

# ASYMPTOTIC THEORY WITH HIERARCHICAL AUTOCORRELATION: ORNSTEIN-UHLENBECK TREE MODELS\*

BY LAM SI TUNG HO<sup>†</sup> AND CÉCILE ANÉ<sup>†,‡</sup>

<sup>†</sup>*Department of Statistics and* <sup>‡</sup>*Department of Botany, University of  
Wisconsin-Madison*

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.

## 1. Introduction and overview of main results.

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 1. Even from

---

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

*AMS 2000 subject classifications:* Primary 62F12, 62M10; secondary 62M30, 92D15, 92B10

*Keywords and phrases:* tree autocorrelation, dependence, microergodic, Ornstein-Uhlenbeck, evolution, phylogenetics

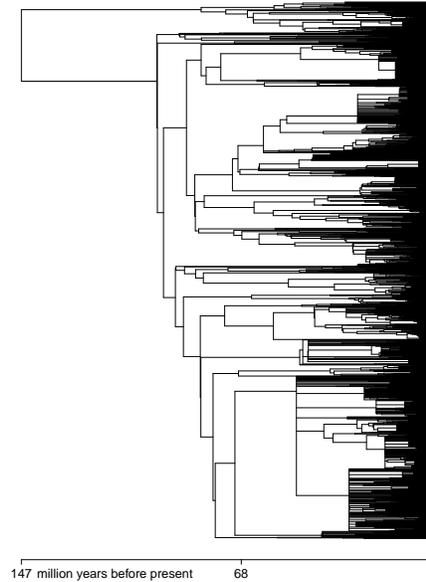


FIG 1. Family tree of 4,507 mammal species (Bininda-Emonds et al., 2007). Branch lengths indicate estimated diversification times on the horizontal axis. The Cretaceous/Tertiary mass extinction event marked the extinction of dinosaurs 65.5 million years ago. Cooper and Purvis (2010) used body mass data available for 77% of these species to infer the mode of evolution: neutral evolution (BM) versus natural selection (OU).

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 ( $\mu$ ) and a selection strength ( $\alpha$ ). 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.

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 evolution-

ary 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

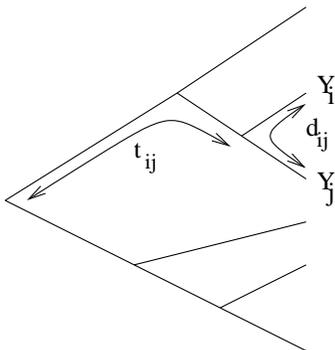


FIG 2. Correlation (or residual correlation) between observations at tips  $i$  and  $j$  are parametrized in the OU model as a function of the tree distance  $d_{ij}$  between  $i$  and  $j$  and of the length  $t_{ij}$  of their shared path from the root. For instance, [Cooper and Purvis \(2010\)](#) considered body mass ( $Y$ ) across 3,473 mammal species ( $i, j = 1 \dots 3473$ ).

variable (or of error term) along the tree results in normally distributed errors and in a covariance matrix governed by the tree, its branch lengths, and a single parameter  $\sigma^2$ . 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). Under the more complex OU process, changes towards a value  $\mu$  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),  $\alpha$  is the selection strength and  $B_t$  is a BM process. In what follows,  $\mu$  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  $\mu$ . In the biology literature,  $\mu$  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  $\alpha$  measures the strength of the pull

back to  $\mu$ . High  $\alpha$  values result in a process narrowly distributed around  $\mu$ , as expected under strong natural selection if the selective fitness of the trait is maximized at  $\mu$  and drops sharply away from  $\mu$ . 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  $\mu$  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  $\mu$  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  $\mu$  along the tree, varying linearly with explanatory variables (Hansen, Pienaar and Orzack, 2008). In our work, we develop an asymptotic theory for the simple case of a constant  $\mu$  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  $\mu$  and variance  $\gamma = \sigma^2/(2\alpha)$ . With this assumption, the observed process  $(Y_i)_{i \in \text{tips}}$  is Gaussian with mean  $\mu$  and variance matrix

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

where  $d_{ij}$  is the tree distance between tips  $i$  and  $j$ , i.e. the length of the path between  $i$  and  $j$ . Therefore, the strength  $\alpha$  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

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

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).

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,  $\sigma^2$  can be consistently estimated but  $\alpha$  cannot under an OU spatial autocorrelation model in dimension  $d \leq 3$  (Zhang, 2004). Recently,  $\alpha$  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.  $\alpha$ ). A parameter is microergodic when two different values for it lead to orthogonal distributions for the complete, asymptotic process (Stein, 1999).

In section 2, we prove that the mean  $\mu$  is non-microergodic under the OU autocorrelation framework, and we provide a lower bound for the variance of the MLE of  $\mu$ . We also give a sufficient condition for the microergodicity of the OU covariance parameters  $\alpha$  and  $\sigma^2$  (or  $\gamma$ ) based on the distribution of internal node ages. The microergodic covariance parameter under spatial infill asymptotics with OU autocorrelation,  $\sigma^2$ , 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 3, we illustrate this theoretically for a symmetric tree asymptotic framework, where we show that the REML estimator of  $\alpha$  converges at a slower rate than that of the generally microergodic parameter. We also illustrate that the ML estimate convergence rate of  $\alpha$  is slower than that of  $\sigma^2$ , 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 (1), because the model matrix is not of full rank under model (2) on an ultrametric tree.

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, Cressie and Gabrosek \(2002\)](#) 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, Peterson and Theobald, 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. 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  $P_{\theta_1}$  and  $P_{\theta_2}$  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  $d \leq 3$ ,  $\alpha$  and  $\gamma$  are not microergodic even though  $\sigma^2$  is ([Zhang, 2004](#)) and the MLE of  $\sigma^2$  is strongly consistent ([Ying, 1991](#)). Also note that the microergodicity of  $(\gamma, \alpha)$  is equivalent to the microergodicity of both  $\gamma$  and  $\alpha$ .

*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  $\sigma$ -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  $\sigma$ -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 \(1978\)](#):

$$(3) \quad r(\mathcal{U}') \leq r(\mathcal{U}'') \quad \text{for } \mathcal{U}' \subset \mathcal{U}''.$$

Consider non singular Gaussian measures  $P_1$  and  $P_2$  on the  $\sigma$ -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:

$$(4) \quad P_1 \perp P_2 \Leftrightarrow r_n \rightarrow \infty, \quad \text{and} \quad P_1 \equiv P_2 \Leftrightarrow r_n \rightarrow r < \infty.$$

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

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

[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  $P_2$ .

$$(5) \quad \lim_{n \rightarrow \infty} \sum_{j=1}^n (\sigma_{j,n}^2 - 1)^2 = \infty.$$

**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 (1) and (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  $\mathcal{S}^{\mathbb{T}_n}$  as the set of all internal nodes of tree  $\mathbb{T}_n$  (including the root) and  $\mathcal{S} = \bigcup_{n=1}^{\infty} \mathcal{S}^{\mathbb{T}_n}$ . Let  $(T_i)_{i \in \mathcal{S}}$  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.  $\mathcal{S}^{\mathbb{T}_n}$  is a subset of  $\mathcal{S}^{\mathbb{T}_{n+1}}$  so  $(T_i)_{i \in \mathcal{S}}$  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  $(T_i)_{i \in \mathcal{S}}$  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  $\mathcal{S}^{\mathbb{T}_n}$  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 2.4 below state general results on the microergodicity of parameters in OU tree models. Our main tool is the equivalence (4) applied to  $r_n = r(\mathbb{T}_n)$ , the entropy distance between  $P_{\theta_1}$  and  $P_{\theta_2}$  for two parameter sets  $\theta_k = (\mu_k, \alpha_k, \gamma_k)$ ,  $k = 1, 2$ , on the  $\sigma$ -algebra generated by  $(Y_j)_{j=1}^n$ .

### 2.3. Microergodicity of the mean $\mu$ .

**THEOREM 2.1.** *Under OU model (1) and condition (C),  $\mu$  is not microergodic.*

The theorem follows directly from (4) and the boundedness of  $(T_i)_{i \in \mathcal{S}}$  once the following upper bound is established:

$$(6) \quad r(\mathbb{T}) \leq (\mu_1 - \mu_2)^2 / (\gamma_1 e^{-2\alpha_1 T})$$

if  $\alpha_1 = \alpha_2$  and  $\gamma_1 = \gamma_2$ , where  $T$  is the age of the root of  $\mathbb{T}$  (appendix B.2). One consequence is that there is no consistent estimator for  $\mu$ . To illustrate this, we consider the MLE of  $\mu$  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 3).

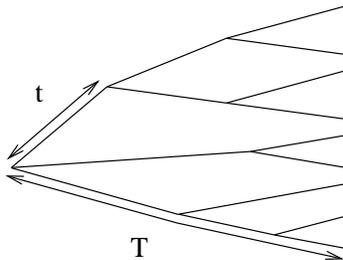


FIG 3. Ultrametric tree with all tips at equal distance  $T$  from the root. The root has  $k = 3$  children here, and  $t$  is the minimum distance from the root to its children.

**THEOREM 2.2.** *Assume OU model (1) on an ultrametric tree. Let  $\hat{\mu}$  be the MLE of  $\mu$  conditional on some possibly wrong value  $\alpha_*$  of  $\alpha$ . Then*

$$(7) \quad \text{var}(\hat{\mu}) \geq \frac{\sigma^2}{2\alpha} e^{-2\alpha T} \left( 1 + \frac{e^{2\alpha t} - 1}{k} \right).$$

*The equality holds if and only if  $\alpha$  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.*

The second part of the theorem follows directly from the lower bound (7). Note that  $\hat{\mu}$  is Gaussian with mean  $\mu$ . Therefore, the lower bound of its variance implies that  $\hat{\mu}$  cannot converge to  $\mu$ . Hence, it is not consistent.

The assumption that  $\alpha > 0$  is trivial. When  $\alpha = 0$ , the OU process reduces to a BM where  $\mu$  has no influence on the process. In that case,  $\mu$  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  $\alpha$ . The precision of  $\hat{\mu}$  is weakest when autocorrelation is strong, i.e. when  $\alpha$  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) predicts unequal variances and most importantly unequal means at the tips. Such trees can carry more information about  $\mu$ . Consider for instance the star tree in figure 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 B.2), thus providing a counterexample to theorem 2.2 when the ultrametric assumption is violated.

**Proof of Theorem 2.2.** To prove (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

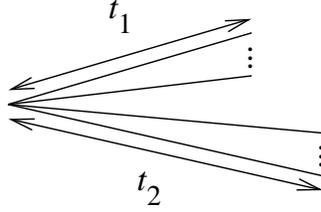


FIG 4. Example of a non-ultrametric tree on which  $\mu$  can be consistently estimated.

$\frac{\sigma^2}{2\alpha}(\mathbf{1}^t V_\alpha^{-1} \mathbf{1})^{-1}$  when  $\alpha_* = \alpha$  is known. Its variance is larger when  $\alpha$  is unknown, by the Gauss-Markov theorem. For this reason, we only need to prove the following lemma (which is done in the appendix B.2).

LEMMA 2.3. 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, (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.

#### 2.4. Microergodicity of the autocorrelation parameter $(\gamma, \alpha)$ .

THEOREM 2.4. Under OU model (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  $(\gamma, \alpha)$  is microergodic. Note that this condition is satisfied if  $(T_i)_{i \in \mathcal{I}}$  has 2 or more limit points.

**Proof of Theorem 2.4.** 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 B.2.

LEMMA 2.5. *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  $\mathcal{S}_S^\mathbb{T} = \{i : T_i \in S, i \in \mathcal{I}^\mathbb{T}\}$  the set of internal nodes of  $\mathbb{T}$  whose ages lie in  $S$ . Let  $(\gamma_1, \alpha_1)$  and  $(\gamma_2, \alpha_2)$  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 B.1 (in appendix B.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  $|\mathcal{C}_n| \geq \frac{1}{2} |\mathcal{S}_{(t_0 - \epsilon_\delta, t_0 + \epsilon_\delta)}^{\mathbb{T}_n}|$ . Let  $r(\mathcal{C}_n)$  be the entropy distance between  $P_{\theta_1}$  and  $P_{\theta_2}$  on the  $\sigma$ -algebra generated by  $\mathcal{C}_n$ . By (3) and direct calculation,

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

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

**Proof of part (b)** First, we consider the case when  $(T_i)_{i \in \mathcal{I}}$  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,  $(\gamma, \alpha)$  is microergodic by the following lemma (proved in appendix B.2):

LEMMA 2.6. *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  $(T_i)_{i \in \mathcal{I}}$  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  $\mathcal{S}_{(t_0, \infty)} = \bigcup_{n=1}^{\infty} \mathcal{S}_{(t_0, \infty)}^{\mathbb{T}_n}$  and  $\mathcal{S}_{[0, t_0]} = \bigcup_{n=1}^{\infty} \mathcal{S}_{[0, t_0]}^{\mathbb{T}_n}$ . The condition in (b) implies that  $\sum_{i \in \mathcal{S}_{(t_0, \infty)}} (T_i - t_0)^2 = \infty$  or  $\sum_{i \in \mathcal{S}_{[0, t_0]}} (T_i - t_0)^2 = \infty$  or both. We now use lemma B.2 (appendix

B.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 (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  $(\gamma, \alpha)$  is microergodic.  $\square$

Theorem 2.4 part (b) gives a very general sufficient condition ensuring the microergodicity of  $(\gamma, \alpha)$ . 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 5). We give below 3 examples in which  $(T_i)_{i \in \mathcal{I}}$  has only one limit point  $t_0$  and the condition in theorem 2.4 part (b) is violated. Two examples illustrate the non-microergodicity of  $(\gamma, \alpha)$ , one in which  $t_0 > 0$  and one in which  $t_0 = 0$ . In the last example the condition in (b) is violated, yet  $(\gamma, \alpha)$  is microergodic.

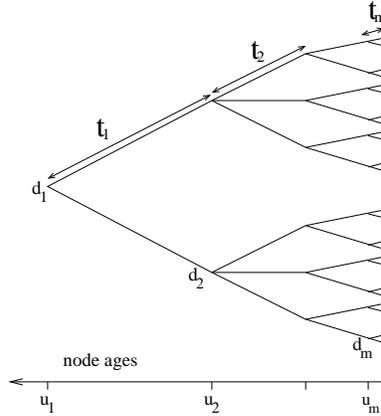


FIG 5. Symmetric trees with  $m = 4$  levels.

**THEOREM 2.7.** *Consider the OU model (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  $(\gamma, \alpha)$  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 2.4 (b).*
  - (i) *If  $t_0 = 0$  then  $(\gamma, \alpha)$  is not microergodic.*
  - (ii) *If  $t_0 > 0$  then  $(\gamma, \alpha)$  is microergodic.*

We discuss here the key ingredients of the proof. The technical details are provided in appendix B.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). In (a), only the multiplicity of the smallest eigenvalue increases to infinity when the tree grows. If  $(\gamma_1, \alpha_1)$  and  $(\gamma_2, \alpha_2)$  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  $t_0 > 0$ .

**3. Different convergence rates of ML estimators for different microergodic parameters.** Section 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 2.4 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  $\sigma^2$  is expected to converge at a faster rate than that of  $\alpha$ , which might not even be consistent. Here we identify cases with unequal convergence rates both theoretically and empirically.

3.1. *Faster convergence of the REML estimator for  $f_{t_0}(\gamma, \alpha)$  than for  $\alpha$  and  $\gamma$ .* We focus here on the symmetric tree growth model from Theorem 2.7 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  $\gamma$  and  $\alpha$ . We show that the REML estimator of  $(\gamma, \alpha)$  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  $\alpha$  or  $\gamma$ , which have stronger requirements to be microergodic.

THEOREM 3.1. *Consider the asymptotic growth model from above with OU model (1). Denote  $\nu = \gamma(1 - e^{-2\alpha t_m})$ . Then the REML estimator  $(\hat{\nu}, \hat{\alpha})$  is consistent and  $\begin{pmatrix} \sqrt{\tilde{n}}(\hat{\nu} - \nu) \\ \sqrt{\tilde{n}}(\hat{\alpha} - \alpha) \end{pmatrix} \xrightarrow{d} N\left(\mathbf{0}, \begin{pmatrix} 8\nu^2 & 0 \\ 0 & v_\alpha \end{pmatrix}\right)$ .*

Moreover, if  $n/\tilde{n} = d_m$  converges to infinity, then  $\sqrt{\tilde{n}}(\hat{\gamma} - \gamma, \hat{\alpha} - \alpha)^\dagger$  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 B.3 gives the expression for  $v_\alpha$ . 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  $\hat{\nu}$  is  $n^{-1/2}$ , which can be much faster than  $\tilde{n}^{-1/2}$ .

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 6 shows the distribution of node ages for this tree, and for a symmetric tree with dense sampling near the tips described in Theorem 2.7 (b), on which  $\alpha$  and  $\gamma$  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,  $\sigma^2$  is microergodic while  $(\gamma, \alpha)$  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  $\sigma^2$  to converge quickly, and the MLE of  $\alpha$  to converge more slowly or not at all.

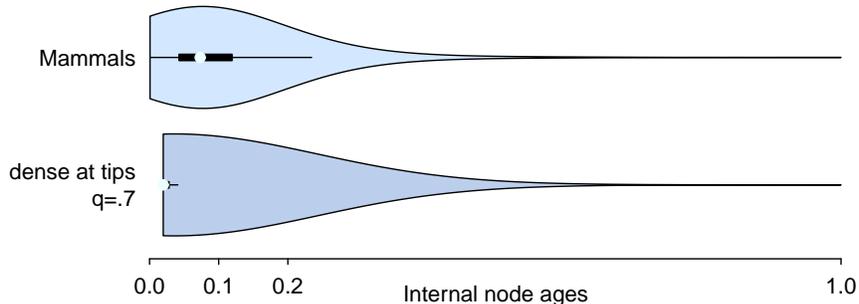


FIG 6. Distribution of node ages in the mammal tree (top) and in a symmetric tree (bottom) of similar size  $n = 2^{12} = 4096$  with  $d = 2$  at each level, levels being added near the tips at ages  $q^m$ . The value  $q = .7 \approx 2^{-1/2}$  is the largest at which  $\alpha$  and  $\gamma$  are not microergodic.

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 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 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 1). Also, the lower bound for the variance of  $\hat{\mu}$  is very close to the true variance (table 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 (7)	0.8517	0.8472	0.8469	0.8468	0.8468	0.8468

TABLE 1

Correlation between  $\log \hat{\alpha}$  and  $\log \hat{\gamma}$  and variance of  $\hat{\mu}$  from simulations. Last line: value of theoretical bound (7) for  $\text{var}(\hat{\mu})$ , averaged over 20 simulation subtrees.

**4. Discussion.** We considered an Ornstein-Uhlenbeck model of hierarchical autocorrelation and showed that the location parameter, here the mean  $\mu$ , 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

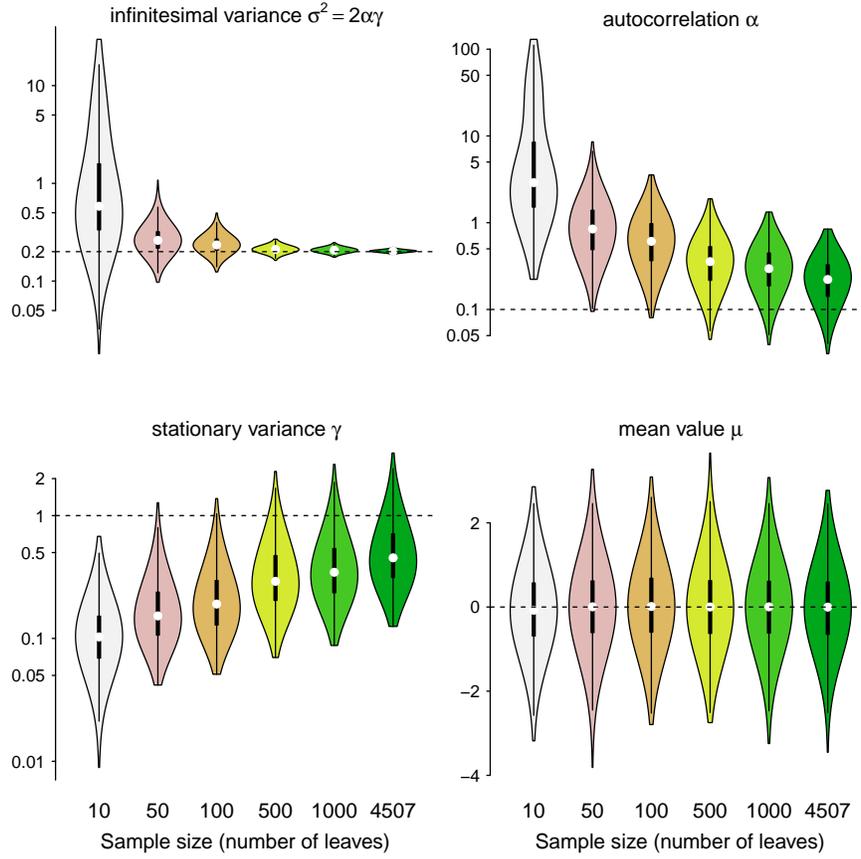


FIG 7. Violin plots showing the distribution of the MLE of  $\mu$ ,  $\gamma$ ,  $\alpha$ , and  $\sigma^2 = 2\alpha\gamma$  on trees subsampled from the mammal phylogeny in *Bininda-Emonds et al. (2007)* with 2000 simulations at each sample size. The true values were  $\mu = 0$ ,  $\gamma = 1$ ,  $\alpha = 0.1$ , and  $\sigma^2 = 0.2$ .

shift, i.e. a change in  $\mu$  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,  $\sigma^2$  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  $\sigma_e^2$  is called a nugget effect in spatial statistics. Measurement error with tree-structured correlation is rarely accounted for in applications (but see [Ives, Midford and Garland, 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 2.6, we can construct a consistent estimator for  $(\gamma, \alpha)$ . 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  $(\gamma, \alpha)$  is microergodic, the REML estimator of  $(\gamma, \alpha)$  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  $\alpha$  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  $\sigma^2$  can be consistently estimated but  $\alpha$  and  $\gamma$  cannot. On the real tree, the MLE of  $\sigma^2$  converges quickly to the true value while that of  $\alpha$  and  $\gamma$  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.

#### APPENDIX A: SPECTRAL DECOMPOSITION OF THE OU COVARIANCE MATRIX ON SYMMETRIC TREES

We consider here symmetric trees (figure 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 (1), the correlation matrix  $\mathbf{V}_\alpha$  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  $\mathbf{V}_\alpha(\mathbf{t}) = \mathbf{V}_\alpha^{\text{BM}}(\mathbf{t}^{\text{BM}})$  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  $\infty$ . Furthermore, Ané (2008) showed that the eigenvectors of  $\mathbf{V}_\alpha^{\text{BM}}$  are independent of the tree's branch lengths, which implies here that the eigenvectors of  $\mathbf{V}_\alpha$  are independent of  $\alpha$ . 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  $\lambda_0$ . This eigenvector is  $\mathbf{1}$  and has multiplicity 1.

## APPENDIX B: SUPPORTING LEMMAS AND TECHNICAL PROOFS

### B.1. Procedures for choosing independent contrasts.

LEMMA B.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  $\mathcal{I}_{(a,b)}^{\mathbb{T}}$  such that  $|\mathcal{C}| \geq \frac{1}{2} |\mathcal{I}_{(a,b)}^{\mathbb{T}}|$ .*

**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  $i_n \in \mathcal{I}_{(a,b)}^{\mathbb{T}_{n-1}}$  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  $\mathcal{I}_{(a,b)}^{\mathbb{T}_n} = \emptyset$ . The procedure guarantees that the paths do not intersect, hence the contrasts are independent. Furthermore,  $\mathcal{I}_{(a,b)}^{\mathbb{T}_n} = \mathcal{I}_{(a,b)}^{\mathbb{T}_{n-1}} \setminus \{i_n, i'_n\}$  where  $i'_n$  is the parent of  $i_n$ , so  $|\mathcal{C}| \geq |\mathcal{I}_{(a,b)}^{\mathbb{T}}|/2$ .

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

*Proof of lemma B.2.* (a) The procedure in the proof of lemma B.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  $\mathcal{S}_{[0,t]}^{\mathbb{T}} \subset \bigcup_{k=1}^m \{i_k, i'_k\}$ , hence

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

(b) Contrasts are chosen by induction, starting with  $\mathcal{C} = \emptyset$ . Let  $r^{\mathbb{T}}$  be the root of  $\mathbb{T}$ . If  $r^{\mathbb{T}} \notin \mathcal{S}_{(t,T]}^{\mathbb{T}}$  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 8). 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  $\leq 3$  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  $r^{\mathbb{T}_k} \in \mathcal{S}_{(t,T]}^{\mathbb{T}}$ . 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  $\mathcal{S}_{(t,T]}^{\mathbb{T}} \subset \{r^{\mathbb{T}}, i_1, i_2\} \cup \bigcup_{k=1}^l \mathcal{S}_{(t,T]}^{\mathbb{T}_k} \cup \{s_k\}$ , and  $\mathcal{C} = \{r^{\mathbb{T}}\} \cup \bigcup_{k=1}^l \mathcal{C}_k$ . Therefore:

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

## B.2. Technical proofs for section 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  $\mathbf{V}_\alpha$  can be expressed in terms of the  $n/2 \times n/2$  identity matrix  $\mathbf{I}$  as  $\mathbf{V}_\alpha = \text{diag}((1-a^2)\mathbf{I}, (1-b^2)\mathbf{I}) + (a\mathbf{1}^t, b\mathbf{1}^t)^t (a\mathbf{1}^t, b\mathbf{1}^t)$ .

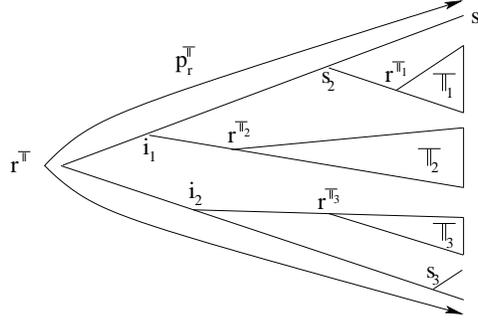


FIG 8. Recursive construction of independent contrasts, taken with respect to the root at each step.

We then get  $\mathbf{V}_\alpha^{-1}$  using Woodbury formula, then

$$\mathbf{1}^t \mathbf{V}_\alpha^{-1} \mathbf{1} = n \left( \frac{1}{1-a^2} + \frac{1}{1-b^2} + \frac{n(a-b)^2}{(1-a^2)(1-b^2)} \right) / \left( 1 + \frac{na^2}{1-a^2} + \frac{nb^2}{1-b^2} \right).$$

If  $t_1 \neq t_2$  then  $a \neq b$  and  $\text{var}(\hat{\mu}) = (\mathbf{1}^t \mathbf{V}_\alpha^{-1} \mathbf{1})^{-1}$  goes to 0 as claimed.

*Proof of lemma 2.3.* We will first prove  $\mathbf{V}_\alpha \geq e^{-2\alpha T} \mathbf{J}_n$  by induction on the number of tips, where  $\mathbf{J}_n = \mathbf{1}\mathbf{1}^t$ . 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  $\mathbf{V}_\alpha^{(1)}, \dots, \mathbf{V}_\alpha^{(k)}$  associated with these subtrees must satisfy  $\mathbf{V}_\alpha^{(i)} \geq e^{-2\alpha T_i} \mathbf{J}_{n_i}$ . Therefore  $\mathbf{V}_\alpha - e^{-2\alpha T} \mathbf{J}_n \geq \text{diag}(\mathbf{V}_\alpha^{(i)} - e^{-2\alpha T_i} \mathbf{J}_{n_i}) \geq 0$  is true for all trees. Now we use the definition of  $t$  and go a step further using that  $\mathbf{V}_\alpha^{(i)} - e^{-2\alpha T} \mathbf{J}_{n_i} \geq (e^{-2\alpha T_i} - e^{-2\alpha T}) \mathbf{J}_{n_i} \geq (e^{-2\alpha(T-t)} - e^{-2\alpha T}) \mathbf{J}_{n_i}$  for all  $i = 1, \dots, k$ . This implies that  $\mathbf{V}_\alpha - e^{-2\alpha T} \mathbf{J}_n \geq (e^{-2\alpha(T-t)} - e^{-2\alpha T}) \text{diag}(\mathbf{J}_{n_1}, \dots, \mathbf{J}_{n_k}) \geq \frac{1}{k} (e^{-2\alpha(T-t)} - e^{-2\alpha T}) \mathbf{J}_n$ , from which lemma 2.3 follows easily.

*Proof of upper bound (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  $P_{\theta_1}$  and  $P_{\theta_2}$ , 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 2.3.

*Proof of lemma 2.5.* First,  $\text{var}(C_i^{p_i}) = \text{var}(Y_1^i) + \text{var}(Y_2^i) - 2\text{cov}(Y_1^i, Y_2^i) = 2\gamma - 2\gamma e^{-2\alpha T_i}$ . 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 2.6.* 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  $h_3(x) = x / (1 - e^{-2xt_2})$  is monotone on  $(0, \infty)$ .

*Proof of theorem 2.7 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  $\lambda_k$  with their multiplicities and  $\mathbf{P}$  contains the eigenvectors of  $\mathbf{V}_\alpha$ , which do not depend of  $\alpha$  (appendix A). Then  $\mathbf{h}$  is orthonormal under  $P_{\theta_1}$ , and orthogonal under  $P_{\theta_2}$  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  $P_{\theta_1}$  and  $P_{\theta_2}$  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 2.7 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:

$$(8) \quad \frac{\lambda_{k,m}(\alpha)}{d^{m-k}} = \sum_{j=0}^{m-k-1} \frac{e^{-2\alpha q^{k+1+j}} - e^{-2\alpha q^{k+j}}}{d^j} + \frac{1 - e^{-2\alpha q^m}}{d^{m-k}}.$$

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  $\alpha$  and  $k$ . To prove the convergence of  $r(\mathbb{T}_m)$  we will need the following lemma, which is proved later.

LEMMA B.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  $\alpha_1, \alpha_2, d$  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 2.4 part a), we can assume  $\gamma_1\alpha_1 = \gamma_2\alpha_2$ . Lemma B.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 (6). This completes the proof.

We now turn to case (ii) with  $t_0 > 0$ . To prove that  $(\gamma, \alpha)$  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 (5). Note that there exists  $w > 0$  such that  $d^w q \geq 1$ . Denote  $k_m = [m/(w+1)]$  where  $[x]$  is a largest integer smaller than  $x$ . The condition in (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  $\alpha$ .

*Proof of lemma B.3.* We first note that for every  $a > 0$  there exists  $x_a > 0$  such that  $e^{-ax} - (1 - ax + a^2x^2/2) = O(a^3x^3)$  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 (8):  $\lambda_{k,m}(\alpha)/d^{m-k} - 2\alpha(1-q)q^k f_1 + 2\alpha^2(1-q^2)q^{2k} f_2 = q^{3k} f_3 O_\alpha(1)$  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  $\frac{\lambda_{k,m}(\alpha_1)}{\lambda_{k,m}(\alpha_2)} = \frac{\alpha_1}{\alpha_2} \left( 1 + (\alpha_2 - \alpha_1)(1+q)q^k f_2/f_1 \right) + q^{2k}O(1)$  where the  $O(1)$  term is bounded uniformly in  $m > k \geq K$ , and the same formula holds when  $\alpha_2$  and  $\alpha_1$  are switched. Lemma B.3 then follows immediately because we assume that  $\gamma_1\alpha_1 = \gamma_2\alpha_2$ .

### B.3. Technical proofs for section 3.

*Criterion for the consistency and asymptotic normality of REML estimators.* In appendix A, we showed that  $\mathbf{1}$  is an eigenvector of  $\mathbf{V}_\alpha$  for symmetric trees, independently of  $\alpha$ . Therefore, the REML estimator of  $(\gamma, \alpha)$  based on  $\mathbf{Y}$  is the ML estimator of  $(\gamma, \alpha)$  based on the transformed data  $\tilde{\mathbf{Y}} = \tilde{\mathbf{P}}^t \mathbf{Y}$  where  $\tilde{\mathbf{P}}$  is the matrix of all eigenvectors but  $\mathbf{1}$ .  $\tilde{\mathbf{Y}}$  is Gaussian centered with variance  $\Sigma_n = \gamma \tilde{\mathbf{\Lambda}}$  where  $\tilde{\mathbf{\Lambda}}$  is the diagonal matrix of all eigenvalues of  $\mathbf{V}_\alpha$  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  $\mathbf{A}_n(\theta)$  such that

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

Under these conditions, the MLE  $\hat{\theta}$  satisfies  $\mathbf{A}_n(\theta)(\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathbf{W}(\theta)^{-1})$ . A standard choice for  $\mathbf{A}_n$  is the inverse of the square-root of the Fisher information matrix  $\mathbf{B}_n = \mathbb{E}(\mathcal{J}_n)$ . 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  $\mathbf{A}_n = \mathbf{B}_n^{1/2}$  and  $\mathbf{W}(\theta) = \mathbf{I}$  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  $\mathbf{B}_n^{-1}$  and  $(\Sigma_n^{-1})_{ij}$  is the  $(i, j)$ -second order derivative of  $\Sigma_n^{-1}$ .

*Proof of Theorem 3.1.* It is convenient here to re-parametrize the model using  $(\nu, \alpha)$ . The diagonal elements in  $\Sigma_n$  are  $\nu \lambda_k(\alpha) / \lambda_m(\alpha)$  with multiplicity  $d_1 \dots d_{k-1} (d_k - 1)$ . The smallest is  $\nu$  (for  $k = m$ ) with multiplicity  $n - \tilde{n}$ , which is conveniently independent of  $\alpha$ . 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 B.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 B.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

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

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

$$\frac{A_n e^{-\alpha A_n} - A_{n-1} e^{-\alpha A_{n-1}}}{e^{-\alpha A_{n-1}} - e^{-\alpha A_n}} > \frac{A_{n+1} e^{-\alpha A_{n+1}} - A_n e^{-\alpha A_n}}{e^{-\alpha A_n} - e^{-\alpha A_{n+1}}}.$$

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 B.4 is easy and left to the reader. The following lemma results directly from lemma B.4.

LEMMA B.5. *With  $m$  fixed and parametrization  $(\nu, \alpha)$ , 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  $\det \mathbf{B}_n \sim 2n\tilde{n}/(\nu^2 v_\alpha)$  with  $v_\alpha$  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  $(\hat{\nu}, \hat{\alpha})$  follows from applying lemma B.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  $\delta$ -method. Its proof is similar to that of the classical  $\delta$ -method (Shao, 1999) and is left to the reader.

LEMMA B.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  $\delta$ -method and the fact that  $\sqrt{\tilde{n}}(\hat{\nu} - \nu)$  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  $\tilde{n} = o(n)$ .

## REFERENCES

- ALDOUS, D. J. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Stat. Sci.* **16** 23–34.
- ANDERES, E. (2010). On the consistent separation of scale and variance for Gaussian random fields. *The Annals of Statistics* **38** 870–893.
- ANÉ, C. (2008). Analysis of Comparative Data with Hierarchical Autocorrelation. *Annals of Applied Statistics* **2** 1078–1102.

- BININDA-EMONDS, O., CARDILLO, M., JONES, K. E., MACPHEE, R. D. E., BECK, R. M. D., GRENYER, R., PRICE, S. A., VOS, R. A., GITTLEMAN, J. L. and PURVIS, A. (2007). The delayed rise of present-day mammals. *Nature* **446** 507-512.
- BUTLER, M. A. and KING, A. A. (2004). Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *The American Naturalist* **164** 683-695.
- COOPER, N. and PURVIS, A. (2010). Body Size Evolution in Mammals: Complexity in Tempo and Mode. *The American Naturalist* **175** 727-738.
- CRESSIE, N. and LAHIRI, S. N. (1993). The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis* **45** 217-233.
- CRESSIE, N. and LAHIRI, S. N. (1996). Asymptotics for REML estimation of spatial covariance parameters. *Journal of Statistical Planning and Inference* **50** 327-341.
- CRESSIE, N., FREY, J., HARCH, B. and SMITH, M. (2006). Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics* **11** 127-150.
- HANSEN, T. F. and MARTINS, E. P. (1996). Translating Between Microevolutionary Processes and Macroevolutionary Patterns: The Correlation Structure of Interspecific Data. *Evolution* **50** 1404-1417.
- HANSEN, T. F., PIENAAR, J. and ORZACK, S. H. (2008). A Comparative Method for Studying Adaptation to a Randomly Evolving Environment. *Evolution* **62** 1965-1977.
- HERSHEY, J. R. and OLSEN, P. A. (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.
- HOEF, J. M. V., PETERSON, E. and THEOBALD, D. (2006). Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics* **13** 449-464.
- HOEF, J. M. V. and PETERSON, E. E. (2010). A Moving Average Approach for Spatial Statistical Models of Stream Networks. *Journal of the American Statistical Association* **105** 6-18.
- HUANG, H.-C., CRESSIE, N. and GABROSEK, J. (2002). Fast, Resolution-Consistent Spatial Prediction of Global Processes From Satellite Data. *Journal of Computational and Graphical Statistics* **11** 63-88.
- IBRAGIMOV, I. A. and ROZANOV, Y. A. (1978). *Gaussian random processes*. Springer-Verlag, New York.
- IKEDA, N. and WATANABE, S. (1981). *Stochastic Differential Equations and Diffusion Processes* **24**. North-Holland, Amsterdam-New York.
- IVES, A. R., MIDFORD, P. E. and GARLAND, T. JR. (2007). Within-Species Variation and Measurement Error in Phylogenetic Comparative Methods. *Systematic Biology* **56** 252-270.
- KINGMAN, J. F. C. (1982a). The coalescent. *Stochastic Processes and their Applications* **13** 235 - 248.
- KINGMAN, J. F. C. (1982b). On the genealogy of large populations. *Journal of Applied Probability* **19** 27-43.
- LANDE, R. (1979). Quantitative Genetic Analysis of Multivariate Evolution, Applied to Brain: Body Size Allometry. *Evolution* **33** 402-416.
- LAVIN, S. R., KARASOV, W. H., IVES, A. R., MIDDLETON, K. M. and JR., T. G. (2008). Morphometrics of the avian small intestine compared with that of nonflying mammals: a phylogenetic approach. *Physiological and Biochemical Zoology* **81** 526-550.
- MARDIA, K. V. and MARSHALL, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71** 135-146.
- MONTEIRO, L. R. and NOGUEIRA, M. R. (2011). Evolutionary patterns and processes in the radiation of phyllostomid bats. *BMC evolutionary biology* **11** 137.
- RAO, C. R. and VARADARAJAN, V. (1963). Discrimination of Gaussian processes. *Sankhyā*:

- The Indian Journal of Statistics, Series A* 303–330.
- SHAO, J. (1999). *Mathematical Statistics*. New York: Springer.
- STEIN, M. L. (1999). *Interpolation of Spatial Data: Some theory for Krigging*. Springer, New York.
- SWEETING, T. J. (1980). Uniform Asymptotic Normality of the Maximum Likelihood Estimator. *The Annals of Statistics* **8** 1375–1381.
- YING, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis* **36** 280–296.
- YULE, G. U. (1925). A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London, Series B* **213** 21–87.
- ZHANG, H. (2004). Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics. *Journal of the American Statistical Association* **99** pp. 250–261.
- ZHANG, H. and ZIMMERMAN, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika* **92** 921–936.

1300 UNIVERSITY AVE., MADISON, WISCONSIN 53706, USA  
E-MAIL: [lamho@stat.wisc.edu](mailto:lamho@stat.wisc.edu), [ane@stat.wisc.edu](mailto:ane@stat.wisc.edu)