Detection of adaptive shifts on phylogenies using shifted stochastic processes on a tree

Paul Bastide[1,2], Mahendra Mariadassou [2] & Stéphane Robin [1]

[1] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France*
[2] *MaIAGE, INRA, Université Paris-Saclay, 78352 Jouy-en-Josas, France*
paul.bastide@agroparistech.fr, mahendra.mariadassou@jouy.inra.fr,
stephane.robin@agroparistech.fr

Summary. Comparative and evolutive ecologists are interested in the distribution of quantitative traits among related species. The classical framework for these distributions consists of a random process running along the branches of a phylogenetic tree relating the species. We consider shifts in the process parameters, which reveal fast adaptation to changes of ecological niches. We show that models with shifts are not identifiable in general. Constraining the models to be parsimonious in the number of shifts partially alleviates the problem but several evolutionary scenarios can still provide the same joint distribution for the extant species. We provide a recursive algorithm to enumerate all the equivalent scenarios and to count the number of effectively different scenarios. We introduce an incomplete-data framework and develop a maximum likelihood estimation procedure based on the EM algorithm. Finally, we propose a model selection procedure, based on the cardinal of effective scenarios, to estimate the number of shifts and for which we prove an oracle inequality.

Keywords. Random process on tree, Ornstein-Uhlenbeck process, Change-point detection, Adaptive shifts, Phylogeny, Model selection

1 Introduction

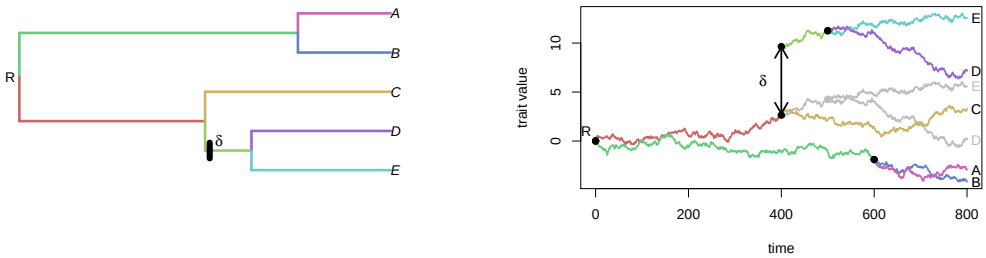
1.1 Motivations: Environmental Shifts

An important goal of comparative and evolutionary biology is to decipher the past evolutionary mechanisms that shaped present day diversity, and more specifically to detect the dramatic changes that occurred in he past (see for instance Losos, 1990; Mahler et al., 2013; Davis et al., 2007; Jaffe et al., 2011). It is well established that related organisms do not evolve independently (Felsenstein, 1985): their shared evolutionary history is well represented by a phylogenetic tree. In order to explain the pattern of traits measured on a set of related species, one needs to take these correlations into account. Indeed, a given species will be more likely to have a similar trait value to her "sister" (a closely related species) than to her "cousin" (a distantly related species), just because of the structure of the tree. On top of that structure, when considering a *functional* trait (*e.g.* a trait directly linked to the fitness of its bearer), such as shell size for turtles (Jaffe et al., 2011), one needs to take into account the effect of the species environment on its traits. Indeed, a change in the environment for a subset of species, like a move to the Galàpagos Islands for turtles, will affect the observed trait distribution, here with a shift towards giant shell sizes compared to mainland turtles. The observed present-day trait distribution hence contains the footprint of adaptive events, and should allow us to detect unobserved past events, like the migration of one ancestral species to a new environment. Our goal here is to devise a statistical method based on a rigorous maximum likelihood framework to automatically detect the past environmental shifts that shaped the present day trait distribution.

1.2 Stochastic Process on a tree

We model the evolution of a quantitative adaptive trait using the framework of stochastic processes on a tree. Specifically, given a rooted phylogenetic tree, we assume that the trait evolves according to a given stochastic process on each branch of the tree. At each speciation event, or equivalently node of the tree, one independent copy with the same initial conditions and parameters is created for each daughter species, or outgoing branches.

Tree Structure. This model is our null model: it accounts for the tree-induced distribution of trait values in the absence of shifts. Depending on the phenomenon studied, several stochastic processes can be used to capture the dynamic of the trait evolution. In the following, we will use the Brownian Motion (BM) and the Ornstein-Uhlenbeck (OU) processes.



(a) A phylogenetic tree.

(b) Trait value evolution.

Figure 1: Trait evolution under a Brownian Motion. The ancestral value of the trait is 0, and the observed values (time 800) range from -4 to 11 for extant species. One shift occurs on the parent branch of (C,D), changing the trajectory of their ancestral trait value from the grey one to the colored one. The shift increases the observed dispersion.

Brownian Motion. Since the seminal article of Felsenstein (1985), the BM has been used as a neutral model of trait evolution. If $(B_t; t \geq 0)$ is the Brownian motion, a character $(W_t; t \geq 0)$ evolves on a lineage according the the stochastic differential equation: $dW_t = \sigma dB_t$, σ^2 being a variance parameter. If μ is the ancestral value at the root of the tree ($t = 0$), then $W_t \sim \mathcal{N}(\mu, \sigma^2 t)$. The variance $\sigma^2 t$ of the trait is proportional to the time of evolution and the covariance $\sigma^2 t_{ij}$ between two species i and j is proportional to their time of shared evolution.

Ornstein-Uhlenbeck Process. An unbounded variance is quite unrealistic for adaptive traits (Butler and King, 2004). For that reason, the OU process, that models stabilizing selection around an adaptive optimum (Hansen, 1997) is usually preferred to the BM. It is defined by the stochastic differential equation:

$$dW_t = -\alpha(W_t - \beta)dt + \sigma dB_t$$

and has stationary distribution $\mathcal{N}(\beta, \sigma^2/2\alpha)$. In this equation, W_t is the *secondary optimum* of a species, a trade-off between all selective constraints – e.g. ecological – on the trait and can be approached by the population mean of that species. The term $-\alpha(W_t - \beta)dt$ of the equation represents the effects of stabilizing selection towards a *primary optimum* β, that depends only on the ecological niche of the species. The selection strength is controlled by the call-back parameter α. For interpretation purpose, we will use the *phylogenetic half-life* $t_{1/2} = \ln(2)/\alpha$, defined as the time needed for the expected trait value to move half the distance from the ancestral state to the primary optimum (Hansen, 1997). The term σdB_t represents the random effects of uncontrolled factors, ranging from genetic drift to environmental fluctuations. We refer to Hansen (1997); Hansen et al. (2008) for further discussion and deeper biological interpretations of the hypothesis underlying this model of evolution.

The aim of our work is to detect environmental shifts.

Environmental Shifts. In addition to the previous mechanisms, we assume that abrupt environmental changes affected the ecological niche of some ancestral species. We model these changes as instantaneous shifts in the parameters of the stochastic process. Shifted parameters are inherited along time and thus naturally create clusters of extant species that share the same

parameters trajectories. In the BM process, shifts affect the mean value of the trait and are thus instantaneously passed on to the trait itself (see Figure 1) whereas in the OU process, shifts affect the primary optimum β. In this case, the trait converges to its new stationary value with an exponential decay of half-life $t_{1/2}$ inducing a lag that makes recent shifts harder to detect (Hansen and Bartoszek, 2012). In the remainder, we assume that all other parameters (σ^2 for the BM and σ^2, α for the OU) are fixed and constant (but see Beaulieu et al., 2012; Rabosky, 2014, for partial relaxations of this hypothesis).

1.3 Scope of this article

State of the Art. Phylogenetics Comparative Methods (PCM) are an active field that has seen many fruitful developments in the last few years (see Pennell and Harmon, 2013, for an extensive review). Several methods have been specifically developed to study adaptive evolution, starting with the work of Butler and King (2004). Butler and King (2004) only considers shifts in the optimal value β whereas Beaulieu et al. (2012) also allow shifts in the selection strength α and the variance σ^2. Both have in common that shift locations are assumed to be known. Several extensions of the model without or with known shifts have also been proposed: Hansen et al. (2008) extended the original work of Hansen (1997) on OU processes to a two-tiered model where $\beta(t)$ is itself a stochastic process (either BM or OU). Bartoszek et al. (2012) extended it further to multivariate traits whereas Hansen and Bartoszek (2012) introduced errors in the observations. Expanding upon the BM, (Landis et al., 2013) replaced fixed shifts, known or unknown, by random jump processes using Levy processes. Non-Gaussian models of trait evolution were also recently considered by Hiscott et al. (2015), who adapted Felsenstein's pruning algorithm for the likelihood computation of these models, using efficient integration techniques. Finally, Ho and Ané (2013a) derived consistency results for estimation of the parameters of an OU on a tree and Bartoszek and Sagitov (2015); Sagitov and Bartoszek (2012) computed confidence intervals of the same parameters by assuming an unknown random tree topology and averaging it out.

The first steps toward automatic detection of shifts, which is the problem of interest in this paper, have been done in a Bayesian framework, for both the BM (Eastman et al., 2013) and the OU (Uyeda and Harmon, 2014). Using RJ-MCMC, they provide the user with the posterior distribution of the number and location of shifts on the tree. Convergence is however severely hampered by the size of the search space. The growing use of PCM in fields where large trees are the norm makes Maximum Likelihood (ML) based point estimates of the shift locations more practical. A stepwise selection procedure for the shifts has been proposed in Ingram and Mahler (2013). The procedure adds shifts one at the time and is therefore rather efficient but the selection criteria is heuristic and has no theoretical grounding for that problem, where observations are correlated through the tree structure. These limitations have been pointed out in Ho and Ané (2014). In this article, the authors describe several identifiability problems that arise when trying to infer the position of the shifts on a tree, and propose a different stepwise algorithm based on a more stringent selection criteria, heuristically inspired by segmentation algorithms.

To rigorously tackle this issue, we introduced a framework where a univariate trait evolving according to a OU process with stationary root state (S) on an Ultrametric tree (U). Furthermore,as the exact position of a shift on a branch is not identifiable for an ultrametric tree, we assume that shifts are concomitant to speciation events and only occur at Nodes (N) of the tree. We refer to this model as OUsun hereafter.

Our contribution. In this work, we make several major contributions to the problem at hand. *First*, we derive a statistical method to find a maximum likelihood estimate of the parameters of the model. When the number of shifts is fixed, we work out an Expectation Maximization (EM) algorithm that takes advantage of the tree structure of the data to efficiently maximize the likelihood. *Second*, we show that, given the model used and the kind of data available, some evolutionary scenarios remain indistinguishable. Formally, we exhibit some

identifiablitity problems in the location of the shifts, even when their number is fixed, and subsequently give a precise characterization of the space of models that can be inferred from the data on extant species. *Third*, we provide a rigorous model selection criterion to choose the number of shifts needed to best explain the data. Thanks to our knowledge of the structure of the spaces of models, acquired through our identifiability study, we are able to mathematically derive a penalization term, together with an oracle inequality on the estimator found. *Fourth* and finally, we implement the method on the statistical software R (R Core Team, 2014), and show that it correctly recovers the structure of the model on simulated datasets. When applied to a biological example, it gives results that are easily interpretable, and coherent with previously developed methods. All the code used in this article is publicly available on GitHub (`https://github.com/pbastide/Phylogenetic-EM`), and should be shortly released as an R-package.

Outline. In Section 2, we present the model, using two different mathematical point of views, that are both useful in different aspects of the inference. In Section 3, we tackle the identifiability problems associated with this model, and describe efficient algorithms to enumerate, first, all equivalent models within a class, and, second, the number of truly different models for a given number of shifts. These two sections form the foundation of Section 4, in which we describe our fully integrated maximum likelihood inference procedure. Finally, in sections 5 and 6, we conduct some numerical experiments on simulated and biological datasets.

2 Statistical Modeling

2.1 Probabilistic Model

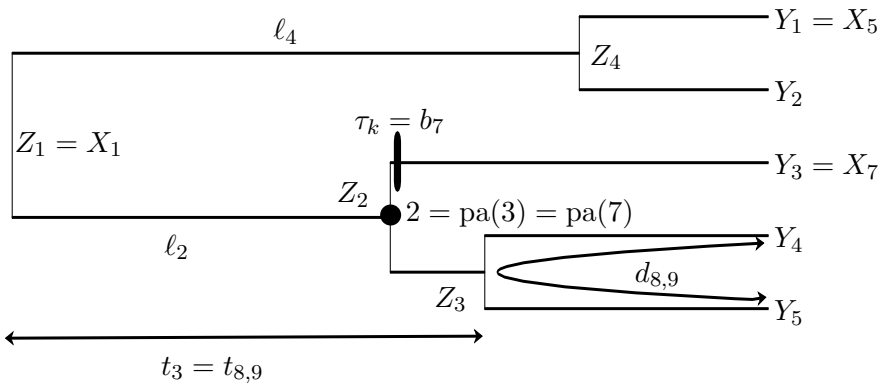


Figure 2: A rooted and time calibrated phylogenetic tree with some notations indicated for the parametrization of the tree and the observed and non-observed variables.

Tree Parametrization. As shown in Figure 2, we consider a rooted tree $\mathcal{T}$ with n tips and m internal nodes ($m = n-1$ for binary trees). The internal nodes are numbered from 1 (the root) to m, and the tips from $m+1$ to $m+n$. Let i be an integer, $i \in [\![2\,, m+n]\!]$. Then $\mathrm{pa}(i)$ is the unique parent of node i. The branch going from $\mathrm{pa}(i)$ to i is noted b_i and has length $\ell_i = t_i - t_{\mathrm{pa}(i)}$ where t_i is the time elapsed between the root and node i. By convention, we set $t_1 = 0$ and $t_{\mathrm{pa}(1)} = -\infty$ for the root. The last convention ensures that the trait follows the stationary distribution (if any) of the process at the root. We denote $\mathrm{Anc}(i) = \{\mathrm{pa}^r(i) : r \geq 0\}$ the set composed of node i and all its ancestors up to the root. For a couple of integers (i, j), $(i, j) \in [\![1\,, m+n]\!]^2$, nodes i and j are at phylogenetic distance d_{ij} and the time of their most recent common ancestor (mrca) is t_{ij}. We consider ultrametric trees, for which $t_{m+1} = \cdots = t_{m+n} =: h$ and note h the tree height. In the following, the tree is fixed and assumed to be known.

Trait Values. We denote by $\mathbf{X}$ the vector of size $m+n$ of the trait values at the nodes of the tree. We split this vector between non-observed values $\mathbf{Z}$ (size m) at the internal nodes, and observed values $\mathbf{Y}$ (size n) at the tips, so that $\mathbf{X}^T = (\mathbf{Z}^T, \mathbf{Y}^T)$. According to our model of trait evolution, the random variable X_i, $i \in [\![1\,, m+n]\!]$, is the result of a stochastic process stopped at time t_i. In the following, we assume that the inference in the BM case is done conditionally to a fixed root value $X_1 = \mu$. In the OUsun case, we assume that the root trait value is randomly drawn from the stationary distribution: $X_1 \sim \mathcal{N}(\mu = \beta_0, \gamma^2 = \frac{\sigma^2}{2\alpha})$, where β_0 is the ancestral optimal value.

Shifts. We assume that K shifts occur on the tree, $K \in \mathbb{N}$. The kth shift, $k \in [\![1\,, K]\!]$, occurs at the beginning of branch τ_k, $\tau_k \in \{b_i, i \in [\![2\,, m+n]\!]\}$, and has intensity δ_k, $\delta_k \in \mathbb{R}$. The interpretation of this intensity depends on the process. In the following, we use the vector $\boldsymbol{\Delta}$ of shifts on the branches, of size $m+n$, with $K+1$ non-zero entries, and defined as follows (see example 2.1):

$$\begin{cases} \Delta_1 = \mu \ (= \beta_0 \text{ for an OUsun}) \\ \Delta_i = \begin{cases} \delta_k & \text{if } \tau_k = b_i \\ 0 & \text{otherwise.} \end{cases} \end{cases} \qquad \forall i \in [\![2\,, n+m]\!]$$

Note that no proper shift occurs on the root branch, but that the root trait value or mean, μ, is formalized as an initial fictive shift on this fictive branch.

Parameters. The parameters needed to describe an OUsun (respectively, a BM) are $\boldsymbol{\theta} = (\gamma, \alpha, \boldsymbol{\Delta})$ (resp. $\boldsymbol{\theta} = (\sigma, \boldsymbol{\Delta})$). Note that, as $\sigma^2 = 2\alpha\gamma^2$, only the two parameters α and γ are needed to describe the OUsun. We denote by OUsun($\boldsymbol{\theta}$) (resp. BM($\boldsymbol{\theta}$)) the OUsun (resp. BM) process running on the tree with parameters $\boldsymbol{\theta}$.

2.2 Incomplete Data Model Point of View

If the trait values were observed at all nodes of the tree, including ancestral ones, shifts would be characterized by unexpectedly large differences between a node and its parent. A way to mimic this favorable case is to use an incomplete data model, as described below. This representation of the model will be useful for the parametric inference using an EM algorithm (Section 4.1).

Brownian Motion. As the shifts occur directly in the mean of the process, we get:

$$\begin{cases} X_1 = \mu \\ X_i | X_{\text{pa}(i)} \sim \mathcal{N}\left(X_{\text{pa}(i)} + \Delta_i,\ \ell_i \sigma^2\right) \quad \forall i \in [\![2\,, m+n]\!] \end{cases} \tag{2.1}$$

The trait value at node i, $i \in [\![2\,, m+n]\!]$, is centered on the value of its parent node $X_{\text{pa}(i)}$, with a variance proportional to the evolution time ℓ_i between i and pa(i). The effect of a non-zero shift Δ_i on branch b_i is simply to translate the trait value by Δ_i.

Ornstein-Uhlenbeck. The shifts occur on the primary optimum $\boldsymbol{\beta}$, which is piecewise constant. As the shifts are assumed to occur at nodes, the primary optimum is entirely defined by its initial value β_0 and its values $\beta_1, \ldots, \beta_{n+m}$ on branches of the tree, where β_i is the value on branch b_i leading to node i.

$$\begin{cases} \beta_1 = \beta_0 \\ \beta_i = \beta_{\text{pa}(i)} + \Delta_i \quad \forall i \in [\![2\,, m+n]\!] \end{cases} \tag{2.2}$$

Assuming that the root node is in the stationary state, we get:

$$\begin{cases} X_1 \sim \mathcal{N}(\mu = \beta_0, \gamma^2 = \frac{\sigma^2}{2\alpha}) \\ X_i | X_{\text{pa}(i)} \sim \mathcal{N}\left(X_{\text{pa}(i)} e^{-\alpha \ell_i} + \beta_i (1 - e^{-\alpha \ell_i}),\ \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha \ell_i})\right) \quad \forall i \in [\![2\,, m+n]\!] \end{cases} \tag{2.3}$$

The trait value at node i depends on both the trait value at the father node $X_{\text{pa}(i)}$ and the value β_i of the primary optimum on branch b_i. Contrary to the BM case, the shifts only appear

indirectly in the distributions of X_is, through the values of $\boldsymbol{\beta}$, and with a shrinkage of $1 - e^{-\alpha d}$ for shifts of age d, which make them harder to detect when d is small compared to $1/\alpha$.

2.3 Linear Regression Model Point of View

A more compact and direct representation of the model is to use the tree matrix to link linearly the observed values (at the tips) with the shift values, as explained below. We will use this linear regression framework for the Lasso (Tibshirani, 1996) initialization of the EM (Section 4.1) and the model selection procedure (Section 4.2). It will also help us to explore identifiability issues raised in the next section.

Matrix of a tree. It follows from its recursive definition that $\mathbf{X}$ is a Gaussian vector. In order to express its mean vector given the shifts, we introduce the tree squared matrix $\mathbf{U}$, of size $(m+n)$, defined by its general term: $U_{ij} = \mathbb{I}\{j \in \mathrm{Anc}(i)\}, \forall (i,j) \in [\![1\,, m+n]\!]^2$. In other words, the column j of this matrix, $j \in [\![1\,, m+n]\!]$, is the indicator vector of the descendants of node j. To express the mean vector of the observed values $\mathbf{Y}$, we also need the sub-matrix $\mathbf{T}$, of size $n \times (m+n)$, composed of the bottom n rows of matrix $\mathbf{U}$, corresponding to the tips (see example 2.1 below). Likewise, the ith row of $\mathbf{T}$, $i \in [\![1\,, n]\!]$, is the indicator vector of the ancestors of node $m+i$.

Brownian Motion. From the tree structure, we get:

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Delta} + \mathbf{E_X} \tag{2.4}$$

$$\mathbf{Y} = \mathbf{T}\boldsymbol{\Delta} + \mathbf{E_Y} \tag{2.5}$$

Here, $\mathbf{E_X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{XX}})$ is a Gaussian error vector with co-variances $[\Sigma_{XX}]_{ij} = \sigma^2 t_{ij}$ for any $1 \leq i, j \leq m+n$, and $\mathbf{E_Y}$ is the vector made of the last n coordinates of $\mathbf{E_X}$.

Ornstein-Uhlenbeck. For the OUsun, shifts occur on the primary optimum, and there is a lag term, so that:

$$\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\Delta} \tag{2.6}$$

$$\mathbf{X} = (\mathbf{U} - \mathbf{AUB})\boldsymbol{\Delta} + \mathbf{E_X} \tag{2.7}$$

where $\mathbf{A} = \mathrm{Diag}(e^{-\alpha t_i}, 1 \leq i \leq m+n)$ and $\mathbf{B} = \mathrm{Diag}(0, e^{\alpha t_{\mathrm{pa}(i)}}, 2 \leq i \leq m+n)$ are diagonal matrices of size $m+n$. As previously, $\mathbf{E_X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{XX}})$, but $\boldsymbol{\Sigma}_{\mathbf{XX}} = \gamma^2 [e^{-\alpha d_{ij}}]_{1 \leq i,j \leq m+n}$. As the tree is ultrametric, this expression simplifies to the following one when considering only observed values:

$$\mathbf{Y} = \mathbf{T}\mathbf{W}(\alpha)\boldsymbol{\Delta} + \mathbf{E_Y} \tag{2.8}$$

where $\mathbf{E_Y}$ is the Gaussian vector made of the last n coordinates of $\mathbf{E_X}$, and $\mathbf{W}(\alpha) = \mathrm{Diag}(1, 1 - e^{-\alpha(h - t_{\mathrm{pa}(i)})}, 2 \leq i \leq m+n)$ is a diagonal matrix of size $m+n$. Note that if α is positive, then $\alpha(h - t_{\mathrm{pa}(i)}) > 0$ for any $i \in [\![1\,, m+n]\!]$, and $\mathbf{W}(\alpha)$ is invertible.

Example 2.1. The tree presented in Figure 2 has five tips and one shift on branch $4+3=7$, with parent node 2, so:

$$\mathbf{U} = \begin{array}{c} \\ Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{array} \begin{array}{c} \begin{array}{ccccccccc} Z_1 & Z_2 & Z_3 & Z_4 & Y_1 & Y_2 & Y_3 & Y_4 & Y_5 \end{array} \\ \left(\begin{array}{ccccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array} \left.\vphantom{\begin{array}{c}1\\1\\1\\1\\1\\1\\1\\1\\1\end{array}}\right\} \mathbf{T} \quad \text{and } \boldsymbol{\Delta} = \begin{pmatrix} \mu \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \delta_1 \\ 0 \\ 0 \end{pmatrix}$$

And, respectively, for a BM or an OUsun ($\mu = \beta_0$):

$$\mathbb{E}[\mathbf{Y}] = \mathbf{T}\boldsymbol{\Delta} = \begin{pmatrix} \mu \\ \mu \\ \mu + \delta_1 \\ \mu \\ \mu \end{pmatrix} \text{ or } \mathbb{E}[\mathbf{Y}] = \mathbf{T}\mathbf{W}(\alpha)\boldsymbol{\Delta} = \begin{pmatrix} \beta_0 \\ \beta_0 \\ \beta_0 + \delta_1(1 - e^{-\alpha(h-t_2)}) \\ \beta_0 \\ \beta_0 \end{pmatrix}$$

Space of Expectations. Expressions 2.5 and 2.8 allow us to link the parameter $\boldsymbol{\theta}$ to the probability distribution of observations $\boldsymbol{Y}$ and to explore identifiability issues. In this linear formulation, detecting shifts boils down to identifying the non-zero components of $\boldsymbol{\Delta}$. The following lemma highlights the parallels between solutions of the BM and OUsun processes:

Lemma 2.1 (Similar Solutions)**.** *Let $\mathbf{m_Y} \in \mathbb{R}^n$ be a vector, $\mathcal{T}$ an ultrametric tree, α a positive real number, and σ, γ non-negative real numbers. Then there exists at least one vector $\boldsymbol{\Delta}^{BM}$, $\boldsymbol{\Delta}^{BM} \in \mathbb{R}^{m+n}$ (respectively, $\boldsymbol{\Delta}^{OU} \in \mathbb{R}^{m+n}$), such that the vector of expectations at the tips of a BM($\sigma, \boldsymbol{\Delta}^{BM}$) (respectively, an OUsun($\gamma, \alpha, \boldsymbol{\Delta}^{OU}$)) running on the tree $\mathcal{T}$ is exactly $\mathbf{m_Y}$.*

Furthermore, $\boldsymbol{\Delta}^{BM}$ is a solution to this problem for the BM if and only if $\boldsymbol{\Delta}^{OU} = \mathbf{W}(\alpha)^{-1}\boldsymbol{\Delta}^{BM}$ is a solution for the OUsun, and $\boldsymbol{\Delta}^{BM}$ and $\mathbf{W}(\alpha)^{-1}\boldsymbol{\Delta}^{BM}$ have the same support. These two vectors are said to be similar.

Proof. The first part of this lemma follows directly from formulas 2.5 (BM) and 2.8 (OU). Indeed, the maps $\boldsymbol{\Delta} \mapsto \mathbf{T}\boldsymbol{\Delta}$ and $\boldsymbol{\Delta} \mapsto \mathbf{T}\mathbf{W}(\alpha)\boldsymbol{\Delta}$ both span $\mathbb{R}^n$. The second part of the lemma is a consequence of $\mathbf{W}(\alpha)$ being diagonal and invertible ($\alpha > 0$). □

Remark 2.1. Lemma 2.1 shows that the OUsun and BM processes that induce a given $\mathbf{m_Y}$ use shifts located on the same branches, although they may differ on other parameters.

3 Identifiability and Complexity of a Model

3.1 Identifiability Issues

As we only have access to $\mathbf{Y}$ and not the whole of $\mathbf{X}$, we only have partial information about the shifts occurrence on the tree. In fact, several different allocations of the shifts can produce the same trait distribution at the tips, and hence are not identifiable. In other words, there exists parameters $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ with the same likelihood function: $p_{\boldsymbol{\theta}}(\cdot) = p_{\boldsymbol{\theta}'}(\cdot)$. Note that the notion of identifiability is intrinsic to the model and affects all estimation methods. Restricting ourselves to the parsimonious allocations of shifts only partially alleviates this issue, and, using a "random cluster model" representation of the problem, we are able to enumerate, first, all the equivalent solutions to a given problem, and, second, all the truly different solutions for a given number of shifts.

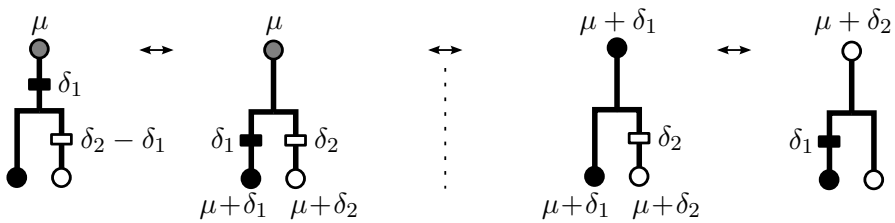


Figure 3: Equivalent allocations in the BM case. Mean tip values are represented by colors and equal for all allocations. The two allocations on the right are parsimonious.

No Homoplasy Assumption. We assume in the following that there is no convergent evolution. This means that each shift creates a new (and unique) mean trait value for extant species that are below them. This assumption is reasonable considering that shifts are real valued and

makes the model similar to "infinite alleles" models in population genetics. This assumption simplifies but does not settle the identifiability issues, as seen in Figure 3.

3.1.1 Definition of the problem

Figure 3 shows a simple example where the model is not identifiable in the BM case. Here, four distinct allocations give the same mean values $(\mu + \delta_1, \mu + \delta_2)$ at the tips. The lack of identifiability is due to the non-invertibility of the tree matrix $\mathbf{T}$.

Proposition 3.1 (Kernel of the Tree Matrix $\mathbf{T}$)**.** *Let i be an internal node, $i \in [\![1\,, m]\!]$, with L_i children nodes $(d_1, \cdots, d_{L_i}) \in [\![2\,, m+n]\!]^{L_i}$. Then the vector $\mathbf{K}^i$ defined as follow:*

$$\forall j \in [\![1\,, m+n]\!], K^i_j = \begin{cases} 1 & \text{if } j = i \\ -1 & \text{if } j \in (d_1, \cdots, d_{L_i}) \\ 0 & \text{otherwise} \end{cases}$$

is in the kernel of $\mathbf{T}$. In addition, the m vectors constructed this way form a basis of the kernel space of $\mathbf{T}$.

These kernel vectors effectively "cancel out" a shift on a branch by balancing it with the opposite shift on all its child branches. Note that the root mean value is treated as a shift.

Proof of Proposition 3.1. Let $i \in [\![1\,, m]\!]$ be an internal node with L_i children nodes $(d_1, \cdots, d_{L_i})$ and $\mathbf{K}^i$ the corresponding vector, defined as in the proposition. Then, for any $j \in [\![1\,, m+n]\!]$:

$$(\mathbf{U}\mathbf{K}^i)_j = U_{ji}K^i_i + \sum_{l=1}^{L_i} U_{jd_l}K^i_{d_l} = \mathbb{I}\{i \in \mathrm{Anc}(j)\} - \sum_{l=1}^{L_i} \mathbb{I}\{d_l \in \mathrm{Anc}(j)\}$$

We can then distinguish three possibilities:

- If $i \notin \mathrm{Anc}(j)$, then $d_l \notin \mathrm{Anc}(j)$ for all $l \in [\![1\,, L_i]\!]$ and $(\mathbf{T}\mathbf{K}^i)_j = 0$.
- If $j = i$, then $i \in \mathrm{Anc}(j)$, and, by definition, $d_l \notin \mathrm{Anc}(j)$ for any $l \in [\![1\,, L]\!]$, so $(\mathbf{U}\mathbf{K}^i)_i = 1$.
- Else, if i is an ancestor of j, with $i \neq j$, then, as i is internal, one (and only one) of its child d_l is also an ancestor of j (potentially j itself), so that the sum cancels out.

This proves that

$$\forall i \in [\![1\,, m]\!], \mathbf{U}\mathbf{K}^i = (\delta_{ij})_{1 \leq j \leq m+n} \tag{3.1}$$

In particular, as $\mathbf{T}\mathbf{K}^i$ is the vector of the last n coordinates of $\mathbf{U}\mathbf{K}^i$, this shows that the vectors $(\mathbf{K}^1, \cdots, \mathbf{K}^m)$ are in the kernel of $\mathbf{T}$.

Then, as we found m independent vectors in the kernel of $\mathbf{T}$, which is a space of dimension lower than m (as the n columns of $\mathbf{T}$ representing tips are linearly independent, and by the rank theorem), this family of vector is a basis of the kernel space. □

The following lemma describes the relationships that exist between these kernel vectors and the tree matrix $\mathbf{U}$ defined in Section 2.3:

Lemma 3.1. *Let b be the canonical basis of $\mathbb{R}^{m+n}$, and S a supplementary space of $\ker(\mathbf{T})$. Then $b' = (\mathbf{K}^1, \cdots, \mathbf{K}^m, \mathbf{b}_{m+1}, \cdots, \mathbf{b}_{m+n})$ is a basis adapted to the decomposition $\ker(\mathbf{T}) \oplus S$, and the matrix $\mathbf{U}$ (as defined in Section 2.3) is the change of basis matrix between b and b'. As a consequence, $\mathbf{U}$ is invertible.*

Proof of lemma 3.1. First, $(\mathbf{b}_{m+1}, \cdots, \mathbf{b}_{m+n})$ is a family of n independent vectors of S of dimension n, so is a basis of S, and b' is a basis adapted to $\ker(\mathbf{T}) \oplus S$.

Let $i \in [\![1\,, m+n]\!]$. Let's show that $\mathbf{U}\mathbf{b}'_i = \mathbf{b}_i$. If $m+1 \leq i \leq m+n$, then $\mathbf{b}'_i = \mathbf{b}_i$, and $\mathbf{U}\mathbf{b}_i = \mathbf{b}_i$ is the ith column of $\mathbf{U}$. Otherwise, if $1 \leq i \leq m$, then $\mathbf{b}'_i = \mathbf{K}^i$, and, from equation 3.1, $\mathbf{U}\mathbf{b}'_i = \mathbf{b}_i$. This shows that $\mathbf{U}$ is the change of basis matrix between b and b'. □

"Random Cluster Model" Representation. When inferring the shifts, we have to keep in mind this problem of non-identifiability, and be able to choose, if necessary, one or several possible allocations among all the equivalent ones. In order to study the properties of the allocations, we use a *random cluster model*, as defined in Mossel and Steel (2004). The following definition states the problem as a node coloring problem.

Definition 3.1 (Node Coloring)**.** Let $\mathcal{C}_K$ be a set of K arbitrary "colors", $K \in \mathbb{N}^*$. For a given shift allocation, the color of each node is given by the application $B : [\![1\,, m+n]\!] \to \mathcal{C}_K$ recursively defined in the following way:

- Choose a color $c \in \mathcal{C}_K$ for the root: $B(1) = c$.
- For a node i, $i \in [\![2\,, m+n]\!]$, set $B(i)$ to $B(\mathrm{pa}(i))$ if there is no shift on branch i, otherwise choose another color c, $c \in \mathcal{C}_K \setminus \{B(\mathrm{pa}(i))\}$, and set $B(i)$ to c.

Hereafter, we identify $(\mathcal{C}_K)^{[\![1,m+n]\!]}$ with $(\mathcal{C}_K)^{m+n}$ and refer to a node coloring indifferently as an application or a vector.

As the shifts only affect $\mathbb{E}[\mathbf{X}]$ and we only have access to $\mathbb{E}[\mathbf{Y}]$, we identify colors with the distinct values of $\mathbb{E}[\mathbf{Y}]$:

Definition 3.2 (Adapted Node Coloring)**.** A node coloring is said to be *adapted* to a shifted random process on a tree if two *tips* have the same color if and only if they have the same mean value under that process.

Proposition 3.2 (Adapted Coloring for BM and OUsun)**.** *Let σ and γ be two non-negative real numbers, and α a positive real number. Then:*

(i) In the BM case, if $\mathcal{C}$ is the set of possible mean values taken by the nodes of the tree, then the knowledge of the node colors is equivalent to the knowledge of $\mathbf{\Delta}$. Furthermore, the associated node coloring is adapted to the original BM.

(ii) In the OUsun case, from lemma 2.1, we can find a similar BM process, i.e. *with shifts on the same branches. Then the knowledge of the node coloring associated to this similar BM process is equivalent to the knowledge of the vector of shifts of the OUsun, and the node coloring obtained is adapted to the original OUsun.*

Proof of Proposition 3.2. The proof of (i) relies on expression 2.4, that states that $\mathbb{E}[\mathbf{X}] = \mathbf{U}\mathbf{\Delta}$. Defining $\mathcal{C}$ as the set off all distinct values of $\mathbb{E}[\mathbf{X}]$, we can identify $\mathbb{E}[\mathbf{X}]$ with the node coloring application that maps any node i with $\mathbb{E}[X]_i$. Since $\mathbf{U}$ is invertible (see lemma 3.1 above), we can go from one formalism to the other.
For (ii), we use lemma 2.1 to find a similar BM, and then use (i). □

From now on, we will study the problem of shifts allocation as a discrete-state coloring problem.

3.1.2 Parsimony

As we saw on Figure 3 there are multiple colorings of the internal nodes that lead to a given tips coloring. Among all these solutions, we choose to study only the *parsimonious* ones. This property can be seen as an optimality condition, as defined below:

Definition 3.3 (Parsimonious Allocation)**.** Given a vector of mean values at the tips produced by a given shifted stochastic process running on the tree, an adapted node coloring is said to be *parsimonious* if it has a minimum number of color changes. We denote by $\mathcal{S}_K^P$ the set of parsimonious allocations of K shifts on the $(m+n-1)$ branches of the tree (not counting the root branch).

As K shifts cannot produce more than $K+1$ colors, we can define an application $\phi : \mathcal{S}_K^P \to (\mathcal{C}_{K+1})^n$ that maps a parsimonious allocation of shifts to its associated tip partition .

Definition 3.4 (Equivalence)**.** Two allocations are said to be *equivalent* (noted $\sim$) if they produce the same partition of the tips and are both parsimonious. Mathematically:

$$\forall s_1, s_2 \in \mathcal{S}_K^P,\ s_1 \sim s_2 \iff \phi(s_1) = \phi(s_2)$$

In other words, two allocations are equivalent if they produce the same tip *coloring* up to a permutation of the colors. Given $d \in (\mathcal{C}_{K+1})^n$ a coloring of the tips of $\mathcal{T}$ with $K+1$ colors, $\phi^{-1}(d)$ is the set of equivalent parsimonious node coloring that coincide with d (up to a permutation of the colors) on the tree leaves.

Several dynamic programming algorithms already exist to compute the minimal number of shifts required to produce a given tips coloring, and to find one associated parsimonious solution (see Fitch, 1971; Sankoff, 1975; Felsenstein, 2004). Here, we need to be a little more precise, as we want to both count and enumerate all possible equivalent node colorings associated with a tip coloring. For simplicity reasons, we only present here the algorithm that counts $\left|\phi^{-1}(d)\right|$, for $d \in (\mathcal{C}_K)^n$. This algorithm can be seen as a corollary of the enumeration algorithm presented and proved in Appendix A. It has $O(K^2Ln)$ time complexity where L is the maximal number of children of the nodes of the tree.

Proposition 3.3 (Size of an equivalence class)**.** *Let d be a coloring of the tips, $d \in (\mathcal{C}_K)^n$, and let i be a node of tree $\mathcal{T}$ with L_i daughter nodes $(i_1, \cdots, i_{L_i})$, $L_i \geq 2$. Denote by $\mathcal{T}_i$ the sub-tree rooted at node i.*
For $k \in \mathcal{C}_K$, $S_i(k)$ is the cost *of starting from node i with color k, i.e. the minimal number of shifts needed to get the coloring of the tips of $\mathcal{T}_i$ defined by d, when starting with node i in color k. Denote by $T_i(k)$ the number of allocations on $\mathcal{T}_i$ that achieve cost $S_i(k)$.*

If i is a tip ($m+1 \leq i \leq m+n$), then,

$$S_i(k) = \begin{cases} 0 & \text{if } d(i) = k \\ +\infty & \text{otherwise} \end{cases} \qquad T_i(k) = \begin{cases} 1 & \text{if } d(i) = k \\ 0 & \text{otherwise} \end{cases}$$

Otherwise, if i is a node, for $1 \leq l \leq L_i$, define:

$$\mathcal{K}_k^l = \underset{p \in \mathcal{C}_K}{\operatorname{argmin}} \left\{ S_{i_l}(p) + \mathbb{I}\{p \neq k\} \right\}$$

As these sets are not empty, let $(p_1, \ldots p_L) \in \mathcal{K}_k^1 \times \ldots \times \mathcal{K}_k^L$. Then:

$$S_i(k) = \sum_{l=1}^{L} S_{i_l}(p_l) + \mathbb{I}\{p_l \neq k\} \quad \text{and} \quad T_i(k) = \prod_{l=1}^{L} \sum_{p_l \in \mathcal{K}_k^l} T_{i_l}(p_l)$$

At the root, if $\mathcal{L} = \operatorname{argmin}_{k \in \mathcal{C}_K} S_1(k)$, then $\left|\phi^{-1}(d)\right| = \sum_{k \in \mathcal{L}} T_1(k)$.

OU Practical Case. We can illustrate this notions on a simple example. We consider an OUsun on a random tree of unit height (total height $h = 1$). We put three shifts on the tree, producing a given trait distribution. Then, using proposition 3.2 and our enumeration algorithm, we can reconstruct the 5 possible allocations of shifts that produce the exact same distribution at the tips. These solutions are shown in Figure 4. Note that the colors are not defined by the value of the optimal regime, but by the mean value of the process at the tips. As a result, the groups shown in blue and red in the first solution have the same optimal value in this configuration, but not in any other. The second solution shown illustrates the fact that all the shifts values are inter-dependent, as changing the position of only one of them can have repercussions on all the others. Finally, the third solution shows that the timing of shifts matters: to have the same impact as an old shift, a recent shift must have a much higher intensity (under constant selection strength such as in the OUsun).

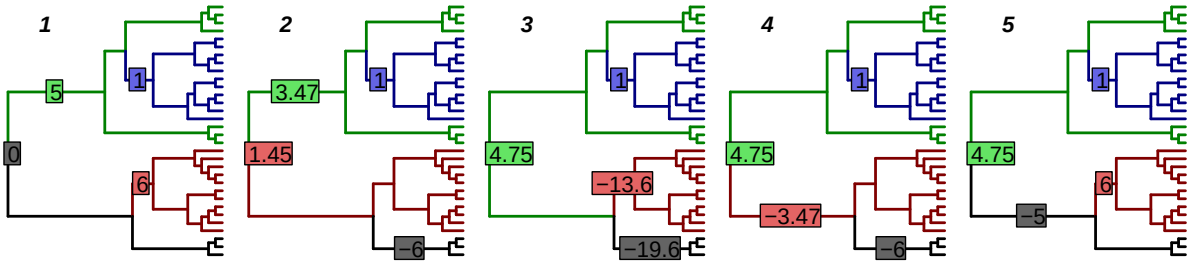


Figure 4: Five equivalent shift allocations that produce colorings that are adapted to an OUsun, with $\alpha = 3$ and $\gamma^2 = 0.1$. The box at the root represents the ancestral optimum β_0, and the boxes on the branches represent the positions and values of the shifts on the optimal value. While accounting for very different evolutionary scenarios, all these models produce the same trait distribution at the tips.

Possible relaxation of the No-homoplasy assumption. Note that the algorithms used for counting and enumerating the configurations of an equivalence class are valid even without the no-homoplasy hypothesis. The no-homoplasy hypothesis is however crucial in the next Section to establish a link between the number of shifts and the number of distinct tips colors.

3.2 Complexity of a Collection of Models

Number of different tips colors. As we make the inference on the parameters with a fixed number of shifts K (see Section 4.1), we need a model selection procedure to choose K. This procedure depends on the *complexity* of the collection of models that use K shifts, defined as the number of *distinct* models. To do that, we count the number of *tree-compatible* partitions of the tips into $K+1$ groups, as defined in the next proposition:

Proposition 3.4. *Under the no homoplasy assumption, an allocation of K shifts on a tree is parsimonious if and only if it creates exactly $K+1$ tip colors. The tip partition into $K+1$ groups associated with this coloring is said to be* tree-compatible*. The set $\mathcal{D}_{K+1} \subset (\mathcal{C}_{K+1})^n$ of such partitions is the image of $\mathcal{S}_K^P$ by the map ϕ defined in the previous section.*

Proof of Proposition 3.4. First, note that K shifts create at most $K+1$ colors. If each shift produces a new tip mean value (no homoplasy), the only way to create K or less colors is to "forget" one of the shifts, i.e. to put shifts on every descendant of the branch where it happens. Such an allocation is not parsimonious, as we could just add the value of the forgotten shift to all its descendant to get the same coloring of the tips with one less shift. So a parsimonious allocation cannot create less than $K+1$ colors, and hence creates exactly $K+1$ colors.

Reciprocally, if an allocation with K shifts that produces p groups is not parsimonious, then we can find another parsimonious one that produces the same p groups with $p-1$ shifts, with $p-1 < K$, i.e. $p < K+1$. So, by contraposition, if the allocation produces $K+1$ groups, then it is parsimonious. □

Using the equivalence relation defined in Definition 3.4, we can formally take the quotient set of $\mathcal{S}_K^P$ by the relation $\sim$ to get the set of parsimonious allocations of K shifts on the $m+n-1$ branches of the tree that are identifiable:

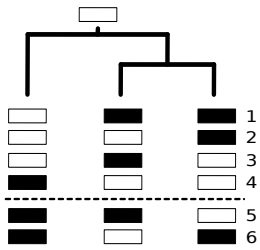
$$\mathcal{S}_K^{PI} = \mathcal{S}_K^P / \sim$$

In other words, the set $\mathcal{S}_K^{PI}$ is constituted of one representative of each equivalence class. Under the no homoplasy assumption, there is thus a bijection between identifiable parsimonious allocations of K shifts and tree-compatible partitions of the tips in $K+1$ groups: $\mathcal{S}_K^{PI} \underset{\sim}{\rightarrow} \mathcal{D}_{K+1}$.

The number $N_{K+1}^{(\mathcal{T})} = |\mathcal{D}_{K+1}|$ is the complexity of the class of models with K shifts defined as the number of distinct identifiable parsimonious possible configurations one can get with K shifts on the tree. To compute $N_K^{(\mathcal{T})}$, we will need $M_K^{(\mathcal{T})}$ the number of *marked* tree-compatible

partitions in K groups. These are composed of all the tree-compatible partitions that have no shift between the root and the marked group, seen as "ancestral" (see example 3.1 below).

Example 3.1 (Difference between $N_2^{(\mathcal{T})}$ and $M_2^{(\mathcal{T})}$).



Partitions in 2 groups.

- If we consider only unmarked partitions, then partitions 1, 2 and 3 are equivalent, respectively, to partitions 4, 5 and 6, and $N_2^{(\mathcal{T})} = 3$.
- If one partition is marked (here for instance, white tips are supposed to be in the "ancestral" state), then partitions 1 and 4 are not equivalent, and $M_2^{(\mathcal{T})} = 4$.
- Note that in the marked case, partitions 5 and 6 do not contribute to $M_2^{(\mathcal{T})}$ as they are not tree-compatible (they require two shifts).

Proposition 3.5 (Computation of the Number of Equivalent Classes)**.** *Let i be a node of tree $\mathcal{T}$, and $K \in \mathbb{N}^*$.*

If i is a tip, then $N_K^{(\mathcal{T}_i)} = M_K^{(\mathcal{T}_i)} = \mathbb{I}\{K = 1\}$.

Else, if i is a node with L_i daughter nodes $(i_1, \cdots, i_{L_i})$, $L_i \geq 2$, then:

$$\begin{cases} N_K^{(\mathcal{T}_i)} = \displaystyle\sum_{\substack{I \subset [\![1, L_i]\!] \\ |I| \geq 2}} \sum_{\substack{k_1 + \cdots + k_{L_i} = K + |I| - 1 \\ k_1, \ldots, k_{L_i} \geq 1}} \prod_{l \in I} M_{k_l}^{(\mathcal{T}_{i_l})} \prod_{l \notin I} N_{k_l}^{(\mathcal{T}_{i_l})} + \sum_{\substack{k_1 + \cdots + k_{L_i} = K \\ k_1, \ldots, k_{L_i} \geq 1}} \prod_{l=1}^{L_i} N_{k_l}^{(\mathcal{T}_{i_l})} \\ M_K^{(\mathcal{T}_i)} = \displaystyle\sum_{\substack{I \subset [\![1, L_i]\!] \\ |I| \geq 1}} \sum_{\substack{k_1 + \cdots + k_{L_i} = K + |I| - 1 \\ k_1, \ldots, k_{L_i} \geq 1}} \prod_{l \in I} M_{k_l}^{(\mathcal{T}_{i_l})} \prod_{l \notin I} N_{k_l}^{(\mathcal{T}_{i_l})} \end{cases} \tag{3.2}$$

In the binary case, this relation becomes, if i has two daughters i_ℓ and i_r:

$$\begin{cases} N_K^{(\mathcal{T}_i)} = \displaystyle\sum_{\substack{k_1 + k_2 = K \\ k_1, k_2 \geq 1}} N_{k_1}^{(\mathcal{T}_{i_\ell})} N_{k_2}^{(\mathcal{T}_{i_r})} + \sum_{\substack{k_1 + k_2 = K + 1 \\ k_1, k_2 \geq 1}} M_{k_1}^{(\mathcal{T}_{i_\ell})} M_{k_2}^{(\mathcal{T}_{i_r})} \\ M_K^{(\mathcal{T})} = \displaystyle\sum_{\substack{k_1 + k_2 = K \\ k_1, k_2 \geq 1}} M_{k_1}^{(\mathcal{T}_{i_\ell})} N_{k_2}^{(\mathcal{T}_{i_r})} + N_{k_1}^{(\mathcal{T}_{i_\ell})} M_{k_2}^{(\mathcal{T}_{i_r})} + \sum_{\substack{k_1 + k_2 = K + 1 \\ k_1, k_2 \geq 1}} M_{k_1}^{(\mathcal{T}_{i_\ell})} M_{k_2}^{(\mathcal{T}_{i_r})} \end{cases} \tag{3.3}$$

Proof. We will prove this proposition in the binary case, the general case being a natural extension of it. If $\mathcal{T}$ is a binary tree with $\mathcal{T}_\ell$ and $\mathcal{T}_r$ as left and right sub-trees, one faces two situations when partitioning the tips in K groups:

- The left and right sub-trees have no group in common. Then, the number of groups in $\mathcal{T}$ is equal to the number of groups in its two sub-trees, and there are $\sum_{k_1+k_2=K} N_{k_1}^{(\mathcal{T}_\ell)} N_{k_2}^{(\mathcal{T}_r)}$ such partitions. This is the first term of the equation on $N_K^{(\mathcal{T})}$ in 3.3.
- The left and right sub-trees have at least one group in common. Then, from the no homoplasy assumption, they have exactly one group in common: the ancestral state of the root. Suppose that this ancestral state is marked. Then it must be present in the two sub-trees, and there are $\sum_{k_1+k_2=K+1} M_{k_1}^{(\mathcal{T}_\ell)} M_{k_2}^{(\mathcal{T}_r)}$ such partitions. This ends the proof of the formula on $N_K^{(\mathcal{T})}$.

To get the formula on $M_K^{(\mathcal{T})}$, we use the same kind of arguments. The second part of the formula is the same as the one for $N_K^{(\mathcal{T})}$, and the first part corresponds to trees for which the marked partition is present in only one of the two sub-trees. □

The complexity of the algorithm described above is $O(2^L(K+L)^L Ln)$. Note that $N_K^{(\mathcal{T})}$ depends on the topology of the tree $\mathcal{T}$ in general. However, if the tree is binary, a closed form solution of the recurrence relation 3.2, which does not depend on the topology, exists.

Corollary 3.1 (Closed Formula Binary Trees)**.** *In the case of a rooted binary tree with n tips, we have:*

$$N_{K+1}^{(\mathcal{T})} = N_{K+1}^{(n)} = \left|\mathcal{S}_K^{PI}\right| = \binom{2n-2-K}{K} \text{ and } M_K^{(\mathcal{T})} = M_K^{(n)} = \binom{2n-K}{K-1}$$

The demonstration of this formula is not immediate, and is based on a Vandermonde-like equality, detailed in Appendix B. The formula is then obtained using a strong induction on the number of tips of the tree.

Remark 3.1. Note that, when K is large compared to $\sqrt{n}$, the average number of configurations per equivalence class goes to infinity. This can be checked by comparing the total number of configurations $\binom{2n-1}{K-1}$ with the total number of classes $\binom{2n-K-1}{K-1}$. As a consequence, we only consider models for which $K < \sqrt{n}$ in the remainder.

Remark 3.2. This formula was already obtained in a different context in Steel (1992) (Proposition 1) and, with a slightly different formulation, in Semple and Steel (2003, Proposition 4.1.4). In these works, the authors are interested in counting the "r-states convex characters on a binary tree". Under the no-homoplasy assumption, this number can be shown to be equal to $\left|\mathcal{S}_{r-1}^{PI}\right|$.

3.3 Another Characterization of Parsimony

The following proposition gives an alternative definition of parsimony under the no-homoplasy hypothesis using the linear formulation of the problem. It will be used for model selection in Section 4.2.

Proposition 3.6 (Equivalence between parsimony and independence)**.** *Let $\mathbf{m_Y}$ be a given mean vector, $\mathbf{m_Y} \in \mathbb{R}^n$, and $\mathbf{\Delta}$ a vector of shifts such that $\mathbf{T\Delta} = \mathbf{m_Y}$, with $\mathbf{T}$ the tree matrix defined in Section 2.3. Under the no homoplasy assumption, the vector of shifts $\mathbf{\Delta}$ is parsimonious if and only if the corresponding column-vectors of the tree matrix $(T_i)_{i\in\mathrm{Supp}(\mathbf{\Delta})}$ are linearly independent.*

Proof of Proposition 3.6. By contraposition, let's first assume that the vector-columns $(T_i)_{i\in\mathrm{Supp}(\mathbf{\Delta})}$ are linearly dependent, and prove that $\mathbf{\Delta}$ is not parsimonious. This means that we can find a vector $\mathbf{E}$, $\mathbf{E} \in \mathbb{R}^{m+n}$, such that $\mathrm{Supp}(\mathbf{E}) \subset \mathrm{Supp}(\mathbf{\Delta})$, and $\mathbf{TE} = 0$. We can hence find $j \in \mathrm{Supp}(\mathbf{\Delta})$, $j > 1$, such that $E_j \neq 0$. Then if $\lambda = -\Delta_j/E_j$, the vector $\mathbf{\Delta}' = \mathbf{\Delta} + \lambda\mathbf{E}$ is a vector of shifts on the tree with one less non-zero coordinate than $\mathbf{\Delta}$ such that $\mathbf{T\Delta}' = \mathbf{m_Y}$. Hence, $\mathbf{\Delta}$ is not parsimonious.

Reciprocally, by contraposition, assume that $\mathbf{\Delta}$ is not parsimonious. Then by proposition 3.4, it produces p groups, with $p \leq K$. Hence the application associated with $(T_i)_{i\in\mathrm{Supp}(\mathbf{\Delta})}$ goes from a space of dimension $K+1$ to a space of dimension $p \leq K$, and hence is not injective, and the family $(T_i)_{i\in\mathrm{Supp}(\mathbf{\Delta})}$ is not independent. □

4 Statistical Inference

4.1 Expectation Maximization

Principle. As shown in Section 2.2, both BM and OUsun models can be seen as incomplete data models. The Expectation Maximization algorithm (EM, Dempster et al., 1977) is a widely used algorithm for likelihood maximization of these kinds of models. It is based on the decomposition:

$$\log p_{\boldsymbol{\theta}}(\mathbf{Y}) = \mathbb{E}_{\boldsymbol{\theta}}[\log p_{\boldsymbol{\theta}}(\mathbf{Z},\mathbf{Y}) \mid \mathbf{Y}] - \mathbb{E}_{\boldsymbol{\theta}}[\log p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y}) \mid \mathbf{Y}]$$

Given an estimate $\boldsymbol{\theta}^{(h)}$ of the parameters, we need to compute some moments of $p_{\boldsymbol{\theta}^{(h)}}(\mathbf{Z} \mid \mathbf{Y})$ (E step), and then find a new estimate $\boldsymbol{\theta}^{(h+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}^{(h)}}[\log p_{\boldsymbol{\theta}}(\mathbf{Z}, \mathbf{Y}) \mid \mathbf{Y}]$ (M step). The parameters are given for the BM and OUsun in subsection 2.1. We assume here that the number of shifts K is fixed.

We only provide the main steps of the EM. Additional details can be found in Appendix C.

E step. As $\mathbf{X}$ is Gaussian, the law of the hidden variables $\mathbf{Z}$ knowing the observed variables $\mathbf{Y}$ is entirely defined by its expectation and variance-covariance matrix, and can be computed using classical formulas for Gaussian conditioning. The needed moments of $\mathbf{Z} \mid \mathbf{Y}$ can also be computed using a procedure that is linear in the number of tips (called "Upward-downward") that takes advantage of the tree structure and bypasses inversion of the variance-covariance matrix (see Lartillot, 2014, for a similar algorithm).

Complete Likelihood Computation. Using the model described in Section 2.2, we can use the following decomposition of the complete likelihood:

$$p_{\boldsymbol{\theta}}(\mathbf{X}) = p_{\boldsymbol{\theta}}(X_1) \prod_{j=2}^{m+n} p_{\boldsymbol{\theta}}\left(X_j \mid X_{\mathrm{pa}(j)}\right)$$

Each term of this product is then known, and we easily get $\mathbb{E}_{\boldsymbol{\theta}^{(h)}}[\log p_{\boldsymbol{\theta}}(\mathbf{Z}, \mathbf{Y}) \mid \mathbf{Y}]$.

M step. The difficulty comes here from the discrete variables (location of shifts on the branches). The maximization is exact for the BM but we only raise the objective function for the OUsun, hence computing a Generalized EM (GEM, see Dempster et al., 1977). This stems from the independent increment nature of the BM: shifts only affect $p_{\boldsymbol{\theta}}\left(X_j \mid X_{\mathrm{pa}(j)}\right)$ on the branches where they occur and the maximization reduces to finding the K highest components of a vector, which has complexity $O(n+K\log(n))$. By contrast, OUsun has autocorrelated increments: shifts affect $p_{\boldsymbol{\theta}}\left(X_j \mid X_{\mathrm{pa}(j)}\right)$ on the branches where they occur and on all subsequent branches. Maximization is therefore akin to segmentation on a tree, which has complexity $O(n^K)$.

Initialization. Initialization is always a crucial step when using an EM algorithm. Here, we use the linear formulation 2.5 or 2.8, and initialize the vector of shifts using a a Lasso regression. The selection strength α is initialized using pairs of tips likely to be in the same group.

4.2 Model Selection

Model Selection in the iid Case with Unknown Variance. Model selection in a linear regression setting has received a lot of attention over the last few years. In Baraud et al. (2009), the authors developed a non-asymptotic method for model selection in the case where the errors are independent and identically distributed (iid), with an unknown variance. In the following, we first recall their main results, and then adapt it to our setting of non-independent errors.

We assume that we have the following model of *independent* observations:

$$\mathbf{Y}' = \mathbf{s}' + \gamma \mathbf{E}' \quad \text{with} \quad \mathbf{E}' \sim \mathcal{N}(0, \mathbf{I_n})$$

and we define a collection $\mathcal{S}' = \{S'_\eta, \eta \in \mathcal{M}\}$ of linear subspaces of $\mathbb{R}^n$ that we call *models*, and that are indexed by a finite or countable set $\mathcal{M}$. For each $\eta \in \mathcal{M}$, we denote by $\hat{\mathbf{s}}'_\eta = \mathrm{Proj}_{S'_\eta} \mathbf{Y}'$ the orthogonal projection of $\mathbf{Y}'$ on S'_η, that is a least-square estimator of $\mathbf{s}'$, and $\mathbf{s}'_\eta = \mathrm{Proj}_{S'_\eta} \mathbf{s}'$ the projection of $\mathbf{s}'$.

We extract from Baraud et al. (2009) the following theorem, that bounds the risk of the selected estimator, and provides us with a non-asymptotic guarantee. It relies on a penalty depending on the EDkhi function, as defined below:

Definition 4.1 (Baraud et al. (2009), Section 4, definitions 2 and 3)**.** Let D, N be two positive integers, and X_D, X_N be two independent χ^2 random variables with degrees of freedom D and

N respectively. For $x \leq 0$, define

$$\mathrm{Dkhi}[D, N, x] = \frac{1}{\mathbb{E}\left[X_D\right]} \mathbb{E}\left[\left(X_D - x\frac{X_N}{N}\right)_+\right]$$

And, for $0 < q \leq 1$ define $\mathrm{EDkhi}[D, N, q]$ as the unique solution of the equation

$$\mathrm{Dkhi}[D, N, \mathrm{EDkhi}[D, N, q]] = q$$

Theorem 4.1 (Baraud et al. (2009), Section 4, theorem 2 and corollary 1)**.** *In the setting defined above, let D_η be the dimension of S'_η, and assume that $N_\eta = n - D_\eta \geq 2$ for all $\eta \in \mathcal{M}$. Let $\mathcal{L} = \{L_\eta\}_{\eta \in \mathcal{M}}$ be some family of positive numbers such that $\Omega' = \sum_{\eta \in \mathcal{M}} (D_\eta + 1)e^{-L_\eta} < +\infty$, and assume that, for $A > 1$,*

$$\mathrm{pen}(\eta) = \mathrm{pen}_{A,\mathcal{L}}(\eta) = A\frac{N_\eta}{N_\eta - 1}\,\mathrm{EDkhi}[D_\eta + 1, N_\eta - 1, e^{-L_\eta}]$$

Then, if $\hat{\eta}$ is the minimizer of the following criterion:

$$\hat{\eta} = \underset{\eta \in \mathcal{M}}{\mathrm{argmin}} \left\|\mathbf{Y}' - \hat{\mathbf{s}}'_\eta\right\|^2 \left(1 + \frac{\mathrm{pen}(\eta)}{N_\eta}\right) \tag{4.1}$$

the following non-asymptotic bound holds:

$$\mathbb{E}\left[\frac{\left\|\mathbf{s}' - \hat{\mathbf{s}}'_{\hat{\eta}}\right\|^2}{\gamma^2}\right] \leq \frac{A}{A-1} \inf_{\eta \in \mathcal{M}} \left\{\frac{\left\|\mathbf{s}' - \mathbf{s}'_\eta\right\|^2}{\gamma^2}\left(1 + \frac{\mathrm{pen}(\eta)}{N_\eta}\right) + \mathrm{pen}(\eta) - D_\eta\right\} + 2A^2\frac{\Omega'}{A-1}$$

In addition, if $\kappa < 1$, $N_\eta \geq 7$ and $\max(L_\eta, D_\eta) \leq \kappa n$ for any $\eta \in \mathcal{M}$ then:

$$\mathbb{E}\left[\frac{\left\|\mathbf{s}' - \hat{\mathbf{s}}'_{\hat{\eta}}\right\|^2}{\gamma^2}\right] \leq C(A, \kappa)\left[\inf_{\eta \in \mathcal{M}}\left\{\frac{\left\|\mathbf{s}' - \mathbf{s}'_\eta\right\|^2}{\gamma^2} + \max(L_\eta, D_\eta)\right\} + \Omega'\right]$$

The penalty used here ensures an oracle inequality: in expectation, the risk of the selected estimator is bounded by the risk of the best possible estimator of the collection of models, up to a multiplicative constant, and a residual term that depends on the dimension of the oracle model. Note that if the collection of models is poor, such an inequality has low value. We refer to Baraud et al. (2009) for a more detailed discussion of this result.

Adaptation to the Tree-Structured Framework. We use the linear formulation described in 2.3, and assume that we are in the OUsun model (this procedure would also work for a BM with a deterministic root). Then, if $\mathbf{V}$ is a matrix of size n, with $V_{ij} = e^{-\alpha d_{ij}}, \forall (i,j) \in [\![1\,, n]\!]^2$, we have:

$$\mathbf{Y} = \mathbf{TW}(\alpha)\mathbf{\Delta} + \gamma\mathbf{E} = \mathbf{s} + \gamma\mathbf{E} \quad \mathbf{E} \sim \mathcal{N}(0, \mathbf{V})$$

We assume that α is fixed, so that the design matrix $\mathbf{TW}(\alpha)$ and the structure matrix $\mathbf{V}$ are known and fixed. A *model* is defined here by the position of the shifts on the branches of the tree, i.e. by the non-zero components of $\mathbf{\Delta}$ (with the constraint that the first component, the root, is always included in the model). We denote by $\mathcal{M} = \bigcup_{K=0}^{p-1} \mathcal{S}_K^{PI}$ the set of allowed (parsimonious) allocations of shifts on branches (see Section 3.2), p being the maximum allowed dimension of a model. From proposition 3.6, for $\eta \in \mathcal{M}$, the columns vectors $\mathbf{T}_\eta$ are linearly independent, and the model $S_\eta = \mathrm{Span}(\mathbf{T_i}, i \in \eta)$ is a linear sub-space of $\mathbb{R}^n$ of dimension $D_\eta = |\eta| = K_\eta + 1$, K_η being the number of shifts in model η. Note that as $\mathbf{W}(\alpha)$ is diagonal invertible, it does not affect the definition of the linear subspaces. The set of models is then $\mathcal{S} = \{S_\eta,\ \eta \in \mathcal{M}\}$.

We define the Mahalanobis norm associated to $\mathbf{V}^{-1}$ by:

$$\|\mathbf{R}\|_{\mathbf{V}^{-1}} = \mathbf{R}^T\mathbf{V}^{-1}\mathbf{R},\ \forall \mathbf{R} \in \mathbb{R}^n$$

The projection on S_η according to the metric defined by $\mathbf{V}^{-1}$ is then:

$$\hat{\mathbf{s}}_\eta = \mathrm{Proj}_{S_\eta}^{\mathbf{V}^{-1}}(\mathbf{Y}) = \underset{\mathbf{a}\in S_\eta}{\mathrm{argmin}}\, \|\mathbf{Y}-\mathbf{a}\|^2_{\mathbf{V}^{-1}} \quad \text{and} \quad \mathbf{s}_\eta = \mathrm{Proj}_{S_\eta}^{\mathbf{V}^{-1}}(\mathbf{s})$$

And, for a given number of shifts K, we define the best model with K shifts as the one maximizing the likelihood, or, equivalently, minimizing the least-square criterion for models with K shifts:

$$\hat{\mathbf{s}}_K = \underset{\eta\in\mathcal{S},|\eta|=K+1}{\mathrm{argmin}} \|\mathbf{Y}-\hat{\mathbf{s}}_\eta\|^2_{\mathbf{V}^{-1}}$$

The idea is then to slice the collection of models by the number of shifts K they employ. Thanks to the EM algorithm above, we are able to select the best model in such a set. The problem is then to select a reasonable number of shifts. To compensate the increase in the likelihood due to over-fitting, using the model selection procedure described above, we select K using the following penalized criterion:

$$\mathrm{Crit}_{LS}(K) = \|\mathbf{Y}-\hat{\mathbf{s}}_K\|^2_{\mathbf{V}^{-1}}\left(1+\frac{\mathrm{pen}(K)}{n-K-1}\right) \tag{4.2}$$

Remark 4.1. As noted in Baraud et al. (2009), the previous criterion can equivalently be rewritten in term of likelihood, as:

$$\mathrm{Crit}_{LL}(K) = \frac{n}{2}\log\left(\frac{\|\mathbf{Y}-\hat{\mathbf{s}}_K\|^2_{\mathbf{V}^{-1}}}{n}\right) + \frac{1}{2}\mathrm{pen}'(K) \tag{4.3}$$

with $\mathrm{pen}'(K) = n\log\left(1+\frac{\mathrm{pen}(K)}{n-K-1}\right)$. As we use maximum-likelihood estimators, we chose this formulation for the implementation.

The following proposition holds:

Proposition 4.1 (Form of the Penalty and guaranties (α known))**.** *Let $\mathcal{L} = \{L_K\}_{K\in[\![0,p-1]\!]}$, with $p \le n-7$ the maximum dimension of a model, and:*

$$L_K = \log\left|\mathcal{S}_K^{PI}\right| + 2\log(K+2), \forall K \in [\![0\,, p-1]\!] \tag{4.4}$$

Let $A > 1$ and assume that

$$\mathrm{pen}_{A,\mathcal{L}}(K) = A\frac{n-K-1}{n-K-2}\,\mathrm{EDkhi}[K, n-K-2, e^{-L_K}]$$

Suppose that $\hat{K}$ is a minimizer of 4.2 or 4.3 with this penalty.
Then, if $\kappa < 1$, and $p \le \min\left(\frac{\kappa n}{2+\log(2)+\log(n)}, n-7\right)$, we get:

$$\mathbb{E}\left[\frac{\left\|\mathbf{s}-\hat{\mathbf{s}}_{\hat{K}}\right\|^2_{\mathbf{V}^{-1}}}{\gamma^2}\right] \le C(A,\kappa)\inf_{\eta\in\mathcal{M}}\left\{\frac{\|\mathbf{s}-\mathbf{s}_\eta\|^2_{\mathbf{V}^{-1}}}{\gamma^2} + (K_\eta+2)\,(3+\log(n))\right\}$$

with $C(A,\kappa)$ a constant depending on A and κ only.

The proof of this proposition can be found in Appendix D. It relies on theorem 4.1, adapting it to our tree-structured observations.

Remark 4.2. With this oracle inequality, we can see that we are missing the oracle by a $\log(n)$ term. This $\log(n)$ term is known to be unavoidable, see Baraud et al. (2009) for further explanations.

Remark 4.3. Note that the chosen penalty may depend on the topology of the tree through the term $\left|\mathcal{S}_K^{PI}\right|$ (see Section 3.2).

Remark 4.4. The penalty involves a constant $A > 1$, that needs to be chosen by the user. Following Baraud et al. (2009) who tested a series of values, we fixed this constant to $A = 1.1$.

5 Simulations Studies

5.1 Simulations Scheme

We tested our algorithm on data simulated according to an OUsun, with varying parameters. The simulation scheme is inspired by the work of Uyeda and Harmon (2014). We first generated three distinct trees with, respectively, 64, 128 and 256 tips, using a pure birth process with birth rate $\lambda = 0.1$. The tree heights were scaled to one, and their topology and branch lengths were fixed for the rest of the simulations. We then used a star-like simulation study scheme, fixing a base scenario, and exploring the space of parameters one direction at the time. The base scenario was taken to be relatively "easy", with $\beta_0 = 0$ (this parameter was fixed for the rest of the simulations), $\alpha_b = 3$ (i.e $t_{1/2,b} = 23\%$), $\gamma_b^2 = 0.5$ and $K_b = 5$. The parameters then varied in the following ranges: the phylogenetic half life $t_{1/2} = \ln(2)/\alpha$ took 11 values in $[0.01\,, 10]$; the root variance $\gamma^2 = \frac{\sigma^2}{2\alpha}$ took 9 values in $[0.05\,, 25]$; the number of shifts K took 9 values in $[0\,, 16]$ (see Figures 6-7 for the exact values taken). The problem was all the more *difficult* that γ^2, $t_{1/2}$ or K were large.

For each simulation, the K shifts were generated in the following way. First, their values were drawn according to a mixture of two Gaussian distributions, $\mathcal{N}(4, 1)$ and $\mathcal{N}(-4, 1)$, in equal proportions. The mixture was chosen to avoid too many shifts of small amplitude. Then, their positions were chosen to be balanced: we first divided the tree in K segments of equal heights, and then randomly drew in each segment an edge where to place a shift. We only kept parsimonious allocations.

Each of these configurations was repeated 200 times, leading to 16200 simulated data sets. An instance of a tree with the generated data is plotted in Figure 5.

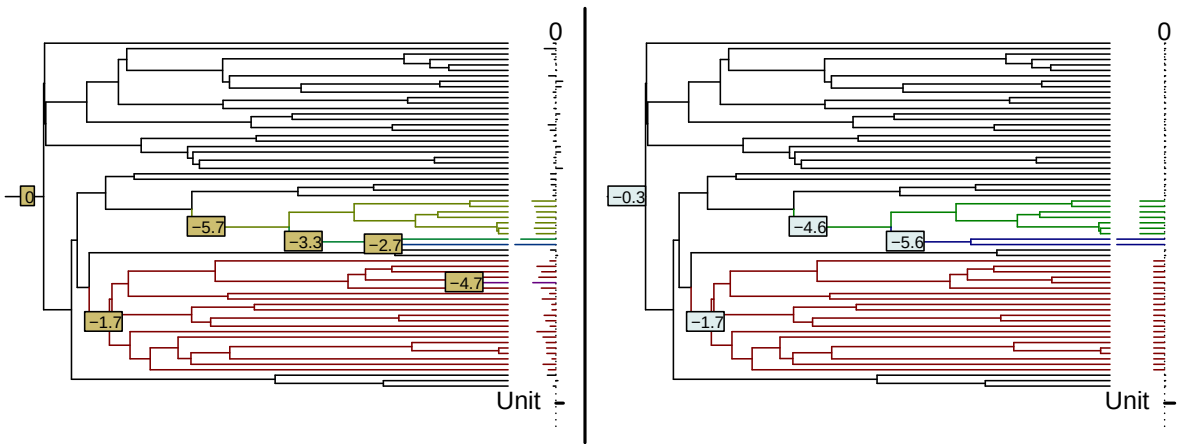


Figure 5: *Left*: Simulated configuration (with $t_{1/2} = 0.75$, $\gamma^2 = 0.5$ and $K = 5$). The shifts positions and values are marked on the tree. The value of the character generated (positive or negative) is represented on the right. The colors of the branches correspond to the true regimes, black being the ancestral state. *Right*: One of the three equivalent allocations of shifts for the model inferred from the data, with corresponding vector of mean tip values. Shifts not recovered are located on pendant edges, and have low influence on the data. The two other equivalent allocations can be easily deducted from this one.

5.2 Inference Procedures

For each generated dataset, we ran our EM procedure with fixed values of $K \in [\![0\,, \lfloor\sqrt{n}\rfloor]\!]$, n being the number of tips of the tree. Remark that for $n = 64$, $\lfloor\sqrt{n}\rfloor = 8$, and we have no hope of detecting true values of K above 8 (see remark 3.1 for an explanation of the bound in $\sqrt{n}$). The number of shifts K_s was chosen thanks to our penalized criterion, and we kept inferences corresponding to both K_s and the true number K_t.

We ran two sets of estimations for α either known or estimated. The computations took respectively 66 and 570 (cumulated) days of CPU time. This amounts to a mean computational time of around 6 minutes (367 seconds) for one estimation when α is fixed, and 52 minutes (3137 seconds) when α is estimated, with large differences between easy and difficult scenarios.

5.3 Scores Used

The convergence of the EM algorithm was assessed through the comparison of the likelihood of the true and estimated parameters, and the comparison of mean number of EM steps needed when α is fixed or estimated. The quality of the estimates of β_0, $t_{1/2}$ and γ^2 was assessed using the coefficient of variation. The model selection procedure was evaluated by comparing the true number of shifts with the estimated one, which should be lower. We do not expect to find the exact number as some shifts, which are too small or too close of the tips, cannot be detected. To evaluate the quality of the clustering of the tips, that is the only observable quantity, we used the Adjusted Rand Index (ARI, Hubert and Arabie, 1985) between the true clustering of the tips, and the one induced by the estimated shifts. The ARI is proportional to the number of concordant pairs in two clusterings and has maximum value of 1 (for identical clusterings) and expected value of 0 (for random clusterings). Note that this score is conservative as shifts of small intensity, which are left aside by our model selection procedure, produce "artificial" groups that cannot be reconstructed.

5.4 Results

The selection strength is notoriously difficult to estimate, with large ranges of values giving similar behaviors (see Thomas et al., 2014). We hence first analyse the impact of estimating α on our estimations, showing that the main behavior of the algorithm stays the same. Then, we study the shifts reconstruction procedure.

Convergence and Likelihood. For α known, all estimations converged in less than 49 iterations, with a median number of 13 iterations. For α estimated, the number of iterations increased greatly, with a median of 69, and a fraction of estimations (around 3.3%) that reached the maximum allowed number (fixed at 1000 iterations) without converging. Unsurprisingly, the more difficult the problem was, the more iterations were needed. The log-likelihoods of the estimated parameters are close to the true ones, even when α is estimated (see supplementary Figure 13 in Appendix E, first row).

Estimation of continuous parameters. Figure 6 (first row) shows that we tend to slightly over-estimate α in general. The estimation is particularly bad for large values of α (with a high variance on the result, see first box of the row), and low values of α. In this regime, the model is "over-confident", as it finds a higher selection strength than the real one and therefore a smaller variance (second row of Figure 6). For smaller and bigger trees, the estimators behave in the same way, but with degraded or improved values, as expected. We also note that taking the true number of shifts instead of the estimated one slightly degrades our estimation of these parameters (see supplementary Figure 13 in Appendix E).

The estimation of β_0 is not affected by the knowledge of α or K (see Figure 7, first row), and only has an increased variance for more difficult configurations. In the remainder, we only show results obtained for estimated α as estimating α does not impact ARI, $\hat{K}$ and $\hat{\beta}_0$ (see supplementary Figure 14 in Appendix E).

Estimation of the number of shifts. The way shifts were drawn ensures that they are not too small in average, and that they are located all along the tree. Still, some shifts have a very small influence on the data, and are hence hard to detect (see Figure 5). The selection model procedure almost always under-estimates the number of shifts, except in very favorable cases (Figure 7, second row). This behavior is nonetheless expected, as allowing more shifts does not guarantee that the right shifts will be found (see supplementary Figures 11 and 12 in Appendix E).

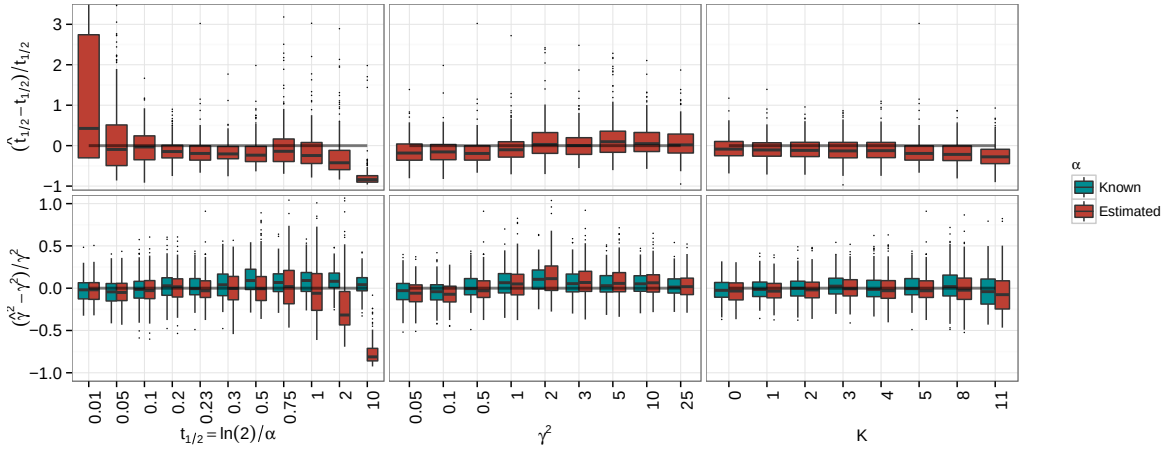


Figure 6: Box plots over the 200 repetitions of each set of parameters, for the phylogenetic half-life (top) and root variance (bottom) with K estimated, and α fixed to its true value (blue) or estimated (red), on a tree with 128 taxa. For better legibility, the y-axis of these two rows were re-scaled, omitting some outliers (respectively, for $t_{1/2}$ and γ^2, 0.82% and 0.46% of points are omitted). The whisker of the first box for $t_{1/2}$ goes up to 7.5.

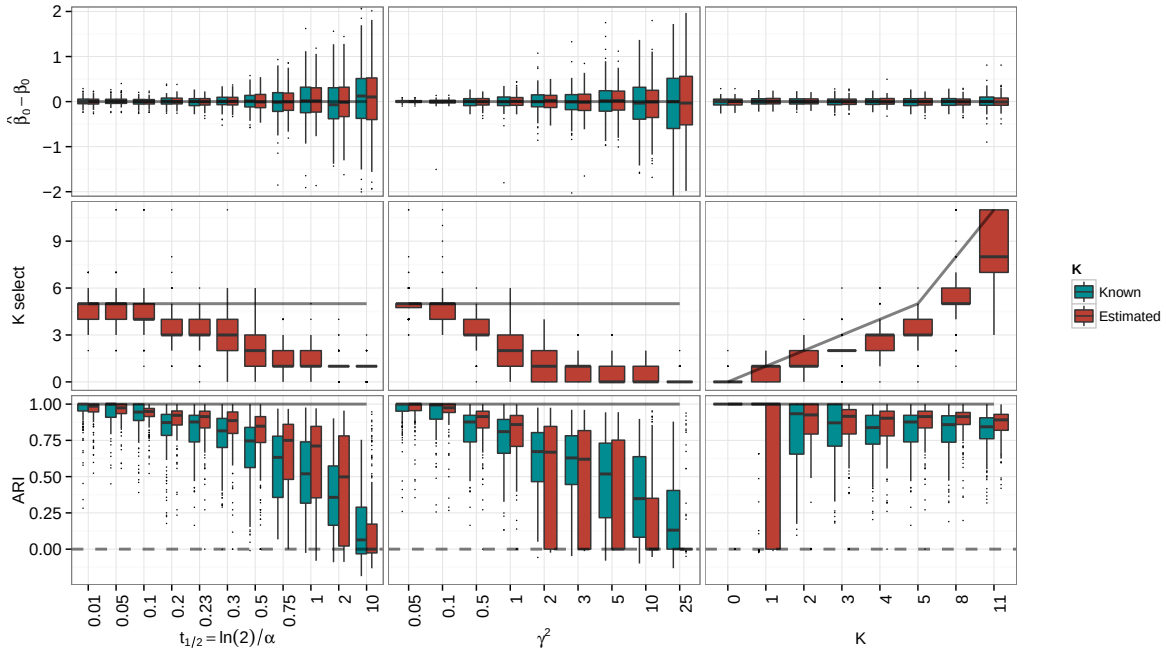


Figure 7: Same for β_0 (top), the number of shifts (middle) and ARI (bottom), with α estimated, and K fixed to its true value (blue) or estimated (red). As previously, the y-axis of the first row (β_0) was re-scaled, omitting some outliers (1.38% of points are omitted).

Clustering of the tips. The ARI tends to be degraded for small values of α or high variance, but remains positive (Figure 7, third row). When only one shift occurs, the ARI is very unstable, but for any other value of K, it stays quite high. Finally, knowing the number of shifts does not improve the ARI.

Equivalent Solutions. When α and K are both estimated, only 5.8% of the configurations have 2 or more equivalent solutions. One inferred configuration with three equivalent solutions

is presented Figure 5.

Comparison with bayou. As mentioned above, our simulation scheme, although not completely equivalent to the scheme used in Uyeda and Harmon (2014), is very similar, so that we can compare our results with theirs. The main differences lies in the facts that we took a grid on $\gamma^2 = \sigma^2/2\alpha$ instead of σ^2, and that we took shifts with higher intensities, making the detection of shifts easier. For each set of parameters, they estimated the posterior distribution using some fixed prior, two independent chains of $200,000$ generations, and a burning of $60,000$ generations. We can see that we get the same qualitative behaviors for our estimators, with the selection strength α over or under estimated, respectively, in small or large values regions. The main difference lies in the estimation of the number of shifts. Maybe because of the priors they used ($K \sim$ Conditional Poisson$(\lambda = 9, K_{max} = n/2)$), they tend to estimate similar numbers of shifts (centered on 9) for any set of parameters. In particular, while our method seems to be quite good at detecting situations where there are no shifts at all, theirs seems unable to catch these kind of configurations, despite the fact that their shifts have low intensity, leading to a possible over-fitting of the data.

Overall, the behavior of the algorithm is quite satisfying. Our model selection procedure avoids over-fitting, while recovering the correct clustering structure of the data. It furthermore allows for a reasonable estimation of the continuous parameters, except for α which is notoriously difficult to estimate.

6 Case Study: Chelonian Carapace Length Evolution

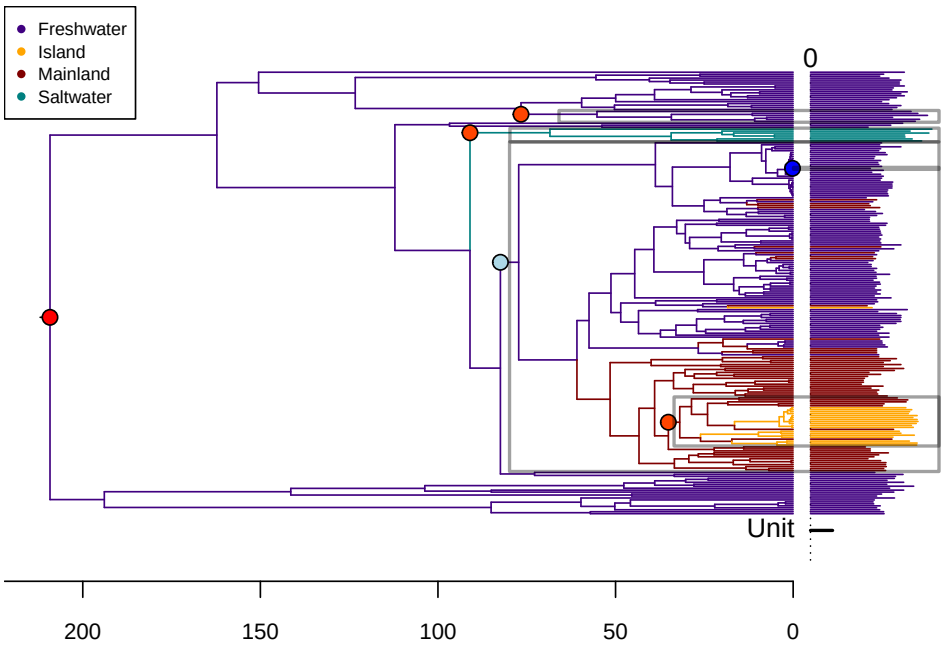


Figure 8: Phylogenetic tree of the Chelonians. Log-transformed traits values are represented on the right. Branch colors represent the habitats. The shifts found by our EM algorithm are shown as circles, with a color indicating the value of the shift, from blue (negative) to red (positive). Boxes highlight the groups induced by the shifts. The x-scale is in million years.

6.1 Description of the Dataset

Extant species of the order Testudines, or Chelonii, are turtles and tortoises, living all across the globe, and exhibiting a wide variation in body size, from the small desert speckled tortoise (*Homopus signatus*, 10 cm), to the large marine leatherback sea turtle (*Dermochelys coriacea*,

244 cm). In order to test the hypothesis of island and marine gigantism, that could explain the extreme variations observed, Jaffe et al. (2011) compiled a dataset containing a measure of the carapace length for 226 species, along with a phylogenetic tree of theses species, spanning 210 million years (my) (see Figure 8). They assigned each species to one of four habitats: mainland-terrestrial, freshwater, marine and island-terrestrial. Then, testing several fixed regimes allocations on the branches of the tree using the method described in Butler and King (2004), they found the best support in favor of a "OU2" model that assigned one regime to each habitat. Following Uyeda and Harmon (2014), we will refer to this model as the "$\mathrm{OU}_{habitat}$" model. Note that this model is ambiguously defined, as it requires to assign a habitat to each ancestral species. Using proposition 3.3, we found that there were 48 equivalent parsimonious ways of doing so that respect the habitats observed at the tips of the tree. One of these habitat reconstruction is presented Figure 8.

6.2 Method

We used the version of the dataset embedded in the package geiger (Harmon et al., 2008), that contains a phylogenetic tree and a vector of log-carapaces lengths. The corresponding habitats are reported in the Appendix of Jaffe et al. (2011).

We ran our algorithm with a number of shifts going from 0 to 20 and α varying on a grid: $\alpha \in \{0.01, 0.02, 0.04, 0.06, 0.08, 0.1\}$, but fixed for each estimation. These $6 \times 20 = 120$ estimations took around 2 hours of CPU time. For each number of shifts, going again from 0 to 20, we kept the solution with the maximal likelihood, and we applied the model selection criterion to them. This method gave a solution with 5 shifts, and a selection strength of 0.06 (i.e. 5.5 % of the total height of the tree). Using a finer grid for α gives highly similar results, allocating shifts to the same edges. These last estimations are given below.

6.3 Results

Our method selected a solution with 5 shifts, a rather strong selection strength ($t_{1/2} = 5.4\%$ of the tree height), and a rather low root variance ($\gamma^2 = 0.22$, see table 1, first column). Two of those shifts are closely related to the habitats defined in Jaffe et al. (2011) (see Figure 8). The ancestral optimal value, that applies here to two clades of freshwater turtles, is estimated to be around 38 cm. A small decrease in size for a large number of mainland and freshwater turtles is found (optimal value 24 cm). Marine turtles (super-family Chelonioidea) are found to have an increased carapace length (with an optimal value of 130 cm), as well as a clade containing soft-shell tortes (family Trionychidae, optimal size 110 cm), and a clade containing almost all island tortoises, including several sub-species of Galápagos tortoises (*Geochelone nigra*). Only the Ryukyu black-breasted leaf turtle (*Geoemyda japonica*), endemic to the Ryukyu Islands in Japan, and distant on the phylogenetic tree, is not included in this group. Note that the group also contains some mainland tortoises of the genus Geochelone, that are closely related to Galápagos tortoises. This is typical of our method: it constructs groups that are both phenotypically and phylogenetically coherent. Finally, one species is found to have its own group, the black-knobbed map turtle (*Graptemys nigrinoda*), with a very low optimal value of 1.4×10^{-20} cm, for a measured trait of 15 cm. The fact that the shift has a very high negative value (-49 in log scale) is just an artifact due to the actualization factor on a very small branch (0.18 my, for an inferred phylogenetic half-life of 11 my). This is a rather unexpected choice of shift location. When considering the linear model as transformed by the cholesky matrix of the variance to get independent errors (as in the proof of proposition 4.1), we find a leverage of 0.94, indicating that this species trait behaves in the transformed space as an outsider.

6.4 Comparison with other methods

Methods. In order to compare our results to previously published ones, we reproduced some of the analysis already conducted on this dataset. We first fitted an $\mathrm{OU}_{habitat}$ model with fixed regimes as in Jaffe et al. (2011), using the R package OUwie (Beaulieu et al., 2012). We tested

all of the 48 possible ways of allocating internal nodes, and took the solution with the highest likelihood. Using the package bayou (Uyeda and Harmon, 2014), we reproduced the Bayesian analysis of the data, using two independent chains of 500000 generations each, discarding the first 150000 generations as burning. We assigned the priors that were used in the original study on the parameters, namely: $P(\alpha) \sim \text{LogNormal}(\ln \mu = -5, \ln \sigma = 2.5)$, $P(\sigma^2) \sim \text{LogNormal}(\ln \mu = 0, \ln \sigma = 2)$, $P(\beta_i) \sim \text{Normal}(\mu = 3.5, \sigma = 1.5)$, $P(K) \sim$ Conditional Poisson$(\lambda = 15, K_{max} = 113)$. The computations took around 2.3 hours of CPU time. We also ran the stepwise-AIC method SURFACE (Ingram and Mahler, 2013), that relies on a forward-backward procedure. This took around 11 hours of CPU time. Finally, we ran the function OUshifts from the package phylolm, that uses a modified BIC criterion and an heuristic stepwise procedure to detect shifts (Ho and Ané, 2014). Thanks to an efficient linear algorithm, detailed in Ho and Ané (2013b), this function is pretty fast, taking only about 8 minutes of CPU time. Note that the model used in these last three methods (bayou, SURFACE and OUshifts) are slightly different from ours, as they assume that the root is fixed to the ancestral optimum state, and not drawn from its stationary distribution.

Results. The shifts allocated on the tree by methods bayou, SURFACE and OUshifts are presented on Supplementary Figure 15 (Appendix E). We can see that 3 among the most strongly supported shifts in the posterior distribution given by bayou, as well as some among the oldest shifts found by SURFACE and OUshifts are similar to the ones found by our method. The bayou method finds equal support for many shifts, all over the tree, and the median of the posterior distribution is 17 shifts, which is pretty close to the mode of the prior put on the number of shifts (15) . The SURFACE and OUshifts methods select respectively 33 and 8 shifts, including many on pendant edges, that are not easily interpretable. The backward step of SURFACE allowed to merge the regimes found for marine turtles and soft-shell tortoises that our method found to have very similar optimal values. The results of the five methods studied are summarized Table 1. Note that these four models are not nested, due to the status assigned to the root state, and to the possible convergences.

Compared to step-wise heuristics, our integrated maximum likelihood based approach allows us to have a more "global" view of the tree, and hence to select a solution that accounts better for the global structure of the trait distribution. Thanks to its rigorous model selection procedure, our model seems to report significant shifts only, that are more easily interpretable than the solutions found by other methods, and that do not rely on any chosen prior.

	Habitat	EM	bayou	SURFACE	OUshifts
Number of shifts	16	5	17	33	8
Number of regimes	4	6	18	13	9
lnL	-133.86	-97.59	-91.54	30.38	-79.79
Marginal lnL	NA	NA	-149.09	NA	NA
α ($\times h$, per my)	9.32	12.76	36.54	1.72	3.25
$\ln 2/\alpha$ (my)	15.56	11.36	3.97	84.28	44.64
σ^2 ($\times h$, per my)	6.21	5.57	11.91	0.72	2.29
γ^2	0.33	0.22	0.16	0.21	0.35
CPU time (min)	65.25	134.49	136.81	634.16	8.28

Table 1: Summary of the results obtained with several methods for the Chelonian Dataset. For bayou, the median of the posterior distributions is given.

Acknowledgments

We would like to thank Cécile Ané for helpful discussions on an early draft of this manuscript. We are grateful to the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.

fr) for providing the computational resources needed for these experiments. We also thank the two anonymous reviewers whose careful and critical reading greatly helped improve this manuscript.

References

Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian Model Selection with an Unknown Variance. *The Annals of Statistics*, 37(2):630–672, April 2009.

Yannick Baraud, Annie Bouvier, Christophe Giraud, and Sylvie Huet. LINselect: R Package for Estimator Selection. 2013.

Krzysztof Bartoszek and Serik Sagitov. Phylogenetic confidence intervals for the optimal trait value. *Journal of Applied Probability*, 52(4):1115–1132, December 2015.

Krzysztof Bartoszek, Jason Pienaar, Petter Mostad, Staffan Andersson, and Thomas F. Hansen. A Phylogenetic Comparative Method for Studying Multivariate Adaptation. *Journal of Theoretical Biology*, 314:204–215, December 2012.

Jeremy M. Beaulieu, Dwueng-Chwuan Jhwueng, Carl Boettiger, and Brian C. O'Meara. Modeling Stabilizing Selection: Expanding The Ornstein–Uhlenbeck Model Of Adaptive Evolution. *Evolution*, 66(8):2369–2383, August 2012.

Marguerite A. Butler and Aaron A. King. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *The American Naturalist*, 164(6):683–695, December 2004.

Clayton E. Cressler, Marguerite A. Butler, and Aaron A. King. Detecting Adaptive Evolution in Phylogenetic Comparative Analysis Using the Ornstein-Uhlenbeck Model. *Systematic biology*, 64(6):953–958, June 2015.

Charles C. Davis, Maribeth Latvis, Daniel L. Nickrent, Kenneth J. Wurdack, and David A. Baum. Floral Gigantism in Rafflesiaceae. *Science*, 315(5820):1812–1812, March 2007.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

Jonathan M. Eastman, Daniel Wegmann, Christoph Leuenberger, and Luke J. Harmon. Simpsonian 'Evolution by Jumps' in an Adaptive Radiation of Anolis Lizards. *ArXiv e-print*, May 2013.

Joseph Felsenstein. Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1–15, January 1985.

Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Suderland, USA, 2004.

Walter M. Fitch. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology*, 20(4):406–416, 1971.

Chris Fraley, Adrian E. Raftery, T. Brendan Murphy, and Luca Scrucca. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical report, 2012.

Yves Grandvalet, Julien Chiquet, and Christophe Ambroise. Sparsity by Worst-Case Quadratic Penalties. *arXiv e-print*, October 2012.

Thomas F. Hansen. Stabilizing Selection and the Comparative Analysis of Adaptation. *Evolution*, 51 (5):1341–1351, October 1997.

Thomas F. Hansen and Krzysztof Bartoszek. Interpreting the Evolutionary Regression: the Interplay between Observational and Biological Errors in Phylogenetic Comparative Studies. *Systematic Biology*, 61(3):413–425, May 2012.

Thomas F. Hansen, Jason Pienaar, and Steven Hecht Orzack. A Comparative Method for Studying Adaptation to a Randomly Evolving Environment. *Evolution*, 62(8):1965–1977, 2008.

Luke J. Harmon, Jason T. Weir, Chad D. Brock, Richard E. Glor, and Wendell Challenger. GEIGER: investigating evolutionary radiations. *Bioinformatics*, 24(1):129–131, 2008.

Gordon Hiscott, Colin Fox, Matthew Parry, and David Bryant. Efficient recycled algorithms for quantitative trait models on phylogenies. *ArXiv e-print*, December 2015.

Lam Si Tung Ho and Cécile Ané. Asymptotic Theory with Hierarchical Autocorrelation: Ornstein-Uhlenbeck Tree Models. *The Annals of Statistics*, 41(2):957–981, 2013a.

Lam Si Tung Ho and Cécile Ané. A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. *Systematic Biology*, 63(3):397–408, 2013b.

Lam Si Tung Ho and Cécile Ané. Intrinsic Inference Difficulties for Trait Evolution with Ornstein-Uhlenbeck Models. *Methods in Ecology and Evolution*, 5(11):1133–1146, November 2014.

Lawrence Hubert and Phipps Arabie. Comparing Partitions. *Journal of Classification*, 2(1):193–218, 1985.

Travis Ingram and D. Luke Mahler. SURFACE: Detecting Convergent Evolution from Comparative Data by Fitting Ornstein-Uhlenbeck Models with Stepwise Akaike Information Criterion. *Methods in Ecology and Evolution*, 4(5):416–425, May 2013.

Alexander L. Jaffe, Graham J. Slater, and Michael E. Alfaro. The Evolution of Island Gigantism and Body Size Variation in Tortoises and Turtles. *Biology letters*, 11(11), November 2011.

Michael J. Landis, Joshua G. Schraiber, and Mason Liang. Phylogenetic Analysis Using Lévy Processes: Finding Jumps in the Evolution of Continuous Traits. *Systematic Biology*, 62(2):193–204, March 2013.

Nicolas Lartillot. A Phylogenetic Kalman Filter for Ancestral Trait Reconstruction Using Molecular Data. *Bioinformatics*, 30(4):488–496, February 2014.

Jonathan B. Losos. A Phylogenetic Analysis of Character Displacement in Caribbean Anolis Lizards. *Evolution*, 44(3):558–569, May 1990.

D. Luke Mahler, Travis Ingram, Liam J. Revell, and Jonathan B. Losos. Exceptional Convergence on the Macroevolutionary Landscape in Island Lizard Radiations. *Science*, 341(6143):292–295, July 2013.

Elchanan Mossel and Mike Steel. A Phase Transition for a Random Cluster Model on Phylogenetic Trees. *Mathematical Biosciences*, 187(2):189–203, February 2004.

Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20(2):289–290, July 2004.

Matthew W. Pennell and Luke J. Harmon. An Integrative View of Phylogenetic Comparative Methods: Connections to Population Genetics, Community Ecology, and Paleobiology. *Annals of the New York Academy of Sciences*, 1289:90–105, June 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.r-project.org.

Daniel L. Rabosky. Automatic Detection of Key Innovations, Rate Shifts, and Diversity-Dependence on Phylogenetic Trees. *PLoS ONE*, 9(2), February 2014.

Peter Rousseeuw, Christophe Croux, Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Tobias Verbeke, Manuel Koller, and Martin Maechler. *robustbase: Basic Robust Statistics.*, 2014.

Serik Sagitov and Krzysztof Bartoszek. Interspecies Correlation for Neutrally Evolving Traits. *Journal of Theoretical Biology*, 309(21):11–19, September 2012.

David Sankoff. Minimal Mutation Trees of Sequences. *SIAM Journal on Applied Mathematics*, 28(1): 35–42, 1975.

Charles Semple and Mike A. Steel. *Phylogenetics*, volume 24 of *Oxford lecture series in mathematics and its applications*. Oxford University Press, 2003.

Tanja Stadler. TreeSim: Simulating Trees under the Birth-Death Model. 2014. URL http://cran.r-project.org/package=TreeSim. R package version 2.1.

Michael Steel. The Complexity of Reconstructing Trees from Qualitative Characters and Subtrees. *Journal of classification*, 9(1):91–116, January 1992.

Gavin Huw Thomas, Natalie Cooper, Chris Venditti, Andrew Meade, and Robert P Freckleton. Bias and measurement error in comparative analyses: a case study with the Ornstein Uhlenbeck model. *bioRxiv*, April 2014.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Josef C. Uyeda and Luke J. Harmon. A Novel Bayesian Method for Inferring and Interpreting the Dynamics of Adaptive Landscapes from Phylogenetic Comparative Data. *Systematic Biology*, 63(6): 902–918, July 2014.

Steve Weston. *doParallel: Foreach parallel adaptor for the parallel package*, 2014a. URL http://cran.r-project.org/package=doParallel. R package version 1.0.8.

Steve Weston. *foreach: Foreach looping construct for R*, 2014b. URL http://cran.r-project.org/package=foreach. R package version 1.4.2.

Hadley Wickham. Reshaping Data with the reshape Package. *Journal of Statistical Software*, 21(12): 1–20, 2007. URL http://www.jstatsoft.org/v21/i12.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Use R! Springer-Verlag New York, 2009.

A Enumeration of Equivalence Classes

Definition A.1 (Coloring concatenation)**.** Let i be a node of tree $\mathcal{T}$ with L_i daughter nodes $(i_1, \cdots, i_{L_i})$, $L_i \geq 2$, and assume that the tips are colored according to the application $d \in (\mathcal{C}_K)^n$. We denote by $\mathcal{T}_i$ the sub-tree rooted at node i, and by $\mathcal{A}_{\mathcal{T}_i}(d)$ the set of parsimonious shifts allocations on $\mathcal{T}_i$ that produce a coloring of the tips compatible with d. At the root, $\mathcal{A}_{\mathcal{T}_1}(d) = \mathcal{A}_{\mathcal{T}}(d) = \phi^{-1}(d)$.

- For $k \in \mathcal{C}_K$, $S_i(k)$ is the *cost* of starting from node i with color k, *i.e.* the minimal number of shifts needed to get the right coloring of the tips of $\mathcal{T}_i$, when starting with color k.
- $S_i^{tot} = \min_{k \in \mathcal{C}_K} S_i(k)$ is the minimal cost of subtree $\mathcal{T}_i$, *i.e.* the number of shifts of a parsimonious coloring. $\mathcal{L}_i = \operatorname{argmin}_{k \in \mathcal{C}_K} S_i(k)$ is the set of colors root i can take in a parsimonious coloring of sub-tree $\mathcal{T}_i$.
- For $k \in \mathcal{C}_K$, $\mathcal{B}^k_{\mathcal{T}_i}$ is the set of colorings of $\mathcal{T}_i$ that respect the colors at the tips, have $S_i(k)$ shift, and start with color k.
- For $\mathcal{K} \subset \mathcal{C}_K$, $\mathcal{B}^{\mathcal{K}}_{\mathcal{T}_i} = \bigcup_{k \in \mathcal{K}} \mathcal{B}^k_{\mathcal{T}_i}$. Hence, $\mathcal{A}_{\mathcal{T}_i}(d) = \mathcal{B}^{\mathcal{L}_i}_{\mathcal{T}_i}$, and the computation of $\mathcal{A}_{\mathcal{T}_i}(d)$ only requires the computation of $S_i(k)$ and $\mathcal{B}^k_{\mathcal{T}_i}$ for any $k \in \mathcal{C}_K$.
- For $(p_1, \cdots, p_{L_i}) \in (\mathcal{C}_K)^{L_i}$, and $(B_1, \cdots, B_{L_i}) \in \mathcal{B}^{p_1}_{\mathcal{T}_{i_1}} \times \cdots \times \mathcal{B}^{p_{L_i}}_{\mathcal{T}_{i_{L_i}}}$ we define, for $k \in \mathcal{C}_K$, the concatenation $B = {}^k\bigoplus_{l=1}^{L_i} B_l, B \in \mathcal{B}^k_{\mathcal{T}_i}$, by: $\begin{cases} B(i) = k \\ B(j) = B_l(j) & \text{if } i \in \mathcal{T}_{i_l} \end{cases}$. As the sub-trees $\mathcal{T}_{i_l}$, $l \in [\![1\,, L_i]\!]$ do not overlap, this application is correctly defined on the nodes of $\mathcal{T}_i$.

Using these definitions, we can state the following recursion formula:

Proposition A.1 (Enumeration Recursion Formula)**.** *Let $k \in \mathcal{C}_K$, and $i \in [\![1\,, m+n]\!]$. If i is a tip of the tree, then*

$$S_i(k) = \begin{cases} 0 & \text{if } d(i) = k \\ +\infty & \text{otherwise} \end{cases} \qquad \mathcal{B}^k_{\mathcal{T}_i} = \begin{cases} \{i \mapsto k\} & \text{if } d(i) = k \\ \emptyset & \text{otherwise} \end{cases}$$

If i is a node of tree $\mathcal{T}$ with L_i daughter nodes $(i_1, \cdots, i_{L_i})$, $L_i \geq 2$, and assuming that $S_{i_l}(k)$ and $\mathcal{B}^k_{\mathcal{T}_{i_l}}$ are known for any $l \in [\![1\,, L_i]\!]$ and $k \in \mathcal{C}_K$, define, for $l \in [\![1\,, L_i]\!]$:

$$\mathcal{K}^l_k = \operatorname*{argmin}_{1 \leq p \leq K} \{S_{i_l}(p) + \mathbb{I}\{p \neq k\}\}$$

As these sets are not empty, let $(p_1, \ldots p_{L_i}) \in \mathcal{K}^1_k \times \ldots \times \mathcal{K}^{L_i}_k$. Then

$$S_i(k) = \sum_{l=1}^{L_i} (S_{i_l}(p_l) + \mathbb{I}\{p_l \neq k\}) \quad \text{and} \quad \mathcal{B}^k_{\mathcal{T}_i} = \left\{ {}^k\bigoplus_{l=1}^{L_i} B_l : \forall l \in [\![1\,, L_i]\!], B_l \in \mathcal{B}^{\mathcal{K}^l_k}_{\mathcal{T}_{i_l}} \right\}$$

Proof. The actualization of $S_i(k)$ is the same as in the Sankoff algorithm (Sankoff, 1975). The set $\mathcal{B}^k_{\mathcal{T}_i}$ is then obtained by enumerating all the possible ways of concatenating children sets $\mathcal{B}^{\mathcal{K}^l_k}_{\mathcal{T}_{i_l}}$, each of which is the ensemble of solutions for the sub-tree $\mathcal{T}_{i_l}$ that realize the minimal number of shifts when starting in state k. □

Remarking that $T_i(k) = \left|\mathcal{B}^k_{\mathcal{T}_i}\right|$, proposition 3.3 of the main text follows immediately.

B A Vandermonde Like Identity

Proposition B.1. *Let $(n, n') \in N$ and $K \in \mathbb{N}$. With the standard convention that $\binom{n}{k} = 0$ if $n < k$,*

$$\binom{n+n'-K}{K} = \sum_{k=0}^{K} \binom{n-k}{k}\binom{n'-K+k}{K-k} + \sum_{k=0}^{K-1} \binom{(n-1)-k}{k}\binom{(n'-1)-(K-1)+k}{(K-1)-k}$$

which can be rewritten in a more symmetric way as:

$$\binom{n+n'-K}{K} = \sum_{k,k'\geq 0: k+k'=K} \binom{n-k}{k}\binom{n'-k'}{k'} + \sum_{k,k'\geq 0: k+k'=K-1} \binom{(n-1)-k}{k}\binom{(n'-1)-k'}{k'} \tag{B.1}$$

Similarly,

$$\binom{n+n'+1-K}{K} = \sum_{k=0}^{K} \binom{n-k}{k}\binom{n'-K+k}{K-k} + \sum_{k=0}^{K-1} \binom{(n-1)-k}{k}\binom{n'-(K-1)+k}{(K-1)-k} + \binom{n-k}{k}\binom{(n'-1)-(K-1)+k}{(K-1)-k}$$

which can be rewritten in a more symmetric way as:

$$\binom{n+n'+1-K}{K} = \sum_{k,k'\geq 0: k+k'=K} \binom{n-k}{k}\binom{n'-k'}{k'} + \sum_{k,k'\geq 0: k+k'=K-1} \binom{(n-1)-k}{k}\binom{n'-k'}{k'} + \binom{n-k}{k}\binom{(n'-1)-k'}{k'} \tag{B.2}$$

Note that Eq (B.1) generalizes in some way the Vandermonde identity which states

$$\binom{n+n'}{K} = \sum_{k=0}^{K} \binom{n}{k}\binom{n'}{K-k} \tag{B.3}$$

Although several proofs of the Vandermonde identity are known (geometric, algebraic and combinatorial), we only provide a geometric proof of this Vandermonde-like identity.

Consider a grid of size $(n+n') \times K$. We are interested in grid-valued paths that can move either by $(1,0)$ or by $(2,1)$. In other words, if the kth position of a path is (x_k, y_k), then its next position (x_{k+1}, y_{k+1}) is either (x_k+1, y_k) or (x_k+2, y_k+1). We are interested in paths starting at $(0,0)$ and ending at $(n+n', K)$.

Such a path consists of K moves of type $(2,1)$ and $n+n'-2K$ moves of type $(1,0)$ and is uniquely determined by the positions of the moves of the former type. There are $\binom{n+n'-2K+K}{K} = \binom{n+n'-K}{K}$ distinct positions and therefore as many such paths.

We now sort the paths according to the value i they take when either reaching the line $x = n$ or reaching the line $x = n+1$ without reaching the line $x = n$ first. We refer to the latter paths as crossing the line $x = n$. Note that this sorting induces a partition of all paths (see Figure 9)

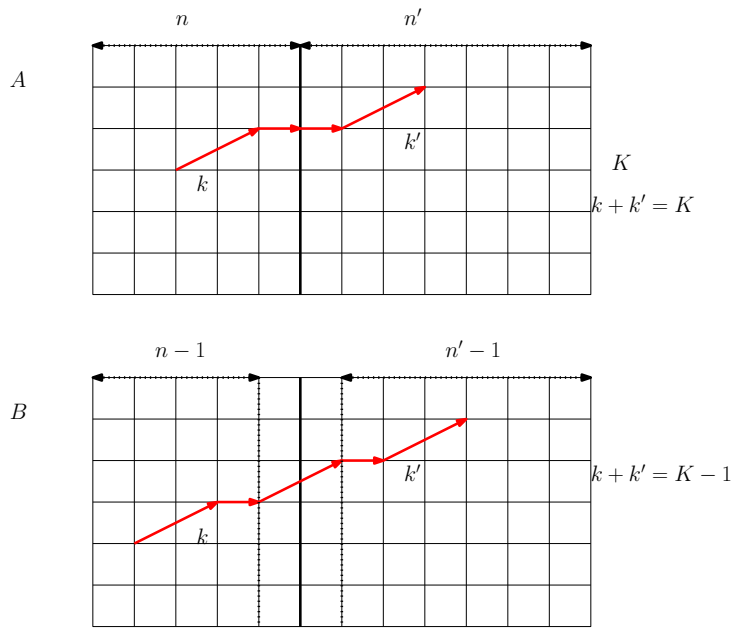


Figure 9: Partition of paths according to whether they reach (A) or cross (B) the line $x = n$

A path reaching $x = n$ at position i uniquely gives rise to two paths: one from $(0, 0)$ to (n, i) and one from (n, i) to $(n + n', K)$ or equivalently from 0 to $(n', K - i)$. There are $\binom{n-i}{i}$ different paths of the first kind and $\binom{n'-K-i}{K-i}$ of the second. There are therefore $\binom{n-i}{i}\binom{n'-K+i}{K-i}$ paths that pass through (n, i).

A path crossing the line $x = n$ and reaching the line $x = n + 1$ at i must do so with a last move of type $(2, 1)$. It therefore uniquely defines a path from $(0, 0)$ to $(n - 1, i - 1)$ and a path from $(n + 1, i)$ to $(n + n', K)$, or equivalently from $(0, 0)$ to $(n' - 1, K - i)$. There are therefore $\binom{n-i}{i-1}\binom{n'-1-K+i}{K-i}$ paths that cross the line $x = n$ and pass through $(n + 1, i)$.

Putting everything together, we get:

$$\begin{aligned}\binom{n+n'-K}{K} &= \sum_{i=0}^{K}\binom{n-i}{i}\binom{n'-K+i}{K-i} + \sum_{i=0}^{K}\binom{n-i}{i-1}\binom{n'-1-K+i}{K-i} \\ &= \sum_{i=0}^{K}\binom{n-i}{i}\binom{n'-K+i}{K-i} + \sum_{i=0}^{K-1}\binom{(n-1)-i}{i}\binom{(n'-1)-(K-1)+i}{(K-1)-i}\end{aligned}$$

which is exactly Eq. (B.1).

To prove Eq. (B.2), we start from a grid of size $(n + n' + 1) \times K$ and are again interested in the paths starting from the bottom left corner and ending in the upper right corner using only (2, 1) and (1, 0) moves. These paths have exactly K moves of type (2, 1) and there are $\binom{n+n'+1-K}{K}$ of them. This time, we partition paths upon the move observed between $x = n$ and $x = (n + 1)$.

The move can be (see also Figure 10):

- (1, 0), in which case k (resp. k') moves of type $(2, 1)$ are used in the interval $[1, n]$ (resp. $[n + 1, n + n'1]$) such that $k + k' = K$;
- (2, 1) starting from $x = n$ and therefore ending at $x = n + 2$, in which case k (resp. k') moves of type $(2, 1)$ are used in the interval $[1, n]$ (resp. $[n + 2, n + n'1]$) such that $k + k' = K - 1$ (one move (2, 1) has already been consumed);
- (2, 1) ending at $x = n + 1$ and therefore starting from $x = n - 1$ in which case k (resp. k') moves of type $(2, 1)$ are used in the interval $[1, n - 1]$ (resp. $[n + 1, n + n'1]$) such that $k + k' = K - 1$ (one move (2, 1) has already been consumed);

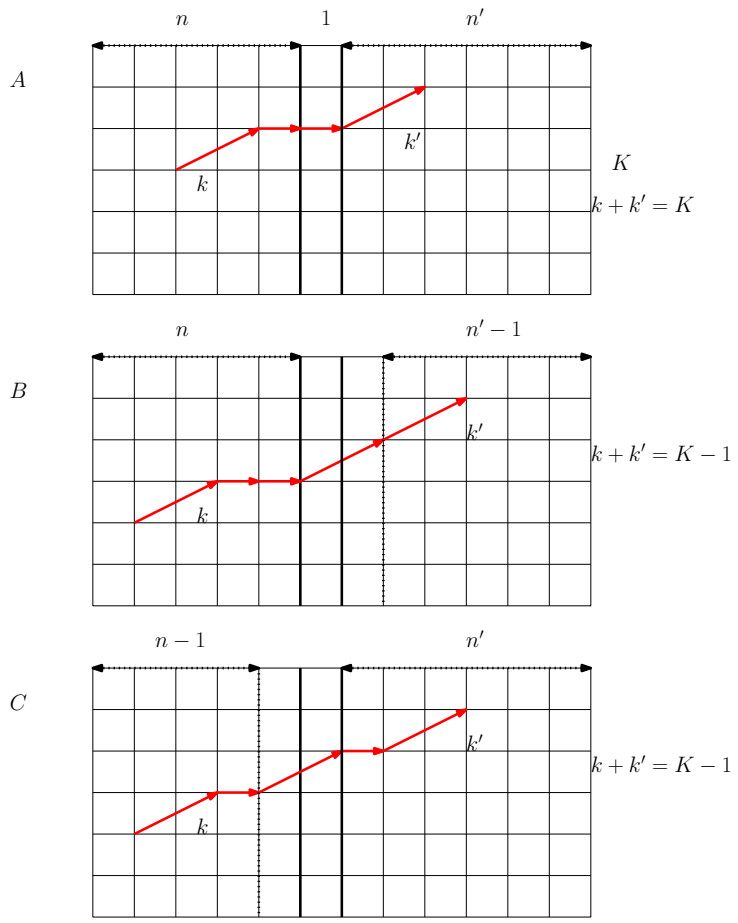


Figure 10: Partition of paths according to whether to the move used between $x = n$ and $x = n + 1$. Cases A, B and C correspond to the items listed in the main text.

Wrapping everything together and using the same arguments as before, we get Eq. (B.2).

C Technical Details of the EM

C.1 E Step

Given a set of parameters $\boldsymbol{\theta}^{(h)}$, we have:

$$\mathbf{X} = (\mathbf{Z}, \mathbf{Y}) \sim \mathcal{N}\left(\mathbf{m}^{(h)} = \begin{pmatrix} \mathbf{m}_{\mathbf{Z}}^{(h)} \\ \mathbf{m}_{\mathbf{Y}}^{(h)} \end{pmatrix}, \ \boldsymbol{\Sigma}^{(h)} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{ZZ}}^{(h)} & \boldsymbol{\Sigma}_{\mathbf{YZ}}^{(h)} \\ \boldsymbol{\Sigma}_{\mathbf{ZY}}^{(h)} & \boldsymbol{\Sigma}_{\mathbf{YY}}^{(h)} \end{pmatrix}\right)$$

hence:

$$\begin{aligned} \mathbf{Z} \mid \mathbf{Y} \sim \mathcal{N}\big(\mathbf{m}_{\mathbf{Z}|\mathbf{Y}}^{(h)} &= \mathbf{m}_{\mathbf{Z}}^{(h)} + \boldsymbol{\Sigma}_{\mathbf{ZY}}^{(h)}(\boldsymbol{\Sigma}_{\mathbf{YY}}^{(h)})^{-1}(\mathbf{Y} - \mathbf{m}_{\mathbf{Y}}^{(h)}), \\ \boldsymbol{\Sigma}_{\mathbf{Z}|\mathbf{Y}} &= \boldsymbol{\Sigma}_{\mathbf{ZZ}}^{(h)} - \boldsymbol{\Sigma}_{\mathbf{ZY}}^{(h)}(\boldsymbol{\Sigma}_{\mathbf{YY}}^{(h)})^{-1}\boldsymbol{\Sigma}_{\mathbf{YZ}}^{(h)}\big) \end{aligned}$$

Remark C.1. We can see that this approach forces us to invert $\boldsymbol{\Sigma}_{\mathbf{YY}}^{(h)}$, a $n \times n$ matrix, which is a costly operation, of order $O(n^3)$. It also computes the complete matrix $\boldsymbol{\Sigma}_{\mathbf{Z}|\mathbf{Y}}$ whereas we only need a linear number of its coefficients: conditional variances and covariances of the form $\mathbb{C}\mathrm{ov}\left[\mathbf{Z}_i; \mathbf{Z}_{\mathrm{pa}(i)} \mid \mathbf{Y}\right]$. Due to the tree structure and to the Gaussian nature of the processes studied, it is possible to compute all the quantities needed in a linear time, using a "forward-backward"-like algorithm (here, "upward-downward", see Lartillot (2014) for a similar algorithm.). The upward step is similar to the pruning algorithm described in Felsenstein (2004, chap. 23). See also Ho and Ané (2013b) for an algorithm linear in the number of iterations.

C.2 Complete Likelihood Computation

Using the incomplete data model described in section 2.2, we can write:

$$p_{\boldsymbol{\theta}}(\mathbf{X}) = p_{\boldsymbol{\theta}}(X_1) \prod_{j=2}^{m+n} p_{\boldsymbol{\theta}}\left(X_j \mid X_{\mathrm{pa}(j)}\right)$$

Taking the expectation, we get for the BM:

$$\begin{aligned} -2\mathbb{E}\left[\log p_{\boldsymbol{\theta}}(\mathbf{X}) \mid \mathbf{Y}\right] = A + \log\gamma^2 + \frac{1}{\gamma^2}\left(\mathbb{V}\mathrm{ar}\left[X_1 \mid \mathbf{Y}\right] + \left(\mathbb{E}\left[X_1 \mid \mathbf{Y}\right] - \mu\right)^2\right) \\ + (m+n-1)\log\sigma^2 + \frac{1}{\sigma^2}\sum_{j=2}^{m+n} \ell_j^{-1}\mathbb{V}\mathrm{ar}\left[X_j - X_{\mathrm{pa}(j)} \mid \mathbf{Y}\right] \\ + \frac{1}{\sigma^2}\sum_{j=2}^{m+n} C_j^{BM}(\boldsymbol{\Delta}) \end{aligned} \tag{C.1}$$

and, for the OUsun:

$$\begin{aligned} -2\mathbb{E}\left[\log p_{\boldsymbol{\theta}}(\mathbf{X}) \mid \mathbf{Y}\right] = B + \sum_{j=2}^{m+n} \log c_j(\alpha) + \frac{1}{\gamma^2}\mathbb{V}\mathrm{ar}\left[X_1 \mid \mathbf{Y}\right] + (m+n)\log\gamma^2 \\ + \frac{1}{\gamma^2}\sum_{j=2}^{m+n} c_j(\alpha)^{-1}\mathbb{V}\mathrm{ar}\left[X_j - X_{\mathrm{pa}(j)}e_j \mid \mathbf{Y}\right] \\ + \frac{1}{\gamma^2}\left(\mathbb{E}\left[X_1 \mid \mathbf{Y}\right] - \beta_0\right)^2 + \frac{1}{\gamma^2}\sum_{j=2}^{m+n} C_j^{OU}(\alpha, \boldsymbol{\Delta}) \end{aligned} \tag{C.2}$$

where A and B are constants, and for each node j, $j \in [\![2\,, m+n]\!]$, we define an actualization factor $c_j(\alpha) = 1 - e_j(\alpha)^2$, with $e_j(\alpha) = e^{-\alpha\ell_j}$, and C_j^{BM} and C_j^{OU} are *costs* associated with branch b_j:

$$\begin{cases} C_j^{BM}(\boldsymbol{\Delta}) = \dfrac{1}{\ell_j}\left(\mathbb{E}\left[X_j \mid \mathbf{Y}\right] - \mathbb{E}\left[X_{\mathrm{pa}(j)} \mid \mathbf{Y}\right] - \Delta_i\right)^2 \\ C_j^{OU}(\alpha, \boldsymbol{\Delta}) = \dfrac{1}{c_j(\alpha)}\left(\mathbb{E}\left[X_j \mid \mathbf{Y}\right] - e_j\mathbb{E}\left[X_{\mathrm{pa}(j)} \mid \mathbf{Y}\right] - \beta_j\left(1 - e_j\right)\right)^2 \end{cases}$$

C.3 M step

Assuming that $p_{\boldsymbol{\theta}^{(h)}}(\mathbf{Z} \mid \mathbf{Y})$ is known, we need to compute $\boldsymbol{\theta}^{(h+1)}$ by maximizing $\mathbb{E}_{\boldsymbol{\theta}^{(h)}}[\log p_{\boldsymbol{\theta}}(\mathbf{Z}, \mathbf{Y}) \mid \mathbf{Y}]$. We have to deal with parameters of different nature, discrete or continuous. For a given vector $\boldsymbol{\Delta}^{(h+1)}$ of K non-zero shifts, we can exhibit closed formulas for $\mu^{(h+1)}$, $\sigma^{(h+1)}$ and $\gamma^{(h+1)}$, for the BM:

$$\begin{cases} \mu^{(h+1)} = \mathbb{E}^{(h)}[Z_1 \mid \mathbf{Y}] \\ \gamma^{2(h+1)} = \mathbb{V}\mathrm{ar}^{(h)}[Z_1 \mid \mathbf{Y}] \\ \sigma^{2(h+1)} = \dfrac{1}{m+n-1}\left[\displaystyle\sum_{j=2}^{m+n} \ell_j^{-1}\mathbb{V}\mathrm{ar}^{(h)}\left[X_j - X_{\mathrm{pa}(j)} \mid \mathbf{Y}\right] + C_j^{BM}\left(\boldsymbol{\Delta}^{(h+1)}\right)\right] \end{cases}$$

and, for the OUsun:

$$\begin{aligned} (m+n)\gamma^{2(h+1)} = \mathbb{V}\mathrm{ar}^{(h)}[X_1 \mid \mathbf{Y}] + \sum_{j=2}^{m+n} c_j(\alpha)^{-1}\mathbb{V}\mathrm{ar}^{(h)}\left[X_j - X_{\mathrm{pa}(j)}e^{-\alpha\ell_j} \mid \mathbf{Y}\right] \\ + \left(\mathbb{E}^{(h)}\left[X_1 \mid \mathbf{Y}\right] - \beta_0^{(h+1)}\right)^2 + \sum_{j=2}^{m+n} C_j^{OU}\left(\alpha^{(h)}, \boldsymbol{\Delta}^{(h+1)}\right) \end{aligned}$$

There is no such closed formula for α. In the implementation we propose, this parameter is actualized after all the others, by doing a numerical maximization of the objective function.

Finally, the vector $\mathbf{\Delta}^{(h+1)}$ of K non-zero shifts can be chosen in an optimal way for the BM thanks to a simple algorithm explained below. In the OUsun case, we can only increase the objective function, and not maximize it. In that case, we hence use a Generalized EM algorithm (GEM, see Dempster et al., 1977).

Optimal Shift Location for the BM. We want to minimize the sum of costs:

$$C^{BM}(\mathbf{\Delta}) = \sum_{j=2}^{m+n} C_j^{BM}(\mathbf{\Delta})$$

Each cost is associated to a branch b_j, $j \in [\![2\,, m+n]\!]$, and, when the sum is minimal, $C_j^{BM}(\mathbf{\Delta})$ can only take two values:

$$\begin{cases} \tilde{C}_j^{BM} = \ell_j^{-1}\left(\mathbb{E}^{(h)}\left[X_j \mid \mathbf{Y}\right] - \mathbb{E}^{(h)}\left[X_{\mathrm{pa}(j)} \mid \mathbf{Y}\right]\right)^2 & \text{if no shift on branch } b_j \\ 0 & \text{if one shift on branch } b_j \end{cases}$$

The sum can hence be minimized in the following way:

1. Compute $\tilde{C}_j^{BM}$ for all $j \in [\![2\,, m+n]\!]$.
2. Find the K highest costs $(j_1, \cdots, j_K) \in [\![2\,, m+n]\!]^K$.
3. Set $\Delta_{j_k}^{(h+1)} = \mathbb{E}^{(h)}\left[X_{j_k} \mid \mathbf{Y}\right] - \mathbb{E}^{(h)}\left[X_{\mathrm{pa}(j_k)} \mid \mathbf{Y}\right]$ for all $k \in [\![1\,, K]\!]$, and $\Delta_j^{(h+1)} = 0$ if $j \notin \{j_1, \cdots, j_K\}$.

This exact and fast algorithm works for the BM because all the costs are independent. Note that it would work for any Levy Process without memory, such as those proposed in Landis et al. (2013) to model evolution of quantitative traits.

GM Step for Shifts Locations for the OU. With $\alpha^{(h)}$ fixed, we want to minimize the sum of costs:

$$C^{OU}(\alpha^{(h)}, \mathbf{\Delta}) = \left(\mathbb{E}^{(h)}\left[X_1 \mid \mathbf{Y}\right] - \beta_0\right)^2 + \sum_{j=2}^{m+n} C_j^{OU}(\alpha^{(h)}, \mathbf{\Delta})$$

The previous algorithm does not work, because the costs are not independent. Solving the problem exactly would require to visit all the possible configurations, and the complexity would be too high, of order $O\left(\binom{m+n}{K}\right) = O(n^K)$. To reduce the execution time of the algorithm, we use heuristics to lower, if not minimize, the sum of costs. We use the following formulation:

$$C^{OU}(\alpha^{(h)}, \mathbf{\Delta}) = \left\|\mathbf{F}^{(h)} - \mathbf{A}^{(h)}\mathbf{U}\mathbf{\Delta}\right\|^2$$

where $\mathbf{U}$ the complete tree matrix given in subsection 2.3, $\mathbf{A}^{(h)} = \mathrm{Diag}\left(1, \sqrt{\frac{1-e^{-\alpha^{(h)}\ell_j}}{1+e^{-\alpha^{(h)}\ell_j}}}; 2 \leq j \leq m+n\right)$ a diagonal matrix depending on $\alpha^{(h)}$, and $\mathbf{F}^{(h)}$ a vector of expectations, with $F_1^{(h)} = \mathbb{E}^{(h)}\left[X_1 \mid \mathbf{Y}\right]$, and, for $2 \leq j \leq m+n$,

$$F_j^{(h)} = \left(1 - e^{-2\alpha^{(h)}\ell_j}\right)^{-1/2}\left(\mathbb{E}^{(h)}\left[X_j \mid \mathbf{Y}\right] - \mathbb{E}^{(h)}\left[X_{\mathrm{pa}(j)} \mid \mathbf{Y}\right] e^{-\alpha^{(h)}\ell_j}\right)$$

We can then use a Lasso algorithm to impose sparsity constraints on $\mathbf{\Delta}$. If $\mathbf{\Delta}_{-1}$ is the vector of shifts without the initial value (intercept), then a Lasso estimator is given by, for $\lambda \geq 0$:

$$\hat{\mathbf{\Delta}}_\lambda = \underset{\mathbf{\Delta}}{\mathrm{argmin}}\left\{\left\|\mathbf{F}^{(h)} - \mathbf{A}^{(h)}\mathbf{U}\mathbf{\Delta}\right\|^2 + \lambda\left|\mathbf{\Delta}_{-1}\right|_1\right\}$$

The estimated vectors $\hat{\mathbf{\Delta}}_\lambda$ have a support that is sparser when λ becomes higher. One then only need to find the right penalty factor λ that ensure that the support has exactly K non zero coordinates, plus the initial value. We ensure that the K shifts are allocated in a parsimonious way by checking their linear independence, using proposition 3.6.

An other method is to take the previous solution $\mathbf{\Delta}^{(h)}$, and test all the configurations where only one shift has moved, and take the best one. In both methods, one also has to ensure that the objective function is increased by the new choice of shifts, so that the GEM algorithm works correctly. This step is generally the longest one in one iteration of the EM.

C.4 Initialization

Initialization is always a crucial step when using an EM algorithm. The vector of shifts $\mathbf{\Delta}$ is initialized thanks to a Lasso procedure. To do that, we use the linear formulation 2.5 or 2.8 of the main text, and we calibrate the penalty so that the initialization vector has a non zero first coordinate (initial value), and K other non-zero coordinates. The variance-covariance matrix is initialized with defaults parameters, and is taken into account thanks to a Cholesky decomposition.

We also initialize the selection strength α. We use the following property: if Y_i and Y_j are two tips in the same group, then, under an OUsun, $\mathbb{E}\left[(Y_i - Y_j)^2\right] = 2\gamma^2(1 - e^{-\alpha d_{ij}})$. Using regression techniques, we can get an initial estimation of α and γ^2 from all these couples. In practice, we first initialize the position of the shifts, and then use only pairs of tips from the same estimated group. Then, as the groups are only approximated, some of the selected pairs (Y_i, Y_j) might not share the same expectation, and we use a robust regression to get more accurate initial estimates.

D Proof of Proposition 4.1 for Model Selection

We prove the proposition using the linear formulation $\mathbf{s} + \gamma\mathbf{E}$, with $\mathbf{E} \sim \mathcal{N}(0, \mathbf{V})$, as derived in the main text for the OUsun (with $\mathbf{s} = \mathbf{TW}(\alpha)\mathbf{\Delta}$, $\gamma^2 = \sigma^2/(2\alpha)$, and $V_{ij} = e^{-\alpha d_{ij}}$, see Formula (2.8)). Note that this framework also holds for the BM with a fixed root (with $\mathbf{s} = \mathbf{T\Delta}$, $\gamma = \sigma$, and $V_{ij} = t_{ij}$, see Formula (2.5)).

We first handle the case where there are no correlations ($\mathbf{V}$ diagonal), and then use a Cholesky decomposition to handle the general case. Note that the case $\mathbf{V}$ diagonal can be seen as the limit of the OUsun when $\alpha = +\infty$, or as a BM on a star tree.

Case $\mathbf{V}$ *diagonal*

In the iid case, we just need to check the conditions of theorem 4.1. This paragraph is highly inspired by the derivation of the bound for the detection of non-zero mean components exposed in Baraud et al. (2009) (sub-section 5.2). Assume that $D_\eta = K_\eta + 1 \leq p \leq n - 7$ for all $\eta \in \mathcal{M}$. The estimator is defined by $\hat{\mathbf{s}}_{\hat{K}}$, with:

$$\hat{K} = \underset{0 \leq K \leq p-1}{\operatorname{argmin}} \|\mathbf{Y} - \hat{\mathbf{s}}_K\|^2_{\mathbf{V}^{-1}} \left(1 + \frac{\operatorname{pen}_{A,\mathcal{L}}(K)}{n - K - 1}\right)$$

From the definition of $\hat{\mathbf{s}}_K$, and as the penalty depends on the model only through its number of shifts, we get that $\hat{\mathbf{s}}_{\hat{K}} = \hat{\mathbf{s}}_{\hat{\eta}}$ the minimizer of the criterion 4.1 of theorem 4.1 (with $N_\eta = n - D_\eta = n - K_\eta - 1$). We then have:

$$\Omega' = \sum_{\eta \in \mathcal{M}} (D_\eta + 1)e^{-L_\eta} = \sum_{K=0}^{p-1} \left|\mathcal{S}_K^{PI}\right| (K+2)e^{-L_K}$$

With the weights L_K defined in equation 4.4 of the proposition, we get:

$$\Omega' = \sum_{K=0}^{p-1} \frac{1}{K+2} \leq \log(p) \leq \log(n)$$

As:

$$\begin{aligned} L_K \leq \log \binom{n+m-1}{K} + 2\log(K+2) &\leq K\log(n+m-1) + 2\log(K+2) \\ &\leq K\log(2n-2) + 2(K+1) \\ &\leq (K+1)(2+\log(2)+\log(n)) \\ &\leq p(2+\log(2)+\log(n)) \end{aligned}$$

if $p \leq \min\left(\frac{\kappa n}{2+\log(2)+\log(n)}, n-7\right)$, then $\max(L_\eta, D_\eta) \leq \kappa n$ for any $\eta \in \mathcal{M}$, and we get the announced bound from the second proposition of theorem 4.1.

Case $\mathbf{V}$ *not diagonal*

Using a Cholesky decomposition, we can find a lower triangular matrix $\mathbf{L}$ such that $\mathbf{V} = \mathbf{L}\mathbf{L}^T$. Then, denoting $\mathbf{Y}' = \mathbf{L}^{-1}\mathbf{Y}$, $\mathbf{s}' = \mathbf{L}^{-1}\mathbf{s}$, and $\mathbf{E}' = \mathbf{L}^{-1}\mathbf{E}$, we have $\mathbf{Y}' = \mathbf{s}' + \gamma\mathbf{E}'$, with $\mathbf{E}' \sim \mathcal{N}(0, \mathbf{I_n})$, and we can apply theorem 4.1 as above. As we changed the metric, the estimators are projections on the linear spaces $S'_\eta = \mathbf{L}^{-1}S_\eta$ for $\eta \in \mathcal{M}$, and we have:

$$\begin{aligned} \hat{\mathbf{s}}'_\eta = \mathrm{Proj}_{S'_\eta}\mathbf{Y}' &= \underset{\mathbf{a}' \in S'_\eta}{\mathrm{argmin}} \left\|\mathbf{Y}' - \mathbf{a}'\right\|^2 = \underset{\mathbf{a}' \in S'_\eta}{\mathrm{argmin}} \left\|\mathbf{L}^{-1}\mathbf{Y} - \mathbf{L}^{-1}\mathbf{L}\mathbf{a}'\right\|^2 \\ &= \underset{\mathbf{a}' \in S'_\eta}{\mathrm{argmin}} \left\|\mathbf{Y} - \mathbf{L}\mathbf{a}'\right\|^2_{\mathbf{V}^{-1}} = \mathbf{L}^{-1} \underset{\mathbf{a} \in S_\eta}{\mathrm{argmin}} \left\|\mathbf{Y} - \mathbf{a}\right\|^2_{\mathbf{V}^{-1}} = \mathbf{L}^{-1}\hat{\mathbf{s}}_\eta \end{aligned}$$

So $\|\mathbf{s} - \hat{\mathbf{s}}_{\hat{\eta}}\|^2_{\mathbf{V}^{-1}} = \left\|\mathbf{s}' - \hat{\mathbf{s}}'_{\hat{\eta}}\right\|^2$ and $\|\mathbf{Y} - \hat{\mathbf{s}}_\eta\|^2_{\mathbf{V}^{-1}} = \left\|\mathbf{Y}' - \hat{\mathbf{s}}'_\eta\right\|^2$, and, as the form of the penalty does not depend on V, by minimizing:

$$\mathrm{Crit}_{LS}(K) = \left\|\mathbf{Y}' - \hat{\mathbf{s}}'_K\right\|^2 \left(1 + \frac{\mathrm{pen}_{A,\mathcal{L}}(K)}{n-K-1}\right) = \|\mathbf{Y} - \hat{\mathbf{s}}_K\|^2_{\mathbf{V}^{-1}} \left(1 + \frac{\mathrm{pen}_{A,\mathcal{L}}(K)}{n-K-1}\right)$$

we get the announced bound on $\mathbb{E}\left[\frac{\|\mathbf{s}-\hat{\mathbf{s}}_{\hat{K}}\|^2_{\mathbf{V}^{-1}}}{\gamma^2}\right] = \mathbb{E}\left[\frac{\left\|\mathbf{s}'-\hat{\mathbf{s}}'_{\hat{K}}\right\|^2}{\gamma^2}\right]$.

E Supplementary Figures

E.1 Simulation Study: Sensitivity and False Positive Rate

Definition of the Scores. We denote by TP the number of True Positives, i.e. the predicted edges on which a shift actually occurred, and FP the number of False Positives. The sensitivity $\frac{TP}{K_t}$ is the proportion of well predicted shifts among all shifts to be predicted, and the False Positive Rate (FPR) $\frac{FP}{n+m-K_t}$ is the proportion of false positive among all edges with no shifts.

Note that here, due to the possible lack of identifiability, the position of the shifts on the tree is not well defined, as a shift can be on a particular edge for one of the equivalent solutions, but not on the others (see Section 3.1). These two scores are hence ill defined for our problem. To avoid such problems, we restrict ourselves to the 91% of unambiguous configurations that occurred during the simulations.

Interpretation of Results. Figure 11 shows that the FPR are systematically worse when using the true number of shifts, indicating that the additional shifts found when compared to the selected number are misplaced. The FPR remains very low, as only a small number of shifts is to be found. Unsurprisingly, the Sensitivity is on the contrary improved when taking the real number of shifts, as shown Figure 12. In addition, the sensitivity is highly degraded when α is small or γ^2 is high, but does not exhibit a clear tendency in the real number of shifts, and the knowledge of the true value of α does not seem to matter.

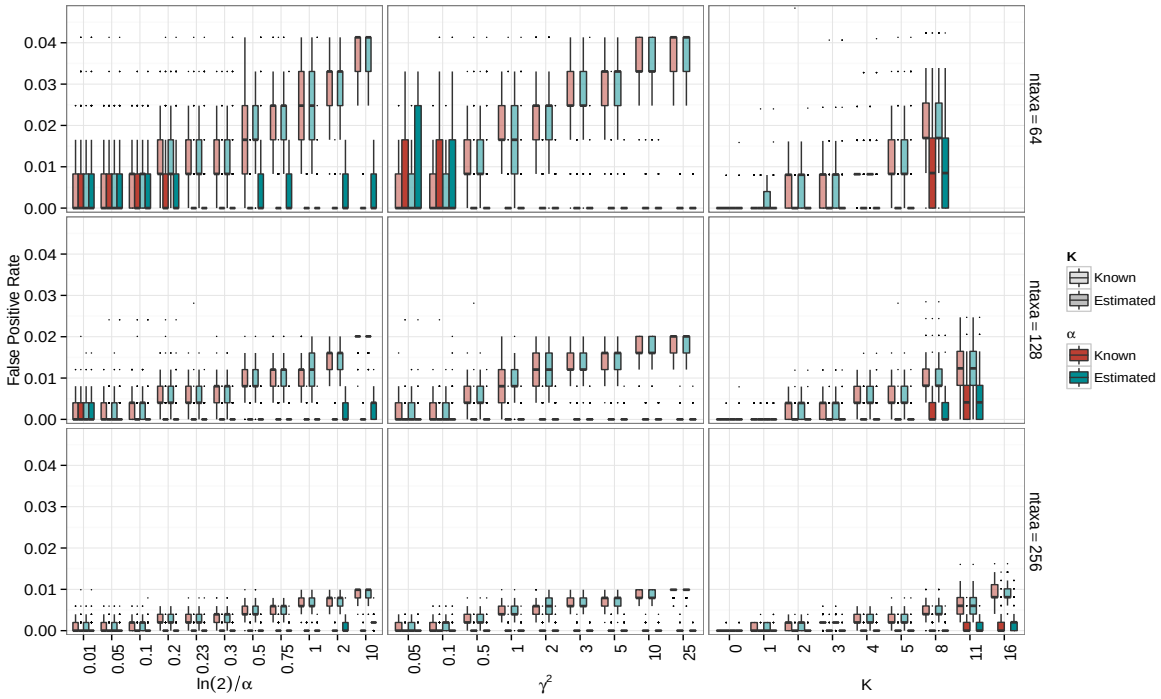


Figure 11: False Positive Rate computed for the different configurations. Note the y scale, that only goes to 0.05.

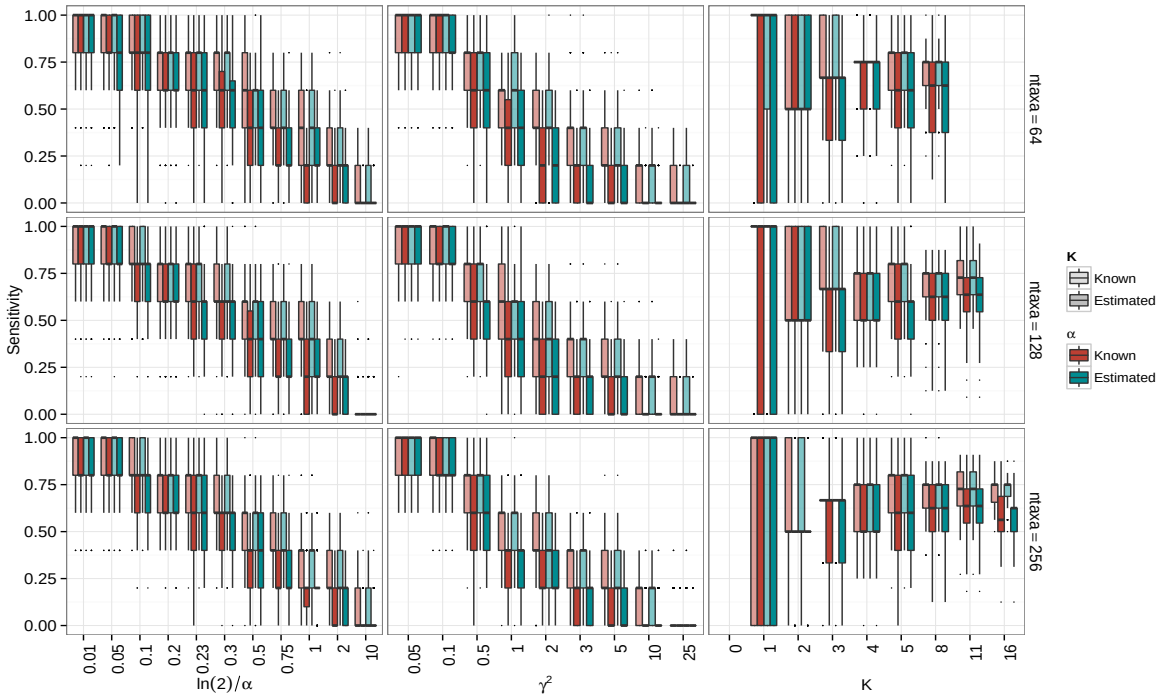


Figure 12: Sensitivity computed for the different configurations, with box-plots over the repetitions.

E.2 Simulation Study: Complementary Analysis

To complete the analysis conducted in the main text, Figure 13 presents the variations of the log-likelihood, phylogenetic half-life, and root variance when α is estimated, and the number of shifts is known or estimated. We can see here that the likelihood is slightly higher when the number of shifts is fixed, which is coherent with the behavior of our model selection procedure, that tends to under-estimate the true number of shifts (see Figure 7). We also note that knowing the true number of shifts has not a great influence on the estimation of α and γ, making the later worse, if anything.

Figure 14, shows the variations of the estimations of β_0, the number of shifts and the ARI when the number of shifts is estimated, and α is fixed or estimated. This confirms our earlier statement, that not knowing α with precision does not have a great impact on the model selection procedure (see also Cressler et al., 2015).

E.3 Chelonia Dataset: Comparison of Infered Shift Locations

On Figure 15, we present and compare the shift locations found by our method, and methods bayou and SURFACE. The differences found are explored deeper in the main text (Section 6.4).

Note on Computation Times. We found that the running time for our method was similar to the running time of previous algorithms (see Table 1 in the main text). However, our computations can be highly parallelized, as each run for a fixed number of shift is independent from the others. For instance, in the previous example, the computation time could be divided by 6, each estimation for a fixed α running on a different core. On the contrary, the SURFACE method cannot be parallelized at all, and only independent chains can be parallelized for Bayesian methods, so that the computation time can only be divided by 2 in our example.

F Practical Implementation

The statistical method described here was implemented on the statistical software R (R Core Team, 2014), and the code is freely available on GitHub (`https://github.com/pbastide/Phylogenetic-EM`). Phylogenetic trees were handled thanks to the package ape (Paradis et al., 2004). Packages TreeSim (Stadler, 2014), robustbase (Rousseeuw et al., 2014) and quadrupen (Grandvalet et al., 2012) were used, respectively, for random tree generation, robust regression and Lasso regression. The penalty described in proposition 4.1 is implemented in package LINselect (Baraud et al., 2013).

Parallelization was achieved thanks to R packages foreach (Weston, 2014b) and doParallel (Weston, 2014a).

Package mclust (Fraley et al., 2012) was used for ARI computations. Plots were made thanks to packages ggplot2 (Wickham, 2009) and reshape2 (Wickham, 2007).

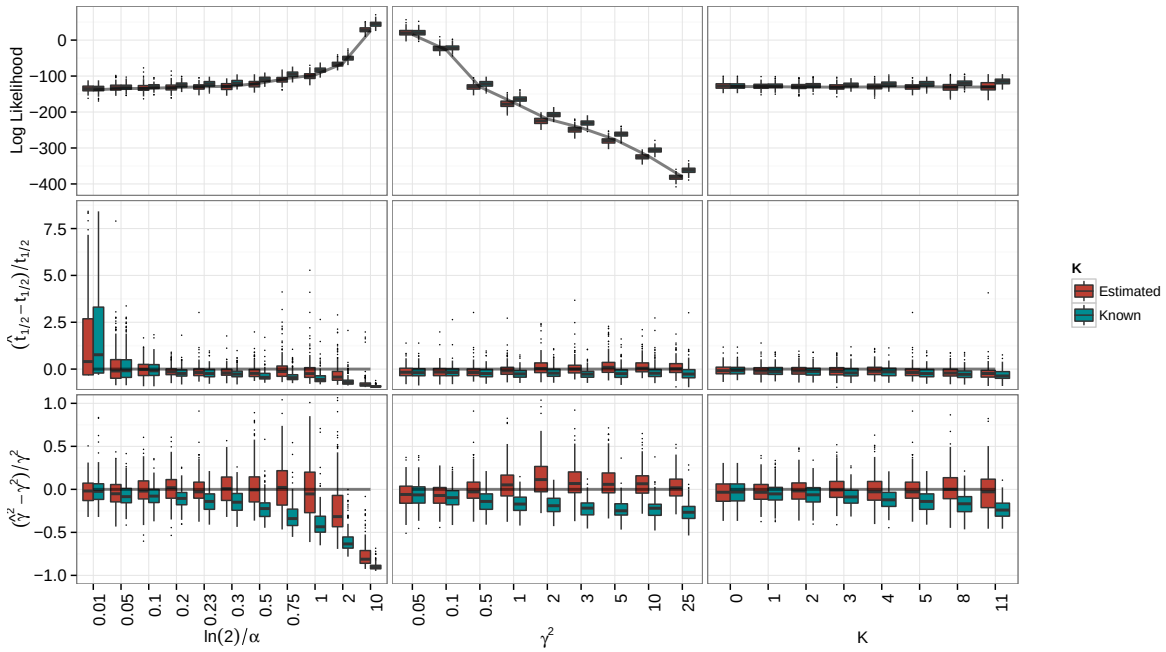


Figure 13: Box plots over the 200 repetitions of each set of parameters, for the log-likelihood (top), phylogenetic half-life (middle) and root variance (bottom) with α estimated, and K fixed or estimated, on a tree with 128 taxa.

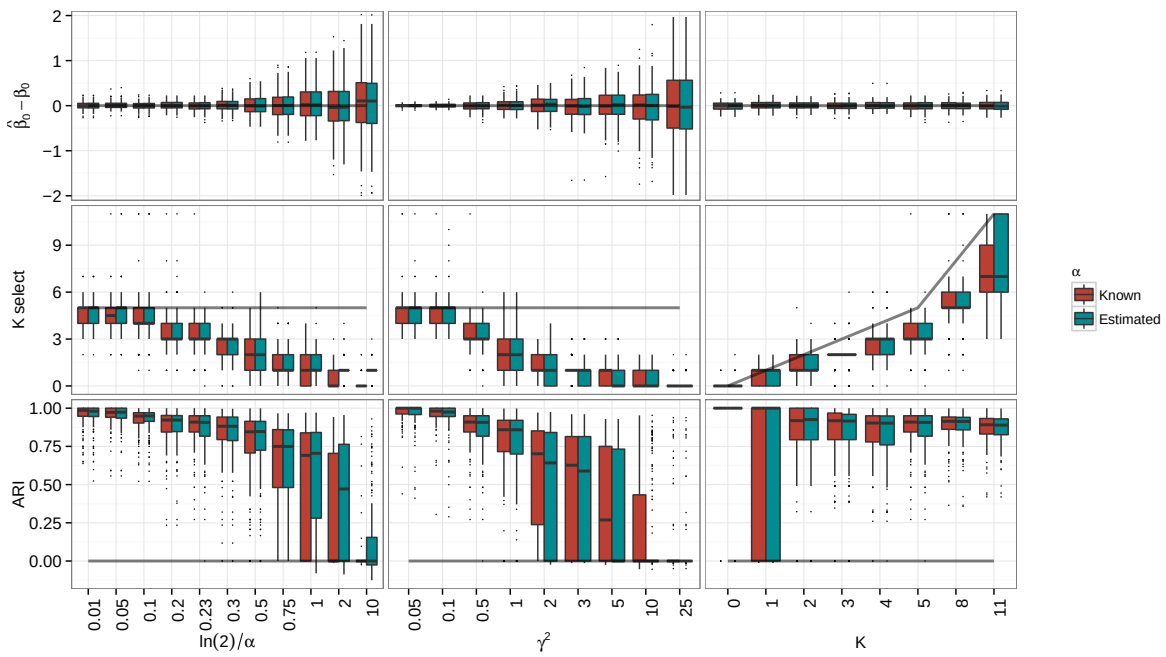


Figure 14: Same for β_0 (top), the number of shifts (middle) and ARI (bottom), with K estimated, and α fixed or estimated.

For better legibility, strips with $t_{1/2}$, γ^2 and β_0 on these two figures were re-scaled, omitting some outliers (respectively, 0.21%, 0.27% and 0.27% of points are omitted). The whisker of the first box for $t_{1/2}$ goes up to 7.5.

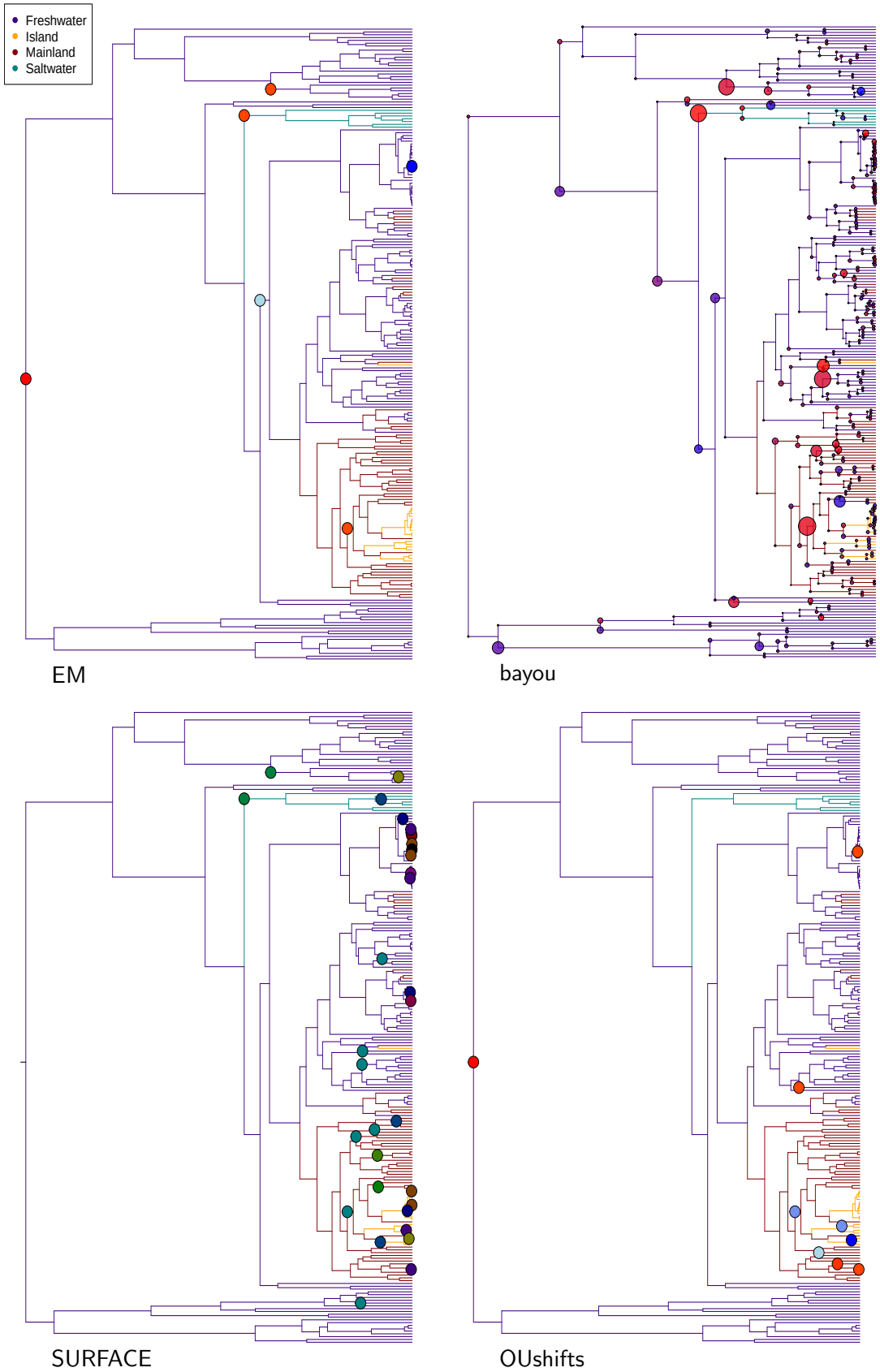


Figure 15: Solutions found by the EM (top left), bayou (top right), SURFACE (bottom left) and OUshifts (bottom right). The branch coloring represents the habitats. For the EM, bayou and OUshifts, the shifts coloring represents their values, from blue (negative) to red (positive). For bayou, the size of the circles are proportional to their posterior probability. For SURFACE, the 13 colors of the shifts represent the regimes.