion

5.7.1 Speed-up for large scale and iterative analyses

The utility of this algorithm comes in when the likelihood of trait data or other score functions need to be evaluated on large data sets or for many iterations. The size of data sets continues to increase as phylogenetic trees continue to become more and more accurate and more taxon-rich. For instance, the goal of the iPlant Tree of Life project (Goff et al., 2011) is to scale up tree inference and post-tree analyses, to hundreds of thousands of taxa. Trait analyses on such large trees are possible with the proposed algorithm. Also, the important sample size here is the number of terminals with trait data, which can be many times larger than the number of species when individual organisms are used instead of species averages (e.g. Revell and Reynolds, 2012).

The need for large numbers of iterations comes from various sources. First, iterations are needed to optimize model parameters and maximize the likelihood, or to search for the area of high posterior probability in MCMC algorithms. Such iterations are needed to optimize Pagel's λ or κ , the rate of variance decrease in an early-burst model, the selection strength α in an OU models, or the estimated ratio of within-species variance versus between-species variance. The number of iterations or MCMC generations that are needed can be substantial as models become very complex, such as with an unbounded number of rate shifts at unknown locations in the tree (Eastman et al., 2011) or an unknown number of selective regime changes at unknown locations Beaulieu et al. (2012). Hadfield (2010) noted that each MCMC step can be sped up considerably by tracking all random effects at non-root internal nodes, at the expense of more iterations for convergence. Second, many iterations are used when tree uncertainty is taken into account, either in an integrated Bayesian framework (as in BayesTraits, Pagel, 1999), or when repeating the analysis on a large sample of phylogenetic trees. Finally, bootstrap procedures are now increasingly used to test hypotheses or assess uncertainty in the resulting parameters (Garland et al., 1993; Freckleton and Harvey, 2006). For instance, Boettiger et al. (2012) advocate the use of parametric bootstrap where data are repeatedly sampled from a given trait evolution model and analyzed like the original data, to perform reliable model selection. Bootstrap methods are powerful but computationally heavy, and have much to gain from a fast likelihood calculation algorithm. Taken together, the number of iterations required to estimate the parameters of the trait evolution model, to integrate over tree uncertainty and to conduct bootstrap procedures can quickly increase the overall computing load.

5.7.2 Model adequacy on very large trees

At very large taxonomic scales, traditional homogeneous models of trait evolution are often likely to be simplistic. To account for changes in modes of trait evolution, complex models are increasingly needed to adequately analyze very large data sets.

To do so, fast algorithms become even more important, to deal both with large trees and models with large numbers of parameters. The computational gains presented here could foster the development of complex models appropriate for large taxonomic scales, which may require special estimation and model selection procedures.

6 DISCUSSION AND FUTURE DEVELOPMENTS

My thesis shows several inference difficulties for the Ornstein-Uhlenbeck (OU) tree models. Some of them can be solved while the others are impossible to overcome without fossil data. However, it does not mean that we should abandon OU models because OU models are best for analysis when the trait truly evolved according to an OU process. Thus, my thesis can inform investigators of situations where increasing present-day observations is not enough and fossil data are needed to test their hypothesis. One possible direction for future development is studying the impact of fossil data on OU models.

In Chapter 2 and 3, I only consider the simplest situation when the parameters μ , α , and γ are assumed to be constant. More work needs to be done in order to generalize theoretical results in this thesis to scenarios where these parameters vary across the tree. Measurement error can also be considered, and in this case, the covariance matrix becomes $\gamma \mathbf{V}_\alpha + \sigma_e^2 \mathbf{I}$, where σ_e^2 is the variance of the measurement error.

Chapter 4 illustrates the breakdown of AIC and BIC to detect changes in the selection optimum along the tree. This phenomenon is identified as a “large p, small n” problem. In the statistic literature, many methods have been developed to solve such problems for independent data. A future direction could be adjusting these methods for OU tree models.

In the future, I also want to develop criteria and methods that can select the best set of taxa for statistical inference under the OU models. To be specific, I consider the problem of selecting an optimal set of k species from an n -species tree for estimating μ and α . Finding such set for both μ and α may be infeasible. Another direction is to focus on each parameter separately.

A APPENDIX A

A.1 Spectral decomposition of the OU covariance matrix on symmetric trees

We consider here symmetric trees (figure 2.5) with m levels of internal nodes, the root being at level 1. Each node at level k is connected to $d_k \geq 2$ children by branches of length t_k . The age of nodes at level k is then $u_k = t_k + \dots + t_m$. Under the OU model (2.1), the correlation matrix V_α is identical to that obtained under a BM model along a tree with an extra branch extending from the root and with transformed branch lengths $\mathbf{t}^{\text{BM}}$:

$$t_k^{\text{BM}}(\alpha) = \begin{cases} 1 - e^{-2\alpha t_m} & \text{for } k = m, \\ e^{-2\alpha u_{k+1}} - e^{-2\alpha u_k} & 1 \leq k \leq m-1, \\ e^{-2\alpha u_1} & k = 0 \text{ (extra root branch)}. \end{cases}$$

Therefore, we can derive the eigen-decomposition of $V_\alpha(\mathbf{t}) = V_\alpha^{\text{BM}}(\mathbf{t}^{\text{BM}})$ as done in Ané (2008). The eigenvalues, from greatest to smallest, are

$$\lambda_k = n \sum_{i=k}^m \frac{t_i^{\text{BM}}(\alpha)}{d_1 \dots d_i} = \sum_{i=k}^m d_{i+1} \dots d_m (e^{-2\alpha u_{i+1}} - e^{-2\alpha u_i})$$

with multiplicity $d_1 \dots d_{k-1}(d_k - 1)$, for $k = 0, \dots, m$ and u_{m+1} set to 0 and u_0 to ∞ . Furthermore, Ané (2008) showed that the eigenvectors of V_α^{BM} are independent of the tree's branch lengths, which implies here that the eigenvectors of V_α are independent of α . Each eigenvector corresponding to $\lambda_k(\alpha)$ represents a contrast between the descendants of a node at level k . One exception is the eigenvector associated with the extra root branch and largest eigenvalue λ_0 . This eigenvector is $\mathbf{1}$ and has multiplicity 1.

A.2 Supporting lemmas and technical proofs

A.2.1 Procedures for choosing independent contrasts

Lemma A.1. *Let $\mathbb{T}$ be an ultrametric tree. For every $a < b$, we can choose a set of independent contrasts $\mathcal{C}$ with respect to some of the internal nodes in $\mathcal{J}_{(a,b)}^{\mathbb{T}}$ such that $|\mathcal{C}| \geq \frac{1}{2}|\mathcal{J}_{(a,b)}^{\mathbb{T}}|$.*

Proof. We choose contrasts as follows, starting with $\mathcal{C} = \emptyset$ and $\mathbb{T}_0 = \mathbb{T}$. At step n , we choose an internal node $i_n \in \mathcal{J}_{(a,b)}^{\mathbb{T}_{n-1}}$ of minimum age, and a path p_{i_n} connecting any two tips having i_n as their common ancestor. We update $\mathcal{C} = \mathcal{C} \cup \{C_{i_n}^{p_{i_n}}\}$ and obtain tree $\mathbb{T}_n$ from $\mathbb{T}_{n-1}$ by dropping all descendants of i_n . We stop when $\mathcal{J}_{(a,b)}^{\mathbb{T}_n} = \emptyset$. The procedure guarantees that the paths do not intersect, hence the contrasts are independent. Furthermore, $\mathcal{J}_{(a,b)}^{\mathbb{T}_n} = \mathcal{J}_{(a,b)}^{\mathbb{T}_{n-1}} \setminus \{i_n, i'_n\}$ where i'_n is the parent of i_n , so $|\mathcal{C}| \geq |\mathcal{J}_{(a,b)}^{\mathbb{T}}|/2$.

Lemma A.2. *Let $\mathbb{T}$ be an ultrametric tree of height T . For all $t \in [0, T]$,*

- (a) *There exists a set of independent contrasts $\mathcal{C}$ with respect to nodes in $\mathcal{J}_{[0,t]}^{\mathbb{T}}$ such that $\sum_{C \in \mathcal{C}} (T_C - t)^2 \geq \frac{1}{2} \sum_{i \in \mathcal{J}_{[0,t]}^{\mathbb{T}}} (T_i - t)^2$.*
- (b) *There exists a set of independent contrasts $\mathcal{C}$ with respect to nodes in $\mathcal{J}_{(t,\infty)}^{\mathbb{T}}$ such that $\sum_{C \in \mathcal{C}} (T_C - t)^2 \geq \frac{1}{4} \left[(T - t)^2 + \sum_{i \in \mathcal{J}_{(t,\infty)}^{\mathbb{T}}} (T_i - t)^2 \right]$.*

Proof of lemma A.2 (a) The procedure in the proof of lemma A.1 gives us a desired set of contrasts. Indeed, let $(i_k)_{k=1}^m$ be the chosen set of nodes and $(i'_k)_{k=1}^m$ be their parents. Then $\mathcal{J}_{[0,t]}^{\mathbb{T}} \subset \bigcup_{k=1}^m \{i_k, i'_k\}$, hence

$$\sum_{i \in \mathcal{J}_{[0,t]}^{\mathbb{T}}} (T_i - t)^2 \leq \sum_{k=1}^m (T_{i_k} - t)^2 + (T_{i'_k} - t)^2 \leq 2 \sum_{k=1}^m (T_{i_k} - t)^2 = 2 \sum_{C \in \mathcal{C}} (T_C - t)^2.$$

(b) Contrasts are chosen by induction, starting with $\mathcal{C} = \emptyset$. Let $r^{\mathbb{T}}$ be the root of $\mathbb{T}$. If $r^{\mathbb{T}} \notin \mathcal{J}_{(t,T]}^{\mathbb{T}}$ then we stop, else we update $\mathcal{C} = \mathcal{C} \cup \{C_{r^{\mathbb{T}}}^{p_{r^{\mathbb{T}}}}\}$ where the path $p_{r^{\mathbb{T}}}$

is chosen carefully as follows. From each child of the root, the path descends toward the tips. Each time an internal node is encountered, a decision needs to be made to either go left or right. Of the two children of the internal node, the path is connected to the youngest (figure A.1). We then remove from $\mathbb{T}$ the path $p_r^\mathbb{T}$ and the edges connected to it. What is left is a forest, a set of subtrees of $\mathbb{T}$, one which we repeat the procedure, recursively extracting one path and its corresponding contrast from each subtree.

We now prove by induction that this procedure gives us a desired set of contrasts. This is easy to see for ≤ 3 tips. Assume that it is true for every tree with $\leq m$ tips, and that $\mathbb{T}$ has $m + 1$ tips. Let i_1 and i_2 be the two children of $r^\mathbb{T}$. Let $(\mathbb{T}_k)_{k=1}^l$ be the subtrees obtained after removing $p_r^\mathbb{T}$ and the edges connected to it, and such that $r^{\mathbb{T}_k} \in \mathcal{J}_{(t, \mathbb{T})}^\mathbb{T}$. Let s_k be the sibling of $r^{\mathbb{T}_k}$ in $\mathbb{T}$ (s_k could be a leaf). By construction, $T_{s_k} \leq T_{r^{\mathbb{T}_k}}$. Let $\mathcal{C}_k$ be the set of contrasts obtained from $\mathbb{T}_k$. We have $\mathcal{J}_{(t, \mathbb{T})}^\mathbb{T} \subset \{r^\mathbb{T}, i_1, i_2\} \cup_{k=1}^l \mathcal{J}_{(t, \mathbb{T})}^{\mathbb{T}_k} \cup \{s_k\}$, and $\mathcal{C} = \{r^\mathbb{T}\} \cup_{k=1}^l \mathcal{C}_k$. Therefore:

$$\begin{aligned} 4 \sum_{C \in \mathcal{C}} (T_C - t)^2 &= 4(T_{r^\mathbb{T}} - t)^2 + 4 \sum_{k=1}^l \sum_{C \in \mathcal{C}_k} (T_C - t)^2 \\ &\geq 2(T_{r^\mathbb{T}} - t)^2 + (\max\{T_{i_1}, t\} - t)^2 + (\max\{T_{i_2}, t\} - t)^2 \\ &\quad + \sum_{k=1}^l \left\{ (T_{r^{\mathbb{T}_k}} - t)^2 + \sum_{i \in \mathcal{J}_{(t, \mathbb{T})}^{\mathbb{T}_k}} (T_i - t)^2 \right\} \geq (T_{r^\mathbb{T}} - t)^2 + \sum_{i \in \mathcal{J}_{(t, \mathbb{T})}^\mathbb{T}} (T_i - t)^2. \end{aligned}$$

A.2.2 Technical proofs for section 2.2

Counter example for theorem 2.2 on non-ultrametric trees Let $a = e^{-\alpha t_1}$ and $b = e^{-\alpha t_2}$. It is easy to see that V_α can be expressed in terms of the $n/2 \times n/2$ identity matrix I as $V_\alpha = \text{diag}((1 - a^2)I, (1 - b^2)I) + (a\mathbf{1}^t, b\mathbf{1}^t)^t (a\mathbf{1}^t, b\mathbf{1}^t)$. We then get V_α^{-1} using Woodbury formula, then

$$\mathbf{1}^t V_\alpha^{-1} \mathbf{1} = n \left(\frac{1}{1 - a^2} + \frac{1}{1 - b^2} + \frac{n(a - b)^2}{(1 - a^2)(1 - b^2)} \right) / \left(1 + \frac{na^2}{1 - a^2} + \frac{nb^2}{1 - b^2} \right). \text{ If } t_1 \neq$$

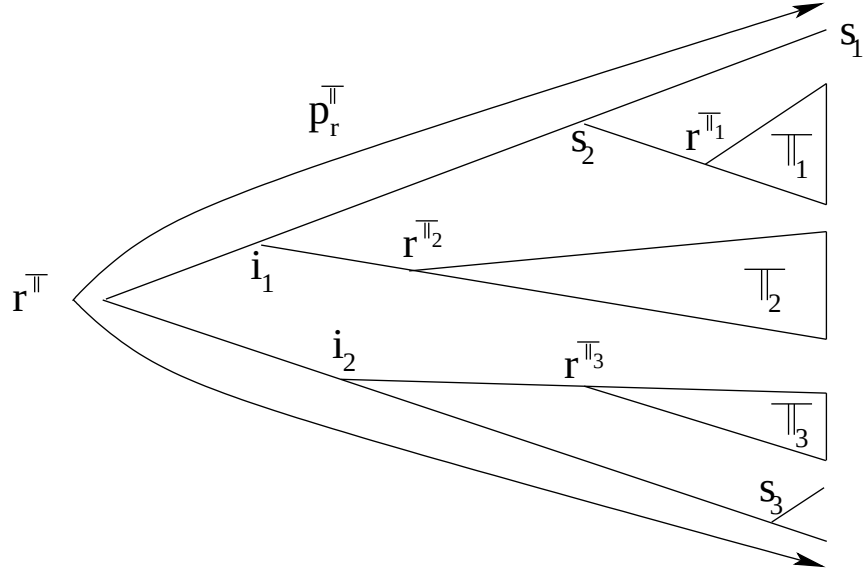


Figure A.1: Recursive construction of independent contrasts, taken with respect to the root at each step.

t_2 then $a \neq b$ and $\text{var}(\hat{\mu}) = (\mathbf{1}^t V_\alpha^{-1} \mathbf{1})^{-1}$ goes to 0 as claimed.

Proof of lemma 2.1 We will first prove $V_\alpha \geq e^{-2\alpha T} J_n$ by induction on the number of tips, where $J_n = \mathbf{1}\mathbf{1}^t$. Clearly, this is true for trees with a single tip. Now consider a tree with n tips, and consider its k subtrees obtained by removing the k branches stemming from the root. Let $T_1, \dots, T_k$ be the heights of these subtrees, that is, the age of their roots. Their number of tips $n_1, \dots, n_k$ is at most $n - 1$. So by induction, the covariance matrices $V_\alpha^{(1)}, \dots, V_\alpha^{(k)}$ associated with these subtrees must satisfy $V_\alpha^{(i)} \geq e^{-2\alpha T_i} J_{n_i}$. Therefore $V_\alpha - e^{-2\alpha T} J_n \geq \text{diag}(V_\alpha^{(i)} - e^{-2\alpha T_i} J_{n_i}) \geq 0$ is true for all trees. Now we use the definition of t and go a step further using that $V_\alpha^{(i)} - e^{-2\alpha T} J_{n_i} \geq (e^{-2\alpha T_i} - e^{-2\alpha T}) J_{n_i} \geq (e^{-2\alpha(T-t)} - e^{-2\alpha T}) J_{n_i}$ for all $i = 1, \dots, k$. This implies that $V_\alpha - e^{-2\alpha T} J_n \geq (e^{-2\alpha(T-t)} - e^{-2\alpha T}) \text{diag}(J_{n_1}, \dots, J_{n_k}) \geq \frac{1}{k} (e^{-2\alpha(T-t)} - e^{-2\alpha T}) J_n$, from which lemma 2.1 follows easily.

Proof of upper bound (2.6) Assume here that $\alpha_1 = \alpha_2$ and $\gamma_1 = \gamma_2 = \gamma$. Since $(Y_i)_{i=1}^n$ have the same covariance matrix $\gamma \mathbf{V}$ under both distributions P_{θ_1} and P_{θ_2} , it is easy to see that $r(\mathbb{T}) = (\mu_1 - \mu_2)^2 \mathbf{1}^t \mathbf{V}^{-1} \mathbf{1} / \gamma$ (Hershey and Olsen, 2007). The bound $r(\mathbb{T}) \leq (\mu_1 - \mu_2)^2 / (\gamma e^{-2\alpha T})$, where T is the age of the root, then follows from lemma 2.1.

Proof of lemma 2.2 First, $\text{var}(C_{i_1}^{p_i}) = \text{var}(Y_1^{i_1}) + \text{var}(Y_2^{i_2}) - 2\text{cov}(Y_1^{i_1}, Y_2^{i_2}) = 2\gamma - 2\gamma e^{-2\alpha T_{i_1}}$. Second, consider two paths p_{i_1} and p_{i_2} that do not intersect. Then, the most recent common ancestor of $Y_j^{i_1}$ and $Y_k^{i_2}$ ($j, k \in \{1, 2\}$) is the most recent common ancestor of internal nodes i_1 and i_2 . Therefore, the distance from $Y_1^{i_1}$ to $Y_1^{i_2}$ equals the distance from $Y_2^{i_1}$ to $Y_1^{i_2}$. Hence $\text{cov}(Y_1^{i_1}, Y_1^{i_2}) = \text{cov}(Y_2^{i_1}, Y_1^{i_2})$. Similarly, $\text{cov}(Y_1^{i_1}, Y_2^{i_2}) = \text{cov}(Y_2^{i_1}, Y_2^{i_2})$. Therefore $\text{cov}(C_{i_1}^{p_i}, C_{i_2}^{p_i}) = \text{cov}(Y_1^{i_1} - Y_2^{i_1}, Y_1^{i_2} - Y_2^{i_2}) = 0$.

Proof of lemma 2.3 Define $h_1(x) = (1 - e^{-2x\alpha_2}) / (1 - e^{-2x\alpha_1})$ and assume $t_1 \neq 0$ and $t_2 \neq 0$. From the system of equations, we have $\gamma_1 / \gamma_2 = h_1(t_1) = h_1(t_2)$. Now $(\log h_1)'(x) / x = h_2(x\alpha_2) - h_2(x\alpha_1)$ where $h_2(x) = xe^{-x} / (1 - e^{-x})$ is monotone on $(0, \infty)$. So $\alpha_1 = \alpha_2$, and $\gamma_1 = \gamma_2$. If $t_2 = 0$ we make a similar argument because $h_3(x) = x / (1 - e^{-2x\alpha_2})$ is monotone on $(0, \infty)$.

Proof of theorem 2.4 part a Under the symmetric tree growth model,

$$r(\mathbb{T}) = \frac{1}{2} \sum_{k=1}^m d_1 \dots d_{k-1} (d_k - 1) \left(\frac{\gamma_2 \lambda_k(\alpha_2)}{\gamma_1 \lambda_k(\alpha_1)} + \frac{\gamma_1 \lambda_k(\alpha_1)}{\gamma_2 \lambda_k(\alpha_2)} - 2 \right) + \left(m_{1,n}^2 + \frac{m_{1,n}^2}{\sigma_{1,n}^2} \right).$$

To show this, we consider $\mathbf{h} = (Y_{j,n})_{j \leq n} = \gamma_1^{-1/2} \mathbf{\Lambda}^{-1/2}(\alpha_1) \mathbf{P}^{-1}(\mathbf{Y} - \mu_1 \mathbf{1})$, where $\mathbf{\Lambda}(\alpha) = \text{diag}(\lambda_k(\alpha))$ contains the eigenvalues λ_k with their multiplicities and $\mathbf{P}$ contains the eigenvectors of $\mathbf{V}_{\alpha}$, which do not depend of α (appendix A.1). Then $\mathbf{h}$ is orthonormal under P_{θ_1} , and orthogonal under P_{θ_2} with variances $(\gamma_2 / \gamma_1) \lambda_k(\alpha_2) / \lambda_k(\alpha_1)$ with multiplicities $d_1 \dots d_{k-1} (d_k - 1)$. Furthermore, $E_2 \mathbf{h} = (\mu_2 - \mu_1) \gamma_1^{-1/2} \mathbf{\Lambda}^{-1/2}(\alpha_1) \mathbf{P}^{-1} \mathbf{1}$ so that $m_{j,n} = 0$ if $j \geq 2$ and $m_{1,n} = (\mu_2 - \mu_1) / \sqrt{n \gamma_1 \lambda_0(\alpha_1)}$, from which $r(\mathbb{T})$ follows.

With increasing node degrees at m levels, it is easy to see that the ratio $\lambda_k(\alpha_1)/\lambda_k(\alpha_2)$ converges to a positive limit for all $k \leq m$. Under the assumption that d_k is fixed for $k < m$, the multiplicity of $\lambda_k(\alpha)$ is constant as n grows, except for $k = m$. $r(\mathbb{T}_m)$ is then expressed as a finite sum where all terms are convergent except for the last term ($k = m$) associated with the smallest eigenvalue $\lambda_m = 1 - e^{-2\alpha t_m}$. This term is bounded if and only if $\gamma_1(1 - e^{-2\alpha_1 t_m}) = \gamma_2(1 - e^{-2\alpha_2 t_m})$, in which case $r(\mathbb{T}_n)$ converges to a finite value. Otherwise, $r(\mathbb{T}_n)$ goes to infinity. Hence P_{θ_1} and P_{θ_2} are equivalent if and only if $\gamma_1(1 - e^{-2\alpha_1 t_m}) = \gamma_2(1 - e^{-2\alpha_2 t_m})$, which completes the proof.

Proof of theorem 2.4 part b We denote here $\lambda_k = \lambda_{k,m}$ to emphasize the dependence of m . We first consider case (i) when $t_0 = 0$. When $d_k = d$ and $u_k = q^k$ the eigenvalues simplify to:

$$\frac{\lambda_{k,m}(\alpha)}{d^{m-k}} = \sum_{j=0}^{m-k-1} \frac{e^{-2\alpha q^{k+1+j}} - e^{-2\alpha q^{k+j}}}{d^j} + \frac{1 - e^{-2\alpha q^m}}{d^{m-k}}. \quad (\text{A.1})$$

It is then easy to see that for all $\alpha \geq 0$ and k , $\lambda_{k,m}(\alpha)/d^{m-k}$ converges to some finite function of α and k . To prove the convergence of $r(\mathbb{T}_m)$ we will need the following lemma, which is proved later.

Lemma A.3. *Let $\gamma_1\alpha_1 = \gamma_2\alpha_2$, that is, $\sigma_1^2 = \sigma_2^2$. Then there exists K , c and C which depend only on α_1, α_2, d and q such that for all $m > k \geq K$*

$$cq^{2k} \leq \frac{\gamma_2\lambda_{k,m}(\alpha_2)}{\gamma_1\lambda_{k,m}(\alpha_1)} + \frac{\gamma_1\lambda_{k,m}(\alpha_1)}{\gamma_2\lambda_{k,m}(\alpha_2)} - 2 \leq Cq^{2k}.$$

Because $\gamma\alpha$ is microergodic (theorem 2.3 part a), we can assume $\gamma_1\alpha_1 = \gamma_2\alpha_2$. Lemma A.3 implies that the first sum in the expression of $r(\mathbb{T}_m)$ (from the proof of part a) is bounded above and below by $\sum_{k=K}^m (dq^2)^k$ up to some multiplicative constant, and so converges to a finite limit because $dq^2 < 1$. The last term with $m_{1,n}^2$ is always bounded as shown in the proof of (2.6). This completes the proof.

We now turn to case (ii) with $t_0 > 0$. To prove that (γ, α) is microergodic, we will show that $P_{\theta_1} \perp P_{\theta_2}$ under the restriction $\gamma_1(1 - e^{-2t\alpha_1}) = \gamma_2(1 - e^{-2t\alpha_2})$. To do so, we only need to check the sufficient condition in (2.5). Note that there exists $w > 0$ such that $d^w q \geq 1$. Denote $k_m = \lfloor m/(w+1) \rfloor$ where $\lfloor x \rfloor$ is a largest integer smaller than x . The condition in (2.5), denoted by z_m , can be written as

$$\begin{aligned} z_m &= \sum_{k=1}^m d^{k-1} (d-1) \left(\frac{\gamma_1 \lambda_{k,m}(\alpha_1)}{\gamma_2 \lambda_{k,m}(\alpha_2)} - 1 \right)^2 \geq d^{k_m-1} \left(\frac{\gamma_1 \lambda_{k_m,m}(\alpha_1)}{\gamma_2 \lambda_{k_m,m}(\alpha_2)} - 1 \right)^2 \\ &\geq d^{k_m-1} \left(\frac{(h_{t_0}(\alpha_1) - h_{t_0}(\alpha_2)) f_{m,1} + O_{\alpha_1, \alpha_2, t_0}(1) q^{k_m}}{1 + h_{t_0}(\alpha_2) f_{m,1} + O_{\alpha_2, t_0}(1) q^{k_m}} \right)^2 \end{aligned}$$

where $h_t(\alpha) = \frac{2\alpha e^{-2\alpha t}}{1 - e^{-2\alpha t}}$ and $f_{m,1}(q) = \sum_{j=0}^{m-k-1} \left(\frac{q}{d}\right)^j + \frac{1}{1-q} \left(\frac{q}{d}\right)^{m-k}$. If $(\gamma_1, \alpha_1) \neq (\gamma_2, \alpha_2)$, then $z_m \rightarrow \infty$ because $h_t(\alpha)$ is monotone in α .

Proof of lemma A.3 We first note that for every $a > 0$ there exists $x_a > 0$ such that $e^{-ax} - (1 - ax + a^2 x^2 / 2) = O(a^3 x^3)$ uniformly for all x in $[0, x_a]$. Therefore there exists $K = K(\alpha, q)$ such that $e^{-2\alpha q^{k+j+1}} - e^{-2\alpha q^{k+j}} - 2\alpha q^{k+j}(1-q) + 2\alpha^2 q^{2k+2j}(1-q^2) = q^{3k+3j} O_\alpha(1)$ where the $O_\alpha(1)$ term is bounded uniformly in $k+j \geq K$. We can now combine this with (A.1): $\lambda_{k,m}(\alpha)/d^{m-k} - 2\alpha(1-q)q^k f_1 + 2\alpha^2(1-q^2)q^{2k} f_2 = q^{3k} f_3 O_\alpha(1)$ where f_1, f_2 and f_3 only depend on $q, d, m-k$ and are defined by $f_1 = f_{m,1}(q)$, $f_2 = f_{m,1}(q^2)$ and $f_3 = f_{m,1}(q^3)$. Because the f values are bounded as $m-k$ grows, we get $\frac{\lambda_{k,m}(\alpha_1)}{\lambda_{k,m}(\alpha_2)} = \frac{\alpha_1}{\alpha_2} (1 + (\alpha_2 - \alpha_1)(1+q)q^k f_2/f_1) + q^{2k} O(1)$ where the $O(1)$ term is bounded uniformly in $m > k \geq K$, and the same formula holds when α_2 and α_1 are switched. Lemma A.3 then follows immediately because we assume that $\gamma_1 \alpha_1 = \gamma_2 \alpha_2$.

A.2.2.1 Technical proofs for section 2.3

Criterion for the consistency and asymptotic normality of REML estimators

In appendix A.1, we showed that $\mathbf{1}$ is an eigenvector of V_α for symmetric trees, independently of α . Therefore, the REML estimator of (γ, α) based on $\mathbf{Y}$ is the ML

estimator of (γ, α) based on the transformed data $\tilde{Y} = \tilde{P}^t Y$ where $\tilde{P}$ is the matrix of all eigenvectors but $\mathbf{1}$. $\tilde{Y}$ is Gaussian centered with variance $\Sigma_n = \gamma \tilde{\Lambda}$ where $\tilde{\Lambda}$ is the diagonal matrix of all eigenvalues of V_α but $\lambda_0(\alpha)$. Following [Mardia and Marshall \(1984\)](#) and like [Cressie and Lahiri \(1993\)](#), we use a general result from [Sweeting \(1980\)](#). The following conditions C1-C2 ensure the consistency and asymptotic normality of the ML estimator (reworded from [Mardia and Marshall, 1984](#)). Assume there exists non-random continuous symmetric matrices $A_n(\theta)$ such that

- (C1) (i) As n goes to infinity A_n^{-1} converges to 0.
(ii) $A_n^{-1} \mathcal{J}_n A_n^{-1}$ converges in probability to a positive definite matrix $W(\theta)$, where $\mathcal{J}_n$ is the second-order derivative of the negative log likelihood function L .
- (C2) Σ_n is twice continuously differentiable on Θ with continuous second derivatives.

Under these conditions, the MLE $\hat{\theta}$ satisfies $A_n(\theta)(\hat{\theta} - \theta) \xrightarrow{d} N(0, W(\theta)^{-1})$. A standard choice for A_n is the inverse of the square-root of the Fisher information matrix $B_n = E(\mathcal{J}_n)$. Because (C1)(ii) is usually difficult to verify, [Mardia and Marshall \(1984\)](#) suggest using a stronger L^2 -convergence condition. This approach was later taken by [Cressie and Lahiri \(1993; 1996\)](#). Unfortunately, their conditions for establishing (C1) do not hold here, because the largest eigenvalues and the ratio of the largest to the smallest eigenvalues are both of order n . In what follows, we will check (C1) for the particular choice of $A_n = B_n^{1/2}$ and $W(\theta) = I$ and where we replace (C1)(ii) by the stronger condition

- (C1) (ii') $\sum_{i,j,k,l=1,2} b^{ki} b^{lj} \text{tr}(\Sigma_n(\Sigma_n^{-1})_{kj} \Sigma_n(\Sigma_n^{-1})_{li})$ converges to 0, where b^{ij} is the (i, j) -element of B_n^{-1} and $(\Sigma_n^{-1})_{ij}$ is the (i, j) -second order derivative of Σ_n^{-1} .

Proof of Theorem 2.5 It is convenient here to re-parametrize the model using (ν, α) . The diagonal elements in Σ_n are $\nu \lambda_k(\alpha) / \lambda_m(\alpha)$ with multiplicity $d_1 \dots d_{k-1}(d_k - 1)$. The smallest is ν (for $k = m$) with multiplicity $n - \tilde{n}$, which

is conveniently independent of α . With this parametrization, the inverse of the Fisher information matrix is the symmetric matrix

$$\mathbf{B}_n^{-1} = \frac{2}{\det \mathbf{B}_n} \begin{pmatrix} \sum_{k=1}^{m-1} d_1 \dots d_{k-1} (d_k - 1) (\Lambda_{k,m} - \Lambda_{m,m})^2 & * \\ -\nu^{-1} \sum_{k=1}^{m-1} d_1 \dots d_{k-1} (d_k - 1) (\Lambda_{k,m} - \Lambda_{m,m}) & (n-1)/\nu^2 \end{pmatrix}$$

where $\Lambda_{k,m} = \lambda'_{k,m}/\lambda_{k,m}$, $\det \mathbf{B}_n = (n-1)^2/(4\nu^2) \text{var}_q(\Lambda_{K,m} - \Lambda_{m,m})$, and the variance is taken with respect to $\mathbb{P}[K = k] = q_{k,n} = d_1 \dots d_{k-1} (d_k - 1)/(n-1)$. When the degree at the last level near the tips d_m becomes large then $q_{m,n} \sim 1$, i.e. the distribution q is concentrated around the high end $K = m$. It is then useful to express

$$\det \mathbf{B}_n = \frac{(n - \tilde{n})(\tilde{n} - 1)}{4\nu^2} \mathbb{E}_p(\Lambda_{K,m} - \Lambda_{m,m})^2 + \frac{(\tilde{n} - 1)^2}{4\nu^2} \text{var}_p(\Lambda_{K,m} - \Lambda_{m,m})$$

where the expectation and variance are now taken with respect to $\mathbb{P}[K = k] = p_{k,n} = d_1 \dots d_{k-1} (d_k - 1)/(\tilde{n} - 1)$ for $k < m$, that is, $p_{k,n} = q_{k,n}(n-1)/(\tilde{n}-1)$. To verify conditions (C1)(i) and (ii'), we will use the following lemmas.

Lemma A.4. $\Lambda_{1,m} < \Lambda_{2,m} < \dots < \Lambda_{m,m}$. Moreover for any fixed T and $\alpha > 0$, $\Lambda_{k,m}$ and $\lambda''_{k,m}/\lambda_{k,m}$ are uniformly bounded. Specifically, $|\Lambda_{k,m}| \leq \max\{2T, 1/\alpha\}$ and $|\lambda''_{k,m}/\lambda_{k,m}| \leq 4 \max\{T^2, T/\alpha\}$.

Proof of lemma A.4 Denote $g(\alpha) = b - (b+c)e^{-c\alpha} + ce^{-(b+c)\alpha}$. It is easy to see that $g' > 0$ then $g > 0$ for all $\alpha, b, c > 0$. It follows that

$$\frac{(a+b)e^{-\alpha b} - a}{1 - e^{-\alpha b}} - \frac{(a+b+c)e^{-\alpha c} - (a+b)}{1 - e^{-\alpha c}} > 0, \forall a \in \mathbb{R}, \alpha, b, c > 0.$$

Now let $a_i > 0$ for $i = 1, \dots, n+1$ and let $\Lambda_k = \sum_{i=1}^k a_i$. By applying the previous inequality with $a = \Lambda_{n-1}$, $b = a_n$ and $c = a_{n+1}$, we get that

$$\frac{\Lambda_n e^{-\alpha \Lambda_n} - \Lambda_{n-1} e^{-\alpha \Lambda_{n-1}}}{e^{-\alpha \Lambda_{n-1}} - e^{-\alpha \Lambda_n}} > \frac{\Lambda_{n+1} e^{-\alpha \Lambda_{n+1}} - \Lambda_n e^{-\alpha \Lambda_n}}{e^{-\alpha \Lambda_n} - e^{-\alpha \Lambda_{n+1}}}.$$

Recall that $\lambda_{k,m} = \sum_{i=k}^m d_{i+1} \dots d_m (e^{-2\alpha u_{i+1}} - e^{-2\alpha u_i})$. The monotonicity of $\Lambda_{k,m}$ in k follows easily from combining the inequality above with the fact that if $x_1/y_1 > \dots > x_n/y_n$ and if $y_i, c_i > 0$, then $\sum_{i=1}^{n-1} c_i x_i / \sum_{i=1}^{n-1} c_i y_i > \sum_{i=1}^n c_i x_i / \sum_{i=1}^n c_i y_i$. The proof of the second part of lemma A.4 is easy and left to the reader. The following lemma results directly from lemma A.4.

Lemma A.5. *With m fixed and parametrization (v, α) , the quantities $(n-1) \sum_{k=1}^m q_{k,n} (\Lambda_{k,m} - \Lambda_{m,m})^2$, $(n-1) \sum_{k=1}^m q_{k,n} (\Lambda_{k,m} - \Lambda_{m,m})$ and the trace of $\Sigma_n (\Sigma_n^{-1})_{kj} \Sigma_n (\Sigma_n^{-1})_{li}$ are bounded in $O(\tilde{n})$ uniformly on any compact subset of $\{T > 0, \alpha > 0\}$. Therefore, (C1)(i) and (ii') are satisfied if $\det \mathbf{B}_n$ is of order greater than $n \tilde{n}^{1/2}$, i.e. if $\det \mathbf{B}_n^{-1} = o(n^{-1} \tilde{n}^{-1/2})$.*

It is easy to see that $\det \mathbf{B}_n \sim 2n\tilde{n}/(v^2 v_\alpha)$ with v_α defined later. Indeed, $p_{k,n}$ converges to 0 when $k < s$, where s is the largest level $\leq m-1$ such that d_s goes to infinity and $d_{s+1}, \dots, d_{m-1}$ are fixed. For $k = s$, $p_{s,n}$ converges to $p_s = 1/(d_{s+1} \dots d_{m-1})$, and $p_{k,n}$ converges to $p_k = (d_k - 1)/(d_k \dots d_{m-1})$ for $s < k < m$. Note that $p_s, \dots, p_{m-1}$ are the asymptotic relative frequencies of node ages at levels $s, \dots, m-1$. If d_m goes to infinity then $v_\alpha = 8 / \sum_{k=s}^{m-1} p_k (\Lambda_k - \Lambda_m)^2$ with $\Lambda_k = \lim_n \Lambda_{k,m}$. If d_m is fixed,

$$v_\alpha = 8 \left(\sum_{k=s}^{m-1} p_k (\Lambda_k - \Lambda_m)^2 - \left(\sum_{k=s}^{m-1} p_k (\Lambda_k - \Lambda_m) \right)^2 / d_m \right)^{-1}.$$

Clearly, $v_\alpha > 0$ because $p_{m-1} = 1 - 1/d_{m-1} > 0$ is fixed and $\Lambda_{m-1} - \Lambda_m > 0$ is easily checked. So $\det \mathbf{B}_n$ is of order $n\tilde{n}$. The consistency and asymptotic normality of $(\hat{v}, \hat{\alpha})$ follows from applying lemma A.5.

For the second part of the theorem, we obtain the asymptotic normality of

$\sqrt{\tilde{n}}(\hat{\gamma} - \gamma, \hat{\alpha} - \alpha)$ through that of $\sqrt{\tilde{n}}(c_1\hat{\gamma} + c_2\hat{\alpha} - c_1\gamma - c_2\alpha)$ for every $c_1, c_2 \in \mathbb{R}$. For this we apply the following δ -method. Its proof is similar to that of the classical δ -method (Shao, 1999) and is left to the reader.

Lemma A.6. *Assume that $(a_n(X_n - x), b_n(Y_n - y))^t$ converges in distribution to $N(\mathbf{0}, \Sigma)$, with $a_n, b_n \rightarrow \infty$, $b_n/a_n \rightarrow 0$ and $\Sigma_{22} > 0$. Suppose that $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a continuous differentiable function such that $\partial g / \partial y(x, y) \neq 0$. Then $b_n(g(X_n, Y_n) - g(x, y))$ also converges to a centered normal distribution with variance $\Sigma_{22}(\partial g / \partial y(x, y))^2$.*

Finally, using the classical δ -method and the fact that $\sqrt{n}(\hat{\gamma} - \gamma)$ is asymptotically normal, we deduce that the asymptotic correlation between $\log \hat{\gamma}$ and $\log(1 - \exp^{-2\hat{\alpha}t_m})$ is -1 if $\tilde{n} = o(n)$.

B APPENDIX B

B.1 Identifiability of multiple selection optima

We first prove that selection optima are not identifiable if each selection regime forms a connected subtree. The OU model for one trait and multiple selection regimes can be written by (4.1) where $\beta = (y_0, \mu_1, \mu_2, \dots, \mu_m)$ now contains multiple selection optima μ_ℓ , and the design matrix $\mathbf{X}$ depends on where in the tree the m selection regimes apply. Specifically, $\mathbf{X}$ is derived from the expectation of trait Y_i at tip i

$$\mathbb{E}[Y_i] = e^{-\alpha T} y_0 + \sum_{k=1}^{\kappa(i)} e^{-\alpha t_{i,k}} (e^{\alpha t_{i,k}} - e^{\alpha t_{i,k-1}}) \theta_{i,k} \quad (\text{B.1})$$

where $\kappa(i)$ be the number of epochs from the root to tip i (as defined by [Beaulieu et al. 2012](#)), and the t values are the transition times between consecutive selective regimes, going forward in time with $t_{i,0} = 0$ at the root and $t_{i,\kappa(i)} = T$ at tip i . Also, $\theta_{i,k}$ is the selection optimum during epoch k along the path to tip i , i.e. $\theta_{i,k} = \mu_\ell$ if epoch $(t_{i,k-1}, t_{i,k})$ is under regime ℓ . If the part under regime ℓ forms a connected component, any two present-day taxa i and j in that selection regime share the same expected value: $\mathbb{E}[Y_i] = \mathbb{E}[Y_j]$. This is because equation (B.1) has the same terms for tips i and j from the root to their common ancestor, and because the connectivity assumption implies that the selection optimum must be μ_ℓ from that ancestor to both i and j . Therefore, the expectation of the observations $\mathbb{E}[\mathbf{Y}] = \mathbf{X}\beta$ contains at most m different values. On the other hand, β is a vector of dimension $m+1$ so $\mathbf{X}$ cannot be of not full rank. If instead at least one selection regime forms a disconnected component in the tree, then two tips i and j in different components of that regime will likely have different means, and all parameters in β will be identifiable (except for some very specific patterns of branch lengths). In that case, $\mathbf{X}$ will be of full rank and all selection optima will be identifiable from a unique MLE.

Next, we turn to situations when the number or age of shifts on a given edge are unknown. We consider shifts $\Delta\mu$ such that $\mu_\ell = \mu_{\ell-1} + (\Delta\mu)_\ell$ at age t_ℓ (now viewed backwards in time). Then (B.1) becomes, for extant taxa i :

$$\mathbb{E}[Y_i] = e^{-\alpha T} y_0 + (1 - e^{-\alpha T}) \mu_0 + \sum_{\ell \in p(\text{root}, i)} (1 - e^{-\alpha t_\ell}) (\Delta\mu)_\ell \quad (\text{B.2})$$

where the sum goes over regime shifts ℓ occurring along the path from the root to taxon i . Therefore, a shift occurring on a given edge affects $\mathbb{E}Y_i$ of all descendant taxa i in the same way, by a mean shift of $\Delta\mathbb{E}Y = (1 - e^{-\alpha t_\ell}) (\Delta\mu)_\ell$. This shift in the expected trait value provides a single equation, which can serve to estimate a single parameter. If the age t_ℓ is known, for instance, then we can solve for the shift in selection optimum: $(\Delta\mu)_\ell = \Delta\mathbb{E}Y / (1 - e^{-\alpha t_\ell})$. The correction term $(1 - e^{-\alpha t_\ell})$ is responsible for phylogenetic inertia ([Hansen and Orzack, 2005](#)). If the age of the shift is unknown, however, t_ℓ and $(\Delta\mu)_\ell$ are not separately identifiable from the single equation above. Similarly, if two (or more) shifts are located on the same branch, then their effects apply to the same set of descendant taxa into a combined mean shift: $\Delta\mathbb{E}Y = (1 - e^{-\alpha t_1}) (\Delta\mu)_1 + (1 - e^{-\alpha t_2}) (\Delta\mu)_2$. The lack of identifiability is even more severe here, with a single equation and 4 parameters if the shift ages are unknown.

B.2 Supplementary Figures

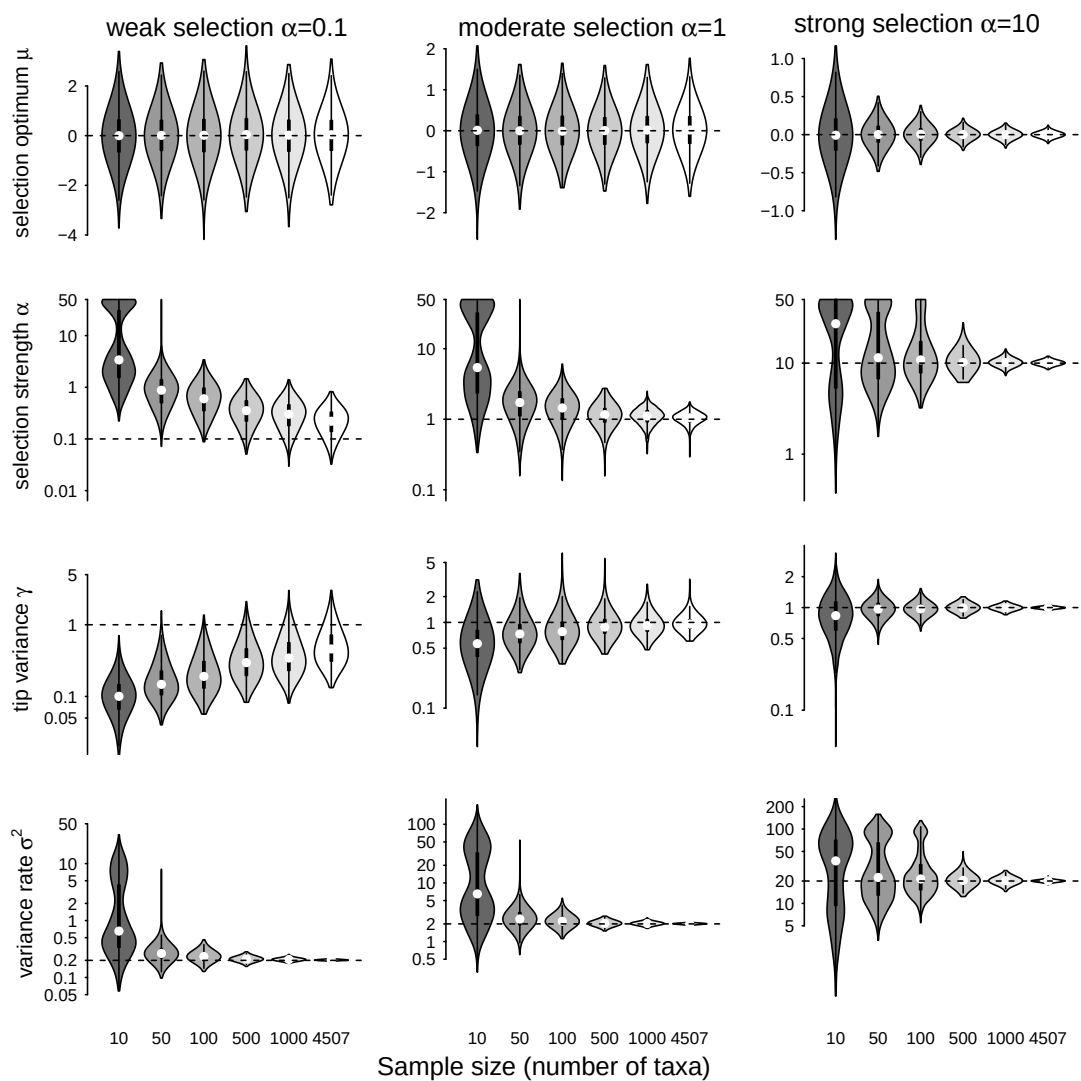


Figure B.1: Distribution of the MLE of μ , γ , α , and $\sigma^2 = 2\alpha\gamma$ on subtrees from the 4507-species mammal tree (Bininda-Emonds et al., 2007). Simulation parameters as in figure 4.3.

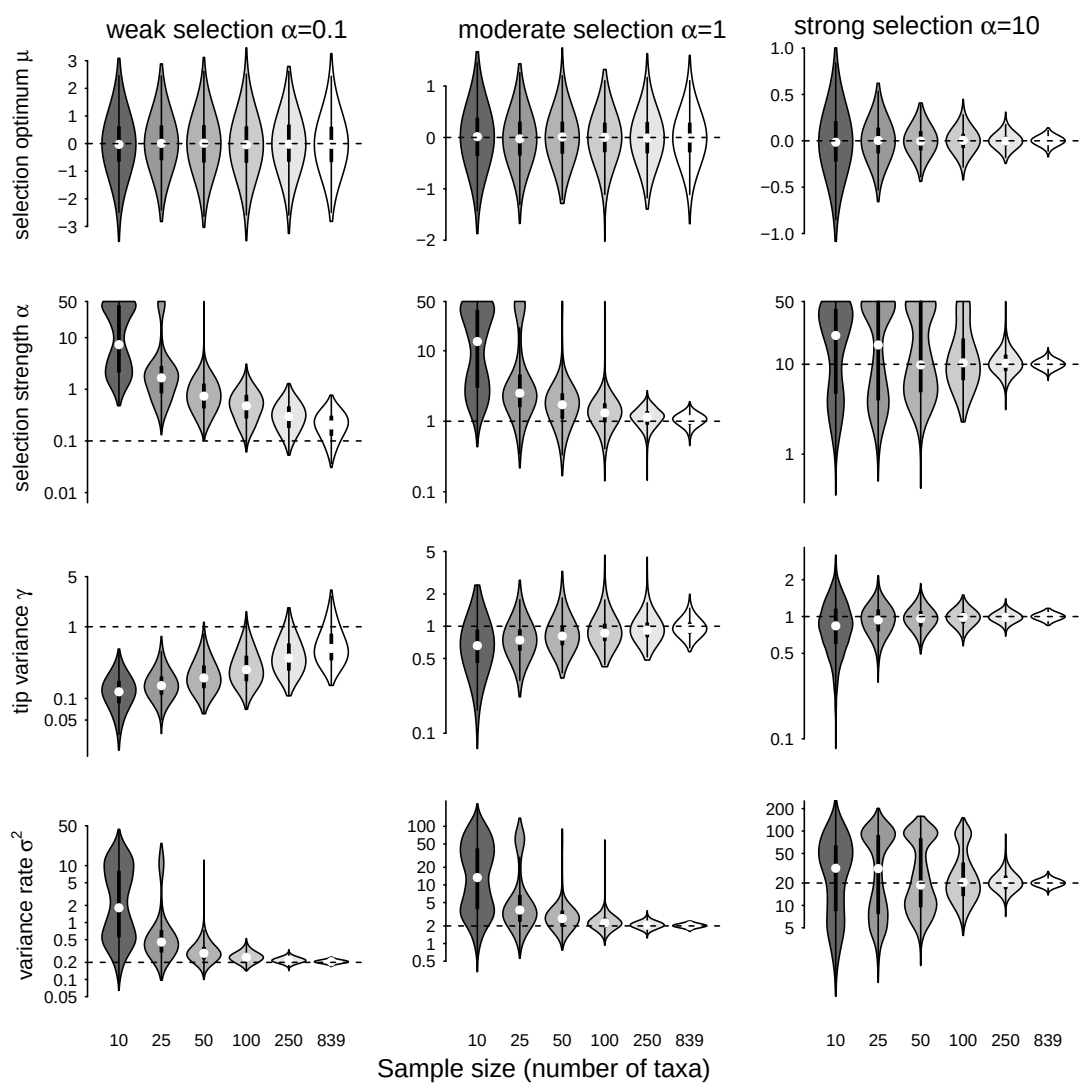


Figure B.2: Distribution of the MLE of μ , γ , α , and $\sigma^2 = 2\alpha\gamma$ on subtrees from the 839-taxon tree on Fabales (Simon et al., 2009). Simulation parameters as in figure 4.3.

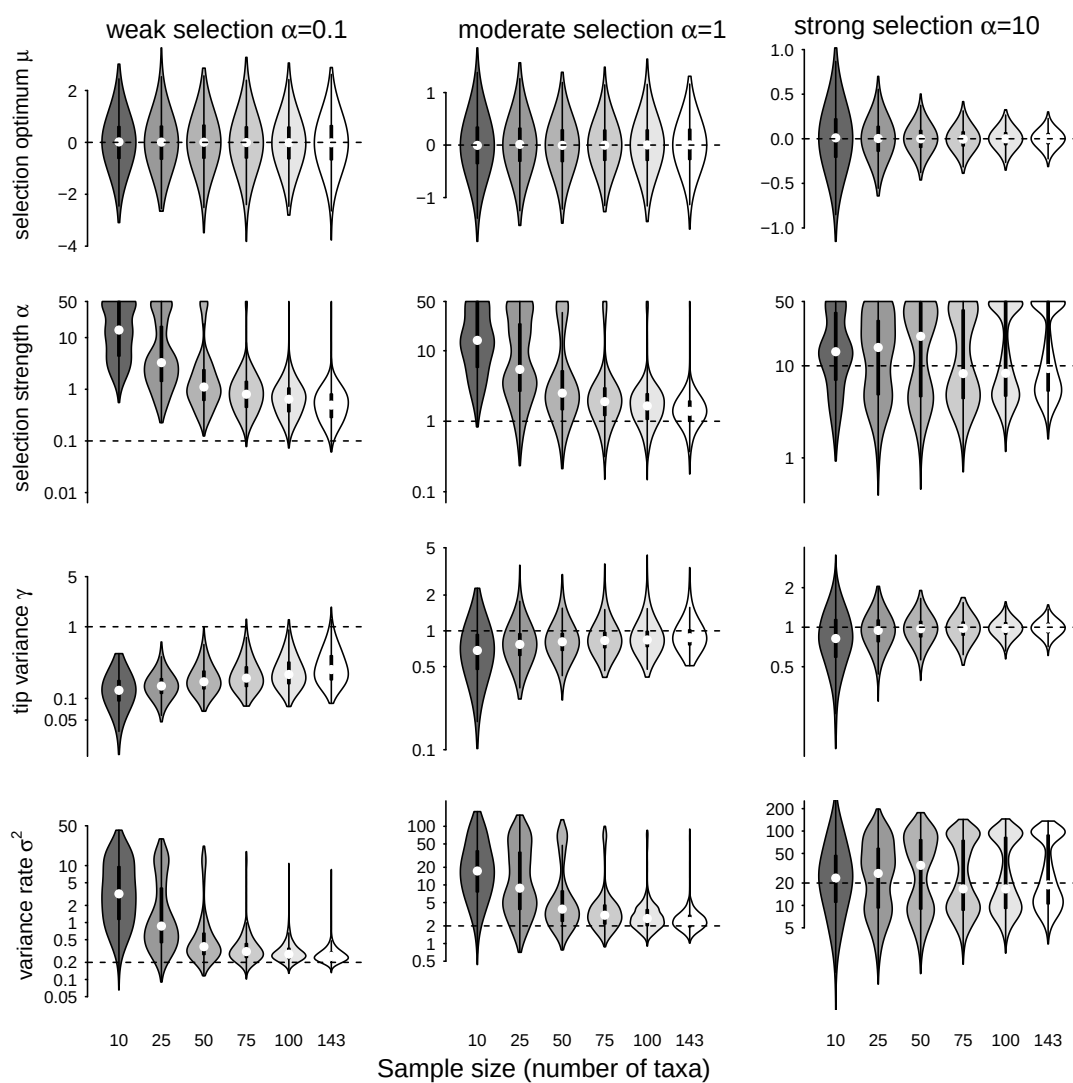


Figure B.3: Distribution of the MLE of μ , γ , α , and $\sigma^2 = 2\alpha\gamma$ on subtrees from the 140-species phylogeny of ants (Moreau et al., 2006). Simulation parameters as in figure 4.3.

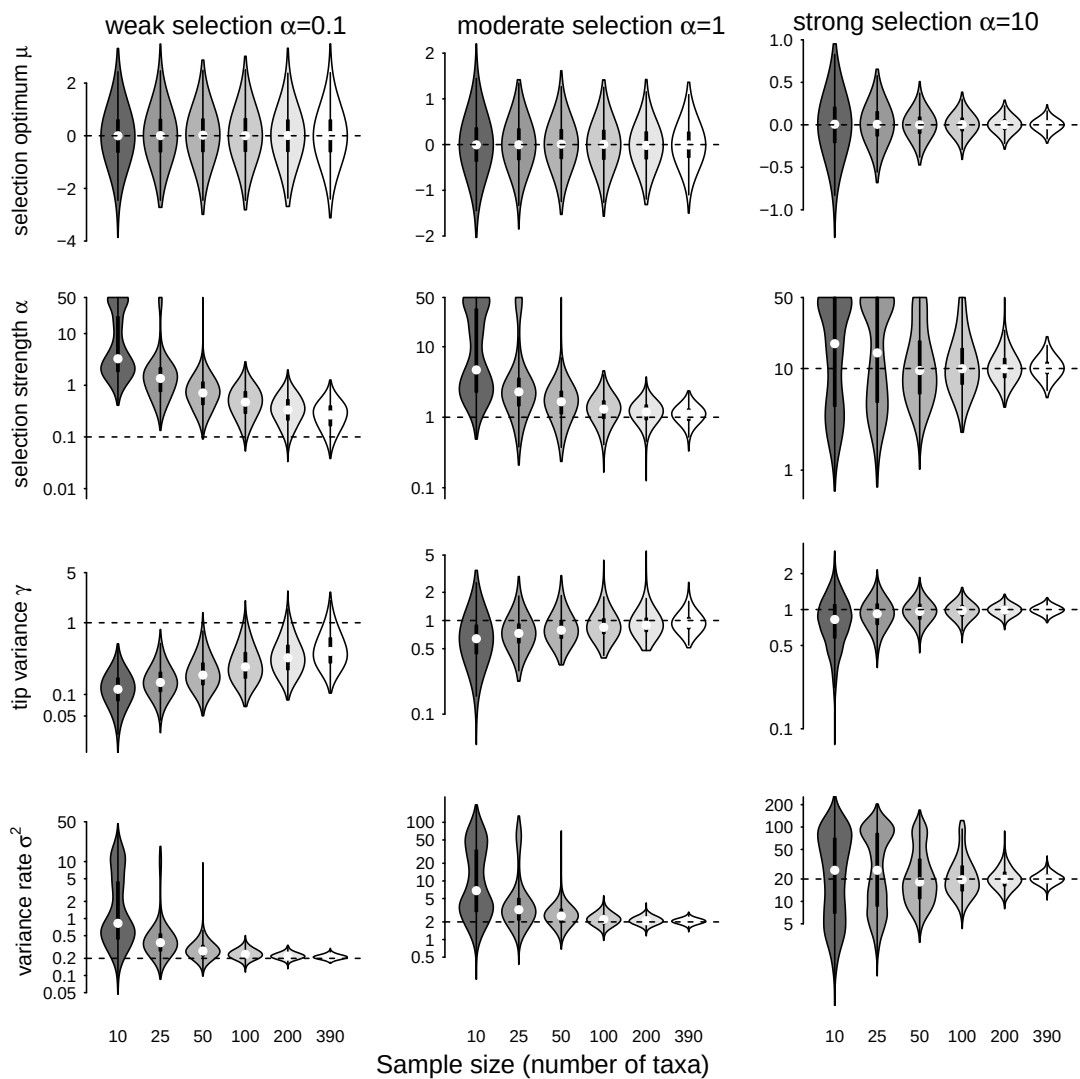


Figure B.4: Distribution of the MLE of μ , γ , α , and $\sigma^2 = 2\alpha\gamma$ on subtrees from the 400-language phylogeny (Gray et al., 2009). Simulation parameters as in figure 4.3.

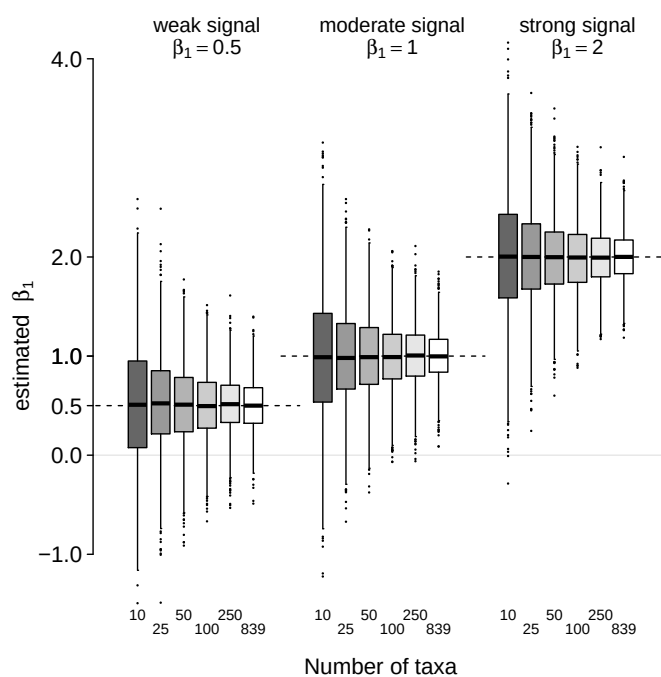


Figure B.5: Boxplots showing the distribution of the MLE of the shift, β_1 , from simulations on the 839-taxon tree on Fabales in fig. 4.4 (right), from 20 sequences of nested trees and 100 data sets per tree.

C.1 Algorithm for multiple traits with different phylogenetic signals

We focus here on the model proposed by [Freckleton \(2012\)](#), where each trait is assumed to follow Pagel's λ model with phylogenetic signal λ_k for trait k , using $\text{cov}(Y_{i,k}, Y_{j,\ell}) = C_{k\ell} V_{ij}(\sqrt{\lambda_k \lambda_\ell})$ for traits k, ℓ and taxa i, j ([Freckleton, 2012](#), Supporting Information). Of interest is $\mathbf{C}$, the evolutionary covariance between the m traits. We will use the fact that $\mathbf{V}(\lambda) = \lambda \mathbf{V} + (1 - \lambda) \mathbf{T}$ where $\mathbf{V}$ is the BM covariance matrix with $V_{ij} = t_{ij}$ and $\mathbf{T}$ is the diagonal matrix with diagonal elements t_{ii} . For this model, the full covariance matrix of all the $\mathbf{Y}$ values stacked as a single vector is $\mathbf{W} = \mathbf{C}_1 \otimes \mathbf{V} + \mathbf{C}_2 \otimes \mathbf{T}$ where $\mathbf{C}_1 = \text{diag}(\sqrt{\lambda_k}) \mathbf{C} \text{diag}(\sqrt{\lambda_k})$ and $\mathbf{C}_2 = \mathbf{C} - \mathbf{C}_1$. Since $\mathbf{V}$ is 3-point structured, we have from (5.1) that $\mathbf{V} = \mathbf{t} \mathbf{1} \mathbf{1}' + \mathbf{A}$ with $\mathbf{A}$ block diagonal, each diagonal block $\mathbf{V}_s$ corresponding to a subtree from the root. We can then write $\mathbf{W} = \mathbf{B} + \mathbf{t} \mathbf{C}_1 \otimes \mathbf{1} \mathbf{1}'$ where $\mathbf{B} = \mathbf{C}_1 \otimes \mathbf{A} + \mathbf{C}_2 \otimes \mathbf{T}$. Up to a permutation of its columns and rows to bring together all traits from the taxa belonging to each subtree s , $\mathbf{B}$ is also block diagonal, with each diagonal block corresponding to a subtree: $\mathbf{W}_s = \mathbf{C}_1 \otimes \mathbf{V}_s + \mathbf{C}_2 \otimes \mathbf{T}_s$. The algorithm can then be adapted to recursively calculate the following quantities: $\mathbf{p} = \mathbf{I} \otimes \mathbf{1}' \cdot \mathbf{W}^{-1} \cdot \mathbf{C}_1 \otimes \mathbf{1}$, $\hat{\boldsymbol{\mu}}_Y = \mathbf{p}^{-1} \cdot \mathbf{I} \otimes \mathbf{1}' \cdot \mathbf{W}^{-1} \mathbf{Y}$, $\tilde{\boldsymbol{\mu}}'_X = \mathbf{X}' \mathbf{W}^{-1} \cdot \mathbf{C}_1 \otimes \mathbf{1} \cdot \mathbf{p}^{-1}$ and finally both $\mathbf{Q} = \mathbf{X}' \mathbf{W}^{-1} \mathbf{Y}$ and $\log |\mathbf{W}|$. The algorithm is adapted as follows.

1. Initialization: for a tree with a single tip i and parent edge of length t , set $\mathbf{p} = (t_{ii} \mathbf{C} \mathbf{C}_1^{-1} - (t_{ii} - t) \mathbf{I})^{-1}$, $\hat{\boldsymbol{\mu}}_Y = \mathbf{C}_1^{-1} \mathbf{Y}$, $\tilde{\boldsymbol{\mu}}'_X = \mathbf{X}'$, $\mathbf{Q} = \tilde{\boldsymbol{\mu}}'_X \mathbf{p} \mathbf{Y}$ and $\log |\mathbf{W}| = \log |t_{ii} \mathbf{C} - (t_{ii} - t) \mathbf{C}_1|$.
2. For a tree with two or more tips and root edge length t , let $\mathbf{p}_s$, $\hat{\boldsymbol{\mu}}_{Y,s}$, $\tilde{\boldsymbol{\mu}}'_{X,s}$ and $\mathbf{Q}_s$ the values obtained recursively from each subtree stemming from the root. Define $\mathbf{p}_A = \sum \mathbf{p}_s$ and weights $\mathbf{w}_s = \mathbf{p}_A^{-1} \mathbf{p}_s$ and $\tilde{\mathbf{w}}_s = \mathbf{p}_s \mathbf{p}_A^{-1}$. Then $\mathbf{p} = \mathbf{p}_A (\mathbf{I} + t \mathbf{p}_A)^{-1}$, $\hat{\boldsymbol{\mu}}_Y = \sum_s \mathbf{w}_s \hat{\boldsymbol{\mu}}_{Y,s}$, $\tilde{\boldsymbol{\mu}}'_X = \sum_s \tilde{\boldsymbol{\mu}}'_{X,s} \tilde{\mathbf{w}}_s$, $\mathbf{Q} = \sum_s \mathbf{Q}_s -$

$$\tilde{\mu}'_X \text{tp}_A^2 (\mathbf{I} + \text{tp}_A)^{-1} \hat{\mu}_Y \text{ and } \log |\mathbf{W}| = \sum_s \log |\mathbf{W}_s| + \log |\mathbf{I} + \text{tp}_A|.$$

3. At the root of the full tree: return $\log |\mathbf{W}|$ and $\mathbf{Q}$ then stop.

C.2 Algorithm for multiple traits with different within and between-species covariances

If the phylogenetic covariance is shared between all traits, but if the within-species phenotypic covariance $\mathbf{C}_w$ is different from the between-species phylogenetic covariance $\mathbf{C}_b$ between traits, the full covariance matrix of all response values (stacked in a single vector) is $\mathbf{W} = \mathbf{C}_b \otimes \mathbf{V} + \mathbf{C}_w \otimes \mathbf{I}$, where $\mathbf{V}$ is the phylogenetic covariance (Felsenstein, 2008). The within-species variance $\mathbf{C}_w$ may also include measurement error, with the model above assuming that this error is independent across species. If $\mathbf{V}$ is 3-point structured, we can again use (5.1) to write $\mathbf{V} = \mathbf{t}\mathbf{t}' + \mathbf{A}$ with $\mathbf{A}$ block diagonal, each diagonal block $\mathbf{V}_s$ corresponding to subtree s . Then we get $\mathbf{W} = \mathbf{B} + \mathbf{t} \mathbf{C}_b \otimes \mathbf{1} \mathbf{I} \otimes \mathbf{1}'$ where $\mathbf{B} = \mathbf{C}_b \otimes \mathbf{A} + \mathbf{C}_w \otimes \mathbf{I}$. Again, up to a permutation of its columns and rows to cluster all data from each subtree, $\mathbf{B}$ is also block diagonal with each diagonal block s corresponding to subtree s : $\mathbf{W}_s = \mathbf{C}_b \otimes \mathbf{V}_s + \mathbf{C}_w \otimes \mathbf{I}$. Therefore, we can recursively calculate the following quantities: $\mathbf{p} = \mathbf{I} \otimes \mathbf{1}' \cdot \mathbf{W}^{-1} \cdot \mathbf{C}_b \otimes \mathbf{1}$, $\hat{\mu}_Y = \mathbf{p}^{-1} \cdot \mathbf{I} \otimes \mathbf{1}' \cdot \mathbf{W}^{-1} \mathbf{Y}$, $\tilde{\mu}'_X = \mathbf{X}' \mathbf{W}^{-1} \cdot \mathbf{C}_b \otimes \mathbf{1} \cdot \mathbf{p}^{-1}$ and as before $\mathbf{Q} = \mathbf{X}' \mathbf{W}^{-1} \mathbf{Y}$ and $\log |\mathbf{W}|$. The same recursion (step 2) applies here as in the previous section, with a different initialization step:

1. For a tree with a single tip (here an individual within a species), set $\mathbf{p} = \mathbf{C}_w^{-1} \mathbf{C}_b$, $\hat{\mu}_Y = \mathbf{C}_b^{-1} \mathbf{Y}$, $\tilde{\mu}'_X = \mathbf{X}'$, $\mathbf{Q} = \mathbf{X}' \mathbf{C}_w^{-1} \mathbf{Y}$ and $\log |\mathbf{V}| = \log |\mathbf{C}_w|$.

REFERENCES

-
- [1] Adamczak, Radosław, and Piotr Miłoś. 2011a. CLT for Ornstein-Uhlenbeck branching particle system. *arXiv preprint arXiv:1111.4559*.
 - [2] ———. 2011b. U-Statistics of Ornstein-Uhlenbeck branching particle system. *Journal of Theoretical Probability* 1–41.
 - [3] Akaike, Hirotugu. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6):716–723.
 - [4] Aldous, David J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Stat. Sci.* 16(1):23–34.
 - [5] Anderes, E. 2010. On the consistent separation of scale and variance for gaussian random fields. *The Annals of Statistics* 38(2):870–893.
 - [6] Anderson, T. W. 1984. *An introduction to multivariate statistical analysis*. 2nd ed. Chichester: Wiley.
 - [7] Ané, Cécile. 2008. Analysis of comparative data with hierarchical autocorrelation. *Annals of Applied Statistics* 2(3):1078–1102.
 - [8] Ané, Cécile, Lam Si Tung Ho, and Sebastien Roch. 2014. Phase transition on the convergence rate of parameter estimation under an Ornstein-Uhlenbeck diffusion on a tree. *Journal of Mathematical Biology*. In review.
 - [9] Athreya, K.B., and P. Ney. 2004. *Branching processes*. Dover Books on Mathematics Series, Dover Publications.
 - [10] Bartoszek, Krzysztof, Jason Pienaar, Petter Mostad, Staffan Andersson, and Thomas F. Hansen. 2012. A phylogenetic comparative method for studying multivariate adaptation. *Journal of Theoretical Biology* 314:204–215.
 - [11] Bartoszek, Krzysztof, and Serik Sagitov. 2012. Phylogenetic confidence intervals for the optimal trait value. *arXiv preprint arXiv:1207.6488*.

- [12] Beaulieu, Jeremy M., Dwueng-Chwuan Jhwueng, Carl Boettiger, and Brian C. O'Meara. 2012. Modeling stabilizing selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution* 66:2369–2383.
- [13] Bininda-Emonds, Olaf, Marcel Cardillo, Kate E. Jones, Ross D. E. MacPhee, Robin M. D. Beck, Richard Grenyer, Samantha A. Price, Rutger A. Vos, John L. Gittleman, and Andy Purvis. 2007. The delayed rise of present-day mammals. *Nature* 446(7135):507–512.
- [14] Blomberg, S. P., Theodore Garland, Jr., and Anthony R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57:717–745.
- [15] Boettiger, Carl, Graham Coop, and Peter Ralph. 2012. Is your phylogeny informative? Measuring the power of comparative methods. *Evolution*. In press.
- [16] Brawand, David, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gabor Csardi, Patrick Harrigan, Manuela Weier, Angelica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W. Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grutzner, Sven Bergmann, Rasmus Nielsen, Svante Pääbo, and Henrik Kaessmann. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478(7369):343–348.
- [17] Butler, Marguerite A., and Aaron A. King. 2004. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *The American Naturalist* 164(6):683–695.
- [18] Cai, X. K., David M. Spooner, and Shelley H. Jansky. 2011. A test of taxonomic and biogeographic predictivity: Resistance to potato virus y in wild relatives of the cultivated potato. *Phytopathology* 101:1074–1080.
- [19] Cooper, Natalie, and Andy Purvis. 2010. Body size evolution in mammals: Complexity in tempo and mode. *The American Naturalist* 175(6):727–738.

- [20] Crawford, Forrest W, and Marc A Suchard. 2013. Diversity, disparity, and evolutionary rate estimation for unresolved Yule trees. *Systematic Biology* 62(3):439–455.
- [21] Cressie, Noel, Jesse Frey, Bronwyn Harch, and Mick Smith. 2006. Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics* 11(2):127–150.
- [22] Cressie, Noel, and Soumendra Nath Lahiri. 1993. The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis* 45:217–233.
- [23] ———. 1996. Asymptotics for REML estimation of spatial covariance parameters. *Journal of Statistical Planning and Inference* 50(3):327–341.
- [24] Cressie, Noel A. C. 1993. *Statistics for Spatial Data*. Revised edition ed. Wiley-Interscience.
- [25] Davis, Charles C., Maribeth Latvis, Daniel L. Nickrent, Kenneth J. Wurdack, and David A. Baum. 2007. Floral gigantism in Rafflesiaceae. *Science* 315(5820):1812.
- [26] Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B* 39(1):1–38.
- [27] Eastman, Jonathan M., Michael E. Alfaro, Paul Joyce, Andrew L. Hipp, and Luke J. Harmon. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution* 65(12).
- [28] Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. Least angle regression. *The Annals of statistics* 32(2):407–499.
- [29] Evans, W. S., C. Kenyon, Y. Peres, and L. J. Schulman. 2000. Broadcasting on trees and the Ising model. *Ann. Appl. Probab.* 10(2):410–433.

- [30] Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American journal of human genetics* 25(5):471.
- [31] ———. 2004. *Inferring phylogenies*. Sinauer Associates.
- [32] Felsenstein, Joseph. 1985. Phylogenies and the comparative method. *The American Naturalist* 125(1):1–15.
- [33] ———. 2008. Comparative methods with sampling error and within-species variation: Contrasts revisited and revised. *The American Naturalist* 171(6):713–725.
- [34] FitzJohn, Richard G. 2012. Diversitree: comparative phylogenetic analyses of diversification in r. *Methods in Ecology and Evolution* 3(6):1084–1092.
- [35] Freckleton, R. P., Paul H. Harvey, and Mark Pagel. 2002. Phylogenetic analysis and comparative data: a test and review of evidence. *The American Naturalist* 160(6):712–726.
- [36] Freckleton, Robert P. 2012. Fast likelihood calculations for comparative analyses. *Methods in Ecology and Evolution* 3(5):940–947.
- [37] Freckleton, Robert P, and Paul H Harvey. 2006. Detecting non-brownian trait evolution in adaptive radiations. *PLoS Biol* 4:e373.
- [38] Freckleton, Robert P, and Walter Jetz. 2009. Space versus phylogeny: disentangling phylogenetic and spatial signals in comparative data. *Proceedings of the Royal Society B: Biological Sciences* 276(1654):21–30.
- [39] Garland, Theodore, Jr., Allan W. Dickerman, Christine M. Janis, and Jason A. Jones. 1993. Phylogenetic analysis of covariance by computer simulation. *Systematic Biology* 42(3):265–292.
- [40] Garland, Theodore, Jr., and Anthony R. Ives. 2000. Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. *The American Naturalist* 155(3):346–364.

- [41] Goff, Stephen A., Matthew Vaughn, Sheldon McKay, Eric Lyons, Ann E. Stapleton, Damian Gessler, Naim Matasci, Liya Wang, Matthew Hanlon, Andrew Lenards, Andy Muir, Nirav Merchant, Sonya Lowry, Stephen Mock, Matthew Helmke, Adam Kubach, Martha Narro, Nicole Hopkins, David Micklos, Uwe Hilgert, Michael Gonzales, Chris Jordan, Edwin Skidmore, Rion Dooley, John Cazes, Robert McLay, Zhenyuan Lu, Shiran Pasternak, Lars Koesterke, William H. Piel, Ruth Grene, Christos Noutsos, Karla Gendler, Xin Feng, Chunlao Tang, Monica Lent, Seung-jin Kim, Kristian Kvilekval, B. S. Manjunath, Val Tannen, Alexandros Stamatakis, Michael Sanderson, Stephen Welch, Karen Cranston, Pamela Soltis, Douglas Soltis, Brian O'Meara, Cécile Ané, Tom Brutnell, Daniel J. Kleibenstein, Jeffrey W. White, Jim Leebens-Mack, Michael J. Donoghue, Edgar P. Spalding, Todd J. Vision, Christopher R. Myers, David Lowenthal, Brian J. Enquist, Brad Boyle, Ali Akoglu, Greg Andrews, Sudha Ram, Doreen Ware, Lincoln Stein, and Dan Stanzione. 2011. The iplant collaborative: Cyberinfrastructure for plant biology. *Frontiers in Plant Science* 2.
- [42] Goldberger, Arthur S. 1962. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association* 57(298):pp. 369–375.
- [43] Grafen, A. 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 326(1233):119–157.
- [44] Gray, Russell D, Alexei J Drummond, and Simon J Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science* 323(5913):479–483.
- [45] Hadfield, J. D., and S. Nakagawa. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology* 23(3):494–508.

- [46] Hadfield, J.D. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software* 33(2):1–22.
- [47] Hager, W. 1989. Updating the inverse of a matrix. *SIAM Review* 31(2): 221–239.
- [48] Hansen, Thomas F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* 51(5):1341–1351.
- [49] Hansen, Thomas F., and Krzysztof Bartoszek. 2012. Interpreting the evolutionary regression: The interplay between observational and biological errors in phylogenetic comparative studies. *Systematic Biology* 61(3):413–425.
- [50] Hansen, Thomas F., and Emilia P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution* 50(4):1404–1417.
- [51] Hansen, Thomas F., and Steven Hecht Orzack. 2005. Assessing current adaptation and phylogenetic inertia as explanations of trait evolution: The need for controlled comparisons. *Evolution* 59(10):2063–2072.
- [52] Hansen, Thomas F., J. Pienaar, and Steven Hecht Orzack. 2008. A comparative method for studying adaptation to a randomly evolving environment. *Evolution* 62(8):1965–1977.
- [53] Harmon, Luke J., Jonathan B. Losos, T. Jonathan Davies, Rosemary G. Gillespie, John L. Gittleman, W. Bryan Jennings, Kenneth H. Kozak, Mark A. McPeck, Franck Moreno-Roark, Thomas J. Near, Andy Purvis, Robert E. Ricklefs, Dolph Schluter, James A. Schulte II, Ole Seehausen, Brian L. Sidlauskas, Omar Torres-Carvajal, Jason T. Weir, and Arne Mooers. 2010. Early bursts of body size and shape evolution are rare in comparative data. *Evolution* 64(8):2385–2396.

- [54] Hershey, John R., and Peder A. Olsen. 2007. Approximating the kullback leibler divergence between gaussian mixture models. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007* 4:IV–317–IV–320.
- [55] Ho, Lam Si Tung, and Cécile Ané. 2013a. Asymptotic theory with hierarchical autocorrelation: Ornstein-Uhlenbeck tree models. *Annals of Statistics* 41(2):957–981.
- [56] ———. 2013b. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology* 63(3):397–408.
- [57] ———. 2014. Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution*. In review.
- [58] Hoef, Jay M. Ver, Erin Peterson, and David Theobald. 2006. Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics* 13(4):449–464.
- [59] Hoef, Jay M. Ver, and Erin E. Peterson. 2010. A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association* 105(489):6–18.
- [60] Huang, Hsin-Cheng, Noel Cressie, and John Gabrosek. 2002. Fast, resolution-consistent spatial prediction of global processes from satellite data. *Journal of Computational and Graphical Statistics* 11(1):63–88.
- [61] Hunt, Gene, Michael A. Bell, and Matthew P. Travis. 2008. Evolution toward a new adaptive optimum: Phenotypic evolution in a fossil stickleback lineage. *Evolution* 62:700–710.
- [62] Ibragimov, I. A., and Yu A. Rozanov. 1978. *Gaussian random processes*. New-York: Springer-Verlag.
- [63] Ikeda, N., and S. Watanabe. 1981. *Stochastic differential equations and diffusion processes*, vol. 24. Amsterdam-New York: North-Holland.

- [64] Ives, Anthony R., and Theodore Garland, Jr. 2010. Phylogenetic logistic regression for binary dependent variables. *Systematic Biology* 59(1):9–26.
- [65] Ives, Anthony R., Peter E. Midford, and Theodore Garland, Jr. 2007. Within-species variation and measurement error in phylogenetic comparative methods. *Systematic Biology* 56:252–270.
- [66] Ives, A.R., and M.R. Helmus. 2011. Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs* 81(3): 511–525.
- [67] Jetz, W, GH Thomas, JB Joy, K Hartmann, and AO Mooers. 2012. The global diversity of birds in space and time. *Nature* 491(7424):444–448.
- [68] King, Aaron A., and Marguerite A. Butler. 2009. *ouch: Ornstein-uhlenbeck models for phylogenetic comparative hypotheses*. (R package), <http://ouch.r-forge.r-project.org>.
- [69] Kingman, J. F. C. 1982a. The coalescent. *Stochastic Processes and their Applications* 13(3):235 – 248.
- [70] ———. 1982b. On the genealogy of large populations. *Journal of Applied Probability* 19(1982):27–43.
- [71] Lande, Russell. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain: Body size allometry. *Evolution* 33(1):402–416.
- [72] Lavergne, Sébastien, Arndt Hampe, and Juan Arroyo. 2013. In and out of africa: how did the strait of gibraltar affect plant species migration and local diversification? *Journal of Biogeography* 40(1):24–36.
- [73] Lavin, Shana R., William H. Karasov, Anthony R. Ives, Kevin M. Middleton, and Theodore Garland Jr. 2008. Morphometrics of the avian small intestine compared with that of nonflying mammals: a phylogenetic approach. *Physiological and Biochemical Zoology* 81(5):526–550.

- [74] Lawler, E.L. 1976. *Combinatorial optimization: Networks and matroids*. Holt, Rinehart and Winston.
- [75] Lawson, Charles L., and Richard J. Hanson. 1995. *Solving least squares problems*. Classics in Applied Mathematics, Society for Industrial and Applied Mathematics.
- [76] Liang, Jiming. 2007. *Linear and generalized linear mixed models and their applications*. Springer Series in Statistics, Springer.
- [77] Lynch, Michael. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution* 45(5):1065–1080.
- [78] Mardia, K. V., and R. J. Marshall. 1984. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* 71(1): 135–146.
- [79] Martins, Emilia P. 2000. Adaptation and the comparative method. *Trends in Ecology and Evolution* 15(7):296–299.
- [80] Martins, Emilia P., and Thomas F. Hansen. 1997. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist* 149(4):646–667.
- [81] Monteiro, L.R., and M.R. Nogueira. 2011. Evolutionary patterns and processes in the radiation of phyllostomid bats. *BMC evolutionary biology* 11(1): 137.
- [82] Moreau, Corrie S, Charles D Bell, Roger Vila, S Bruce Archibald, and Naomi E Pierce. 2006. Phylogeny of the ants: diversification in the age of angiosperms. *Science* 312(5770):101–104.
- [83] Mossel, E., and M. Steel. 2014. Majority rule has transition ratio 4 on yule trees under a 2-state symmetric model. Submitted.

- [84] Mossel, Elchanan, Sébastien Roch, and Allan Sly. 2013. Robust estimation of latent tree graphical models: Inferring hidden states with inexact parameters. *IEEE transactions on information theory* 59(7):4357–4373.
- [85] Motani, Ryosuke, and Lars Schmitz. 2011. Phylogenetic versus functional signals in the evolution of form-function relationships in terrestrial vision. *Evolution* 65(8):2245–2257.
- [86] O’Meara, Brian C., Cécile Ané, Michael J. Sanderson, and Peter C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. *Evolution* 60(5):922–933.
- [87] Pagel, Mark. 1997. Inferring evolutionary processes from phylogenies. *Zoologica Scripta* 26(4):331–348.
- [88] ———. 1999. Inferring the historical patterns of biological evolution. *Nature* 401(6756):877–884.
- [89] Paradis, Emmanuel, and Julien Claude. 2002. Analysis of comparative data using generalized estimating equations. *Journal of Theoretical Biology* 218(2): 175–185.
- [90] Pennell, Matthew W, and Luke J Harmon. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Annals of the New York Academy of Sciences* 1289(1):90–105.
- [91] Peres, Yuval. 1999. Probability on trees: An introductory climb. In *Lectures on probability theory and statistics*, ed. Pierre Bernard, vol. 1717 of *Lecture Notes in Mathematics*, 193–280. Springer Berlin Heidelberg.
- [92] Rabosky, Daniel L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PloS ONE* 9(2).
- [93] Rao, C.R., and VS Varadarajan. 1963. Discrimination of gaussian processes. *Sankhyā: The Indian Journal of Statistics, Series A* 303–330.

- [94] Revell, Liam J. 2009. Size-correction and principal components for interspecific comparative studies. *Evolution* 63(12):3258–3268.
- [95] Revell, Liam J., and Luke J. Harmon. 2008. Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. *Evolutionary Ecology Research* 10:311–331.
- [96] Revell, Liam J., D. L. Mahler, Pedro R. Peres-Neto, and Benjamin D. Redelings. 2012. A new phylogenetic method for identifying exceptional phenotypic diversification. *Evolution* 66(1):135–146.
- [97] Revell, Liam J., and R. Graham Reynolds. 2012. A new Bayesian method for fitting evolutionary models to comparative data with intraspecific variation. *Evolution* 66(9):2697–707.
- [98] Rigai, Guillem, E Lebarbier, and S Robin. 2012. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and computing* 22(4):917–929.
- [99] Rohlf, Rori V, Patrick Harrigan, and Rasmus Nielsen. 2014. Modeling gene expression evolution with an extended Ornstein-Uhlenbeck process accounting for within-species variation. *Molecular biology and evolution* 31(1): 201–211.
- [100] Scales, Jeffrey A, Aaron A King, and Marguerite A. Butler. 2009. Running for your life or running for your dinner: What drives fiber-type evolution in lizard locomotor muscles? *The American Naturalist* 173(5):543–553.
- [101] Schluter, Dolph, Trevor Price, Arne O. Mooers, and Donald Ludwig. 1997. Likelihood of ancestor states in adaptive radiation. *Evolution* 51(6):1699–1711.
- [102] Schwarz, Gideon. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6(2):461–464.

- [103] Semple, Charles, and Mike Steel. 2003. *Phylogenetics*. Oxford Lecture Series in Mathematics and its Applications 24, New York: Oxford University Press.
- [104] Shao, J. 1999. *Mathematical statistics*. New York: Springer.
- [105] Simon, Marcelo F, Rosaura Grether, Luciano P de Queiroz, Cynthia Skema, R Toby Pennington, and Colin E Hughes. 2009. Recent assembly of the cerrado, a neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. *Proceedings of the National Academy of Sciences* 106(48):20359–20364.
- [106] Slater, Graham J. 2013. Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the cretaceous-palaeogene boundary. *Methods in Ecology and Evolution* 4(8):734–744.
- [107] Slater, Graham J., and Luke J. Harmon. 2013. Unifying fossils and phylogenies for comparative analyses of diversification and trait evolution. *Methods in Ecology and Evolution* 4(8):699–702.
- [108] Smith, Stacey DeWitt, Cécile Ané, and David A Baum. 2008. The role of pollinator shifts in the floral diversification of *Iochroma* (Solanaceae). *Evolution* 62(4):793–806.
- [109] Smith, Stephen A., and Michael J. Donoghue. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science* 322(5898): 86–89.
- [110] Stein, Michael L. 1999. *Interpolation of spatial data: Some theory for kriging*. New York: Springer.
- [111] Sweeting, T. J. 1980. Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics* 8(6):1375–1381.
- [112] Thomas, Gavin H., Robert P. Freckleton, and Tamás Székely. 2006. Comparative analyses of the influence of developmental mode on phenotypic

- diversification rates in shorebirds. *Proceedings of the Royal Society B: Biological Sciences* 273(1594):1619–1624.
- [113] Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.
- [114] Venditti, Chris, Andrew Meade, and Mark Pagel. 2011. Multiple routes to mammalian diversity. *Nature* 479(7373):393–396.
- [115] Ying, Zhiliang. 1991. Asymptotic properties of a maximum likelihood estimator with data from a gaussian process. *Journal of Multivariate Analysis* 36(2):280–296.
- [116] Yule, G Udny. 1925. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B* 213:21–87.
- [117] Zhang, H., and D.L. Zimmerman. 2005. Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika* 92(4):921–936.
- [118] Zhang, Hao. 2004. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* 99(465):pp. 250–261.
- [119] Zhang, Nancy R, and David O Siegmund. 2007. A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* 63(1):22–32.
- [120] ———. 2012. Model selection for high-dimensional, multi-sequence change-point problems. *Statistica Sinica* 22(4):1507.
- [121] Zhao, Peng, and Bin Yu. 2006. On model selection consistency of lasso. *The Journal of Machine Learning Research* 7:2541–2563.