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

Paul Bastide,
AgroParisTech, Paris, Institut National de la Recherche Agronomique, Jouy-en-Josas and Paris, and Université Paris-Saclay, France

Mahendra Mariadassou
Institut National de la Recherche Agronomique, Jouy-en-Josas and Paris, and Université Paris-Saclay, France

and Stéphane Robin
AgroParisTech, Paris, Institut National de la Recherche Agronomique, Jouy-en-Josas and Paris, and Université Paris-Saclay, France

[Received August 2015. Final revision July 2016]

Summary. Comparative and evolutive ecologists are interested in the distribution of quantitative traits between 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 expectation–maximization 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: Adaptive shifts; Change point detection; Model selection; Ornstein–Uhlenbeck process; Phylogeny; Random process on tree

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 the past (see for instance Losos (1990), Mahler et al. (2013), Davis et al. (2007) and 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. To explain the pattern of traits measured on a set of related species, we...
P. Bastide, M. Mariadassou and S. Robin

need to take these correlations into account. Indeed, a given species will be more likely to have a similar trait value to that of her ‘sister’ (a closely related species) than to that of her ‘cousin’ (a distantly related species), just because of the structure of the tree. On top of that structure, when considering a functional trait (i.e. a trait that is directly linked to the fitness of its bearer), such as shell size for turtles (Jaffe et al., 2011), we need 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 Galápagos Islands for turtles, will affect the observed trait distribution, here with a shift towards giant shell sizes compared with 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 detect the past environmental shifts that shaped the present day trait distribution automatically.

1.2. Stochastic process on a tree

We model the evolution of a quantitative adaptive trait by 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.

1.2.1. Tree structure

The tree structure model is our null model: it accounts for the tree-induced distribution of trait values in the absence of shifts. Depending on the phenomenon being studied, several stochastic processes can be used to capture the dynamics of the trait evolution. In what follows, we shall use Brownian motion (BM) and Ornstein–Uhlenbeck (OU) processes.

1.2.2. Brownian motion

Since the seminal article of Felsenstein (1985), BM has been used as a neutral model of trait evolution. If \((B_t; t \geq 0)\) is BM, a character \((W_t; t \geq 0)\) evolves on a lineage according to the stochastic differential equation \(dW_t = \sigma dB_t\), \(\sigma^2\) being a variance parameter. If \(\mu\) is the ancestral value at the root of the tree \((t = 0)\), then \(W_t \sim N(\mu, \sigma^2 t)\). 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.

1.2.3. Ornstein–Uhlenbeck process

An unbounded variance is quite unrealistic for adaptive traits (Butler and King, 2004). For that reason, the OU process, which models stabilizing selection around an adaptive optimum (Hansen, 1997), is usually preferred to the BM process. It is defined by the stochastic differential equation \(dW_t = -\alpha(W_t - \beta)dt + \sigma dB_t\) and has stationary distribution \(N(\beta, \sigma^2/2\alpha)\). In this equation, \(W_t\) is the secondary optimum of a species, which is 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 \(\beta\), which depends only on the ecological niche of the species. The selection strength is controlled by the call-back parameter \(\alpha\). For interpretation, we shall use the phylogenetic half-life \(t_{1/2} = \ln(2)/\alpha\), which is defined as the time that is needed for the expected trait value to move half the distance from the ancestral state to the primary optimum (Hansen,
Figure 1. Trait evolution under BM (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 (D,E), changing the trajectory of their ancestral trait value from the dotted one to the full one; the shift increases the observed dispersion): (a) a phylogenetic tree; (b) trait value evolution

1.2.4. 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 parameter trajectories. In the BM process, shifts affect the mean value of the trait and are thus instantaneously passed on to the trait itself (Fig. 1) whereas, in the OU process, shifts affect the primary optimum $\beta$. 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 more difficult to detect (Hansen and Bartoszek, 2012). In the remainder of the paper, we assume that all other parameters ($\sigma^2$ for the BM and $\sigma^2$ and $\alpha$ for the OU process) are fixed and constant (but see Beaulieu et al. (2012) and Rabosky (2014) for partial relaxations of this hypothesis).

1.3. Scope of this paper
1.3.1. State of the art
Phylogenetics comparative methods 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) considered only shifts in the optimal value $\beta$ whereas Beaulieu et al. (2012) also allowed for shifts in the selection strength $\alpha$ and the variance $\sigma^2$. 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 on the BM process, 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. (2016), who adapted Felsenstein’s pruning algorithm for the likelihood computation of these models, using efficient integration techniques. Finally, Ho and Ané (2013) derived consistency results for estimation of the parameters of an OU process on a tree and Bartoszek and Sagitov (2015) and Sagitov and Bartoszek (2012) computed confidence intervals of the same parameters by assuming an unknown random-tree topology and averaging over it.

The first steps towards automatic detection of shifts, which is the problem that is of interest in this paper, have been done in a Bayesian framework, for both the BM (Eastman et al., 2013) and the OU processes (Uyeda and Harmon, 2014). Using reversible jump Markov chain Monte Carlo sampling, 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 phylogenetics comparative methods in fields where large trees are the norm makes maximum-likelihood-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 a time and is therefore quite efficient but the selection criterion 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), who described several identifiability problems that arise when trying to infer the position of the shifts on a tree and proposed a different stepwise algorithm based on a more stringent selection criterion, heuristically inspired by segmentation algorithms.

To tackle this issue rigorously, we introduce a framework where a univariate trait evolves according to an 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 with speciation events and occur only at nodes ($N$) of the tree. We call this model OUsun hereafter.

1.3.2. 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 maximize the likelihood efficiently. Second, we show that, given the model that is used and the kind of data that are available, some evolutionary scenarios remain indistinguishable. Formally, we exhibit some identifiability 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 that are needed to explain the data best. Thanks to our knowledge of the structure of the spaces of models, acquired through our identifiability study, we can mathematically derive a penalization term, together with an oracle inequality on the estimator that is found. Fourth and finally, we implement the method with the statistical software R (R Core Team, 2014) and show that it correctly recovers the structure of the model on simulated data sets. When applied to a biological example, it gives results that are easily interpretable and coherent with previously developed methods. All the code that is used in this paper is publicly available on GitHub (https://github.com/pbastide/PhylogeneticEM).

1.3.3. Outline
In Section 2, we present the model, using two different mathematical points of view that are both useful in different aspects of the inference. In Section 3, we tackle the identifiability problems that are associated with this model, and we 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 data sets.

2. Statistical modelling

2.1. Probabilistic model

2.1.1. Tree parameterization
As shown in Fig. 2, we consider a rooted tree $T$ 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 $\text{pa}(i)$ is the unique parent of node $i$. The branch leading to $i$ from $\text{pa}(i)$ is denoted $b_i$ and has length $l_i = t_i - t_{\text{pa}(i)}$ where $t_i$ is the time elapsed between the root and node $i$. By convention, we set $t_1 = 0$ and $t_{\text{pa}(1)} = -\infty$ for the root. The last convention ensures that the trait follows the stationary distribution (if any) of the process at the root. We denote $\text{Anc}(i) = \{\text{pa}'(i): r \geq 0\}$ the set composed of node $i$ and of 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 is $t_{ij}$. We consider ultrametric trees, for which $t_{m+1} = \ldots = t_{m+n} = h$ and denote $h$ the tree height. In what follows, the tree is fixed and assumed to be known.

**Fig. 2.** A rooted and time-calibrated phylogenetic tree with the notation used to parameterize the tree $(l, t, d, b)$ and the observed, $Y$, and non-observed, $Z$, variables
2.1.2. Trait values

We denote by $X$ the vector of size $m+n$ of the trait values at the nodes of the tree. We split this vector into non-observed values $Z$ (size $m$) at the internal nodes and observed values $Y$ (size $n$) at the tips, so that $X^T = (Z^T, Y^T)$. 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 what follows, we assume that the inference in the BM case is done conditionally on 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: $X_1 \sim N\{\mu = \beta_1, \gamma^2 = \sigma^2/2\alpha\}$, where $\beta_1$ is the ancestral optimal value.

2.1.3. Shifts

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

$$\Delta_1 = \mu \quad (= \beta_1 \text{ for OUsun}) \quad \text{and} \quad \forall i \in \{2, m+n\}, \Delta_i = \begin{cases} \delta_k & \text{if } \tau_k = b_i, \\ 0 & \text{otherwise.} \end{cases}$$

Note that no proper shift occurs on the root branch, but that the root trait value or mean $\mu$ is formalized as an initial notional shift on this notional branch.

2.1.4. Parameters

The parameters that are needed to describe OUsun or BM are respectively $\theta = (\gamma, \alpha, \Delta)$ or $\theta = (\sigma, \Delta)$. As $\sigma^2 = 2\alpha\gamma^2$, only the two parameters $\alpha$ and $\gamma$ are needed to describe the OUsun model. We denote respectively by OUsun($\theta$) and BM($\theta$) the OUsun and BM process running on the tree with parameters $\theta$.

2.2. Incomplete-data model point of view

If the trait values were observed at all nodes of the tree, including ancestral nodes, shifts would be characterized by unexpectedly large differences between a node and its parent. A way to mimic this favourable 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).

2.2.1. Brownian motion

As the shifts occur directly in the mean of the process, we obtain

$$X_1 = \mu \quad \text{and} \quad \forall i \in \{2, m+n\}, X_i | X_{pa(i)} \sim N(X_{pa(i)} + \Delta_i, l_i \sigma^2). \quad (2.1)$$

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

2.2.2. Ornstein–Uhlenbeck process

The shifts occur on the primary optimum $\beta$, which is piecewise constant. As the shifts are assumed to occur at nodes, the primary optimum is entirely defined by its initial value $\beta_1$ and
its values $\beta_2, \ldots, \beta_{n+m}$ on branches of the tree, where $\beta_i$ is the value on branch $b_i$ leading to node $i$:

$$\beta_i \in \mathbb{R} \ (= \mu \text{ for OUsun}) \quad \text{and} \quad \forall i \in [2, m+n], \beta_i = \beta_{\text{pa}(i)} + \Delta_i. \quad (2.2)$$

Assuming that the root node is in the stationary state, we obtain

$$X_1 \sim \mathcal{N}\{\mu = \beta_1, \gamma^2 = \sigma^2/(2\alpha)\},$$

$$X_i|X_{\text{pa}(i)} \sim \mathcal{N}\{X_{\text{pa}(i)} \exp(-\alpha l_i) + \beta_i \{1 - \exp(-\alpha l_i)\}, \sigma^2/(2\alpha)\} \{1 - \exp(-2\alpha l_i)\} \quad (2.3)$$

$$\forall i \in [2, m+n].$$

The trait value at node $i$ depends on both the trait value at the father node $X_{\text{pa}(i)}$ and the value $\beta_i$ of the primary optimum on branch $b_i$. Contrary to the BM case, the shifts appear only indirectly in the distributions of $X_i$s, through the values of $\beta$, and with a shrinkage of $1 - \exp(-\alpha d)$ for shifts of age $d$, which makes recent shifts ($d$ small compared with $1/\alpha$) more difficult to detect.

### 2.3. Linear regression model point of view

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

#### 2.3.1. Matrix of a tree

It follows from the recursive definition of $X$ that it is a Gaussian vector. To express its mean vector given the shifts, we introduce the tree squared matrix $U$, of size $(m+n)$, defined by its general term: $U_{ij} = \mathbb{1}\{j \in \text{Anc}(i)\}, \forall (i, j) \in [1, m+n]^2$. In other words, the $j$th column 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 $Y$, we also need the submatrix $T$, of size $n \times (m+n)$, composed of the bottom $n$ rows of matrix $U$, corresponding to the tips (see example 1 below). Likewise, the $i$th row of $T$, $i \in [1, n]$, is the indicator vector of the ancestors of leaf $m+i$.

#### 2.3.2. Brownian motion

From the tree structure, we obtain

$$X = U\Delta + E_X,$$

$$Y = T\Delta + E_Y. \quad (2.4)$$

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

#### 2.3.3. Ornstein–Uhlenbeck process

For OUsun, shifts occur on the primary optimum, and there is a lag term, so

$$\beta = U\Delta,$$

$$X = (U - AUB)\Delta + E_X \quad (2.5)$$

where $A = \text{diag}\{\exp(-\alpha l_i), 1 \leq i \leq m+n\}$ and $B = \text{diag}\{0, \exp(\alpha t_{\text{pa}(i)}), 2 \leq i \leq m+n\}$ are diagonal matrices of size $m+n$. As previously, $E_X \sim \mathcal{N}(0, \Sigma_{XX})$, but $\Sigma_{XX} = \gamma^2[\exp(-\alpha d_{ij})]_{1 \leq i, j \leq m+n}$. As the tree is ultrametric, this expression simplifies to the following equation when considering only observed values:
\[ Y = \mathbf{TW}(\alpha) \Delta + \mathbf{E}_Y \tag{2.6} \]

where \( \mathbf{E}_Y \) is the Gaussian vector made of the last \( n \) co-ordinates of \( \mathbf{E}_X \), and \( \mathbf{W}(\alpha) = \text{diag}[1, 1 - \exp\{-\alpha(h - t_{\text{pa}(i)})\}, 2 \leq i \leq m + n] \) is a diagonal matrix of size \( m + n \). If \( \alpha \) is positive, then \( \alpha(h - t_{\text{pa}(i)}) > 0 \) for any \( i \in \llbracket 1, m + n \rrbracket \), and \( \mathbf{W}(\alpha) \) is invertible.

2.3.3.1. Example 1. The tree that is presented in Fig. 2 has five tips and one shift on branch 4 + 3 = 7, so

\[
\begin{pmatrix}
Z_1 & Z_2 & Z_3 & Z_4 & Y_1 & Y_2 & Y_3 & Y_4 & Y_5 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{pmatrix}
\]

and

\[
\Delta = \begin{pmatrix}
\mu \\
0 \\
0 \\
0 \\
0 \\
\delta_1 \\
0 \\
0 \\
0 \\
\end{pmatrix}.
\]

And, respectively for BM or OUsun (\( \mu = \beta_1 \)),

\[
\mathbb{E}[Y] = T \Delta = (\mu, \mu, \mu + \delta_1, \mu, \mu)^T,
\]

\[
\mathbb{E}[Y] = \mathbf{TW}(\alpha) \Delta = (\beta_1, \beta_1, \beta_1 + \delta_1[1 - \exp \{-\alpha(h - t_2)\}], \beta_1, \beta_1)^T.
\]

2.3.4. Space of expectations

Expressions (2.4) and (2.6) allow us to link the parameter \( \theta \) to the probability distribution of observations \( Y \) and to explore identifiability issues. In this linear formulation, detecting shifts boils down to identifying the non-zero components of \( \Delta \). The following lemma highlights the parallels between solutions of the BM and OUsun processes.

**Lemma 1** (similar solutions). Let \( \mathbf{m}_Y \in \mathbb{R}^n \) be a vector, \( T \) an ultrametric tree, \( \alpha \) a positive real number and \( \gamma \) non-negative real numbers. Then there is at least one vector \( \Delta_{\text{BM}}, \Delta_{\text{BM}} \in \mathbb{R}^{m+n} \) and \( \Delta_{\text{OU}} \in \mathbb{R}^{m+n} \), such that the vector of expectations at the tips of a BM(\( \sigma, \Delta_{\text{BM}} \)) and an OUsun(\( \gamma, \alpha, \Delta_{\text{OU}} \)) process running on the tree \( T \) is exactly \( \mathbf{m}_Y \).

Furthermore, \( \Delta_{\text{BM}} \) is a solution to this problem for the BM if and only if

\[
\Delta_{\text{OU}} = \mathbf{W}(\alpha)^{-1} \Delta_{\text{BM}}
\]

is a solution for OUsun, and \( \Delta_{\text{BM}} \) and \( \mathbf{W}(\alpha)^{-1} \Delta_{\text{BM}} \) have the same support. These two vectors are said to be similar.

**Proof.** The first part of lemma 1 follows directly from formulae (2.4) (BM) and (2.6) (OU). Indeed, the maps \( \Delta \mapsto T \Delta \) and \( \Delta \mapsto \mathbf{TW}(\alpha) \Delta \) both span \( \mathbb{R}^n \). The second part of lemma 1 is a consequence of \( \mathbf{W}(\alpha) \) being diagonal and invertible (for \( \alpha > 0 \)). \( \square \)

**Remark 1.** Lemma 1 shows that the OUsun and BM processes that induce a given \( \mathbf{m}_Y \) use shifts on the same branches, although they may differ on other parameters.
3. Identifiability and complexity of a model

3.1. Identifiability issues

As we have access to $Y$ only, and not $X$, we have only 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 are parameters $\theta \neq \theta'$ with the same likelihood function: $p_\theta(\cdot) = p_{\theta'}(\cdot)$. 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 can enumerate, first, all the equivalent solutions to a given problem, and, second, all the equivalence classes for a given number of shifts.

3.1.1. No homoplasy assumption

We assume in what follows 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 it. This assumption is reasonable considering that shifts are real valued and makes the model similar to ‘infinite alleles’ models in population genetics. This assumption confines but does not eliminate the identifiability issue, as seen in Fig. 3.

**Proposition 1** (kernel of the tree matrix $T$). Let $i$ be an internal node, $i \in \{1, m\}$, with $L_i$ children nodes $(d_1, \ldots, d_{L_i}) \in \{2, m + n\}^{L_i}$. Then the vector $K^i$ defined as

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

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

These kernel vectors effectively ‘cancel out’ a shift on a branch by balancing it with the opposite shift on all immediate child branches. Note that the root mean value is treated as a shift. The following lemma describes the relationships that exist between these kernel vectors and the tree matrix $U$ that was defined in Section 2.3. The technical proofs of proposition 1 and lemma 2 are postponed to the on-line appendix A.

**Lemma 2.** Let $b$ be the canonical basis of $\mathbb{R}^{m+n}$, and $S$ a supplementary space of $\ker(T)$. Then $b' = (K^1, \ldots, K^m, b_{m+1}, \ldots, b_{m+n})$ is a basis adapted to the decomposition $\ker(T) \oplus S$, where

$$
\ker(T) = \{ \sum_{i=1}^{m} a_i K^i : a_i \in \mathbb{R} \}.
$$
and the matrix $U$ (as defined in Section 2.3) is the change of basis matrix between $b$ and $b'$. As a consequence, $U$ is invertible.

3.1.1.2. ‘Random-cluster model’ representation. When inferring the shifts, we must keep in mind this problem of non-identifiability and be able to choose, if necessary, one or several possible allocations among all the equivalent allocations. 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 colouring problem.

Definition 1 (node colouring). Let $C_K$ be a set of $K$ arbitrary ‘colours’, $K \in \mathbb{N}^*$. For a given shift allocation, the colour of each node is given by the application $B : [1, m+n] \rightarrow C_K$ recursively defined in the following way.

(a) Choose a colour $c \in C_K$ for the root: $B(1) = c$.

(b) For a node $i, i \in [2, m+n]$, set $B(i)$ to $B\{\text{pa}(i)\}$ if there is no shift on branch $i$; otherwise choose another colour $c, c \in C_K \setminus \{B\{\text{pa}(i)\}\}$, and set $B(i)$ to $c$.

Hereafter, we identify $(C_K)^{[1,m+n]}$ with $(C_K)^{m+n}$ and refer to a node colouring indifferently as an application or a vector.

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

Definition 2 (adapted node colouring). A node colouring is said to be adapted to a shifted random process on a tree if two tips have the same colour if and only if they have the same mean value under that process.

Proposition 2 (adapted colouring for BM and OUsun). Let $\sigma$ and $\gamma$ be two non-negative real numbers, and $\alpha$ a positive real number.

(a) In the BM case, if $C$ is the set of possible mean values taken by the nodes of the tree, then the knowledge of the node colours is equivalent to knowledge of $\Delta$. Furthermore, the associated node colouring is adapted to the original BM.

(b) In the OUsun case, from lemma 1, we can find a similar BM process, i.e. with shifts on the same branches. Then knowledge of the node colouring associated with this similar BM process is equivalent to knowledge of the vector of shifts of OUsun, and the node colouring that is obtained is adapted to the original OUsun.

Proof. The proof of part (a) relies on expression (2.4), which states that $\mathbb{E}[X] = U \Delta$. Defining $C$ as the set of all distinct values of $\mathbb{E}[X]$, we can identify $\mathbb{E}[X]$ with the node colouring application that maps any node $i$ with $\mathbb{E}[X]$. Since $U$ is invertible (see lemma 2 above), we can go from one formalism to the other.

For part (b), we use lemma 1 to find a similar BM, and then use part (a).

From now on, we shall study the problem of shifts allocation as a discrete state colouring problem.

3.1.2. Parsimony

As we saw in Fig. 3 there are multiple colourings of the internal nodes that lead to a given tips colouring. Among all these solutions, we choose to study only the parsimonious solutions. This property can be seen as an optimality condition, as defined below:

Definition 3 (parsimonious allocation). Given a vector of mean values at the tips that are produced by a given shifted stochastic process running on the tree, an adapted node colouring
is said to be *parsimonious* if it has a minimum number of colour changes. We denote by $S_P^K$ 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 $K + 1$ colours, we can define an application $\phi: S_P^K \rightarrow (C_K)^n$ that maps a parsimonious allocation of shifts to its associated tip partition.

**Definition 4** (equivalence). Two allocations are said to be equivalent (denoted by ‘\sim’) if they produce the same partition of the tips and are both parsimonious. Mathematically:

$$\forall s_1, s_2 \in S_P^K, s_1 \sim s_2 \iff \phi(s_1) = \phi(s_2).$$

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

Several dynamic programming algorithms already exist to compute the minimal number of shifts required to produce a given tips colouring, and to find one associated parsimonious solution (see Fitch (1971), Sankoff (1975) and Felsenstein (2004)). Here, we need to be a little more precise, as we want both to count and to enumerate all possible equivalent node colourings that are associated with a tip colouring. For brevity, we present only the algorithm that counts $|\phi^{-1}(d)|$, for $d \in (C_K)^n$. This algorithm can be seen as a corollary of the enumeration algorithm (which is presented and proved in the on-line appendix B) and an extension of Fitch’s algorithm where we keep track of both the cost of an optimal colouring and the number of such colourings. It has $O(K^2 L n)$ time complexity where $L$ is the maximal number of children of the nodes of the tree.

**Proposition 3** (size of an equivalence class). Let $d$ be a colouring of the tips, $d \in (C_K)^n$, and let $i$ be a node of tree $T$ with $L_i$ daughter nodes $(i_1, \ldots, i_{L_i})$, $L_i \geq 2$. Denote by $T_i$ the subtree rooted at node $i$.

For $k \in C_K$, $S_i(k)$ is the cost of starting from node $i$ with colour $k$, i.e. the minimal number of shifts needed to obtain the colouring of the tips of $T_i$ defined by $d$, when starting with node $i$ in colour $k$. Denote by $T_i(k)$ the number of allocations on $T_i$ that achieve cost $S_i(k)$.

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

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

$$T_i(k) = \begin{cases} 1 & \text{if } d(i) = k, \\ 0 & \text{otherwise}. \end{cases}$$

Otherwise, if $i$ is a node, for $1 \leq l \leq L_i$, define the set of admissible colours for daughter $i_l$:

$$K^l_k = \arg \min_{p \in C_K} \{ S_{i_l}(p) + \mathbb{1}\{ p \neq k \} \}.$$

As these sets are not empty, let $(p_1, \ldots, p_L) \in K^1_k \times \cdots \times K^L_k$. Then

$$S_i(k) = \sum_{l=1}^L S_{i_l}(p_l) + \mathbb{1}\{ p_l \neq k \},$$

$$T_i(k) = \prod_{l=1}^L \sum_{p_l \in K^l_k} T_{i_l}(p_l).$$
At the root, if $\mathcal{L} = \arg \min_{k \in C_K} S_1(k)$, then $|\phi^{-1}(d)| = \Sigma_{k \in \mathcal{L}} T_1(k)$.

3.1.2.1. Ornstein–Uhlenbeck practical case. We can illustrate this notion on a simple example. We consider 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 2 and our enumeration algorithm, we can reconstruct the five possible allocations of shifts that produce the exact same distribution at the tips. These solutions are shown in Fig. 4. Note that the colours are not defined by the values of the optimal regime $\beta$, but by the mean values $\mathbb{E}[Y]$ of the process at the tips. As a result, the groups that are shown in light grey and dark grey 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 interdependent, 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 effect as an old shift, a recent shift must have a much higher intensity (under constant selection strength such as in OUsun).

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

3.2. Complexity of a collection of models

3.2.1. Number of different tips colours

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 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 colours. The tip partition into $K+1$ groups that is associated with this colouring is said to be tree compatible. The set $\mathcal{D}_{K+1} \subset (C_{K+1})^n$ of such partitions is the image of $\mathcal{S}_K^P$ by the map $\phi$ that was defined in the previous section.

Proof. First, note that $K$ shifts create at most $K+1$ colours. If each shift produces a new tip mean value (no homoplasy), the only way to create $K$ or fewer colours 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 obtain the same colouring of the tips with one fewer shift. So a parsi-
monious allocation cannot create fewer than $K + 1$ colours, and hence creates exactly $K + 1$
colours.

Reciprocally, if an allocation with $K$ shifts that produces $p$ groups is not parsimonious, then
we can find another parsimonious allocation 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 relationship that is defined in definition 4, we can formally take the
quotient set of $S^p_K$ by the relationship ‘∼’ to obtain the set of parsimonious allocations of $K$
shifts on the $m + n − 1$ branches of the tree that are identifiable: $S^p_K = S^p_K / ∼$. In other words, the
set $S^p_K$ 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: $S^p_K → D_{K+1}$.

The number $N^p_T / K$ is the complexity of the class of models with $K$ shifts defined as
the number of distinct identifiable parsimonious possible configurations one can obtain with $K$
shifts on the tree. To compute $N^p_T / K$, we shall need $M^p_T$, the number of marked tree compatible
partitions of the tips in $K + 1$ groups. These are composed of all the tree compatible partitions where one
group, among those that could be in the same state as the root, is distinguished with a mark
(see the following example).

3.2.1.1. Example 2 (difference between $N^p_T$ and $M^p_T$).

(a) If we consider only unmarked partitions, then colourings 1, 2 and 3 (Fig. 5) induce the
same partitions as respectively colourings 4, 5 and 6, and $N^p_T = 3$.

(b) For marked partitions, fix the root state to an arbitrary colour, for instance white, and
consider the white group as marked. Then colourings 5 and 6 are not tree compatible
(their two shifts). And, although they induce the same partition, colourings 1 and 4 correspond to different marked partitions: each marks a different group of leaves.
Therefore $M^p_T = 4$.

Proposition 5 (computation of the number of equivalent classes). Let $i$ be a node of tree $T$, and $K ∈ \mathbb{N}^*$.

If $i$ is a tip, then $N^p_T = M^p_T = \emptyset \{ K = 1 \}$. Otherwise, if $i$ is a node with $L_i$ daughter nodes
$(i_1, \ldots, i_{Li})$, $L_i ≥ 2$, then
\[ N_K^{(T)} = \sum_{I \subseteq \{1, \ldots, L\}} \sum_{k_1, \ldots, k_L \geq 1} \prod_{i \notin I} M_{k_i}^{(T_i)} \prod_{i \in I} N_{k_i}^{(T_i)} + \sum_{k_1, \ldots, k_L \geq 1} \prod_{i \notin I} M_{k_i} \prod_{i \in I} N_{k_i}, \]

(3.1)

\[ M_K^{(T)} = \sum_{I \subseteq \{1, \ldots, L\}} \sum_{k_1, \ldots, k_L \geq 1} \prod_{i \notin I} M_{k_i}^{(T_i)} \prod_{i \in I} N_{k_i}^{(T_i)}. \]

In the binary case, this relationship becomes, if \( i \) has two daughters \( i_1 \) and \( i_r \),

\[ N_K^{(T_i)} = \sum_{k_1+k_2=K} \sum_{k_1, k_2 \geq 1} N_{k_1}^{(T_{i1})} N_{k_2}^{(T_{i_r})} + \sum_{k_1+k_2=K+1} M_{k_1}^{(T_{i1})} M_{k_2}^{(T_{i_r})}, \]

(3.2)

\[ M_K^{(T_i)} = \sum_{k_1+k_2=K} \sum_{k_1, k_2 \geq 1} M_{k_1}^{(T_{i1})} M_{k_2}^{(T_{i_r})} + \sum_{k_1+k_2=K+1} M_{k_1}^{(T_{i1})} M_{k_2}^{(T_{i_r})}. \]

\textbf{Proof.} We shall prove this proposition in the binary case, the general case being a natural extension of it. If \( T \) is a binary tree with \( T_l \) and \( T_r \) as left and right subtrees, we face two situations when partitioning the tips in \( K \) groups.

(a) The left and right subtrees have no group in common. Then, the number of groups in \( T \) is equal to the number of groups in its two subtrees, and there are \( \sum_{k_1+k_2=K} N_{k_1}^{(T_{i1})} N_{k_2}^{(T_{i_r})} \) such partitions. This is the first term of the equation on \( N_K^{(T)} \) in expression (3.2).

(b) The left and right subtrees 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 subtrees, and there are \( \sum_{k_1+k_2=K+1} M_{k_1}^{(T_{i1})} M_{k_2}^{(T_{i_r})} \) such partitions. This ends the proof of the formula on \( N_K^{(T)} \).

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

The complexity of the algorithm described above is \( O(2^L(K+L)Ln) \). Note that \( N_K^{(T)} \) depends on the topology of the tree \( T \) in general. However, if the tree is binary, a closed form solution of the recurrence relationship (3.1), which does not depend on the topology, exists.

\textbf{Corollary 1 (closed formula binary trees).} For a rooted binary tree with \( n \) tips, we have

\[ N_K^{(T)} = N_{K+1}^{(n)} = |S_{K+1}^n| = \binom{2n-2-K}{K} \]

and

\[ M_K^{(T)} = M_K^{(n)} = \binom{2n-K}{K-1}. \]

The demonstration of this formula is not straightforward and is based on a Vandermonde-like equality, detailed in the on-line appendix C. The formula is then obtained by using a strong induction on the number of tips of the tree.

\textbf{Remark 2.} When \( K \) is large compared with \( \sqrt{n} \), the average number of configurations per
equivalence class goes to $\infty$. This can be checked by comparing the total number of configurations $\frac{2^{n-K-1}}{K-1}$ with the total number of classes $\frac{2^{n-K-1}}{K-1}$. As a consequence, we consider only models for which $K < \sqrt{n}$ in the remainder.

Remark 3. This formula has already been 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. Steel (1992) and Semple and Steel (2003) were 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 $|S_{r-1}^{PI}|$.

3.3. Another characterization of parsimony

The following proposition gives an alternative definition of parsimony under the no-homoplasy hypothesis by using the linear formulation of the problem. It will be used for model selection in Section 4.2. Its technical proof is postponed to the on-line appendix A.

Proposition 6 (equivalence between parsimony and independence). Let $m_Y$ be a given mean vector, $m_Y \in \mathbb{R}^n$, and $\Delta$ a vector of shifts such that $T\Delta = m_Y$, with $T$ the tree matrix defined in Section 2.3. Under the no-homoplasy assumption, the vector of shifts $\Delta$ is parsimonious if and only if the corresponding column vectors of the tree matrix $(T_i)_{i \in \text{ supp } (\Delta)}$ are linearly independent.

4. Statistical inference

4.1. Expectation–maximization

4.1.1. Principle

As shown in Section 2.2, both BM and OUsun models can be seen as incomplete-data models. The EM algorithm (Dempster et al., 1977) is a widely used algorithm for likelihood maximization of these kinds of model. It is based on the decomposition $\log \{ p_{\theta}(Y) \} = \mathbb{E}_{\theta} [ \log \{ p_{\theta}(Z, Y) \} | Y ] - \mathbb{E}_{\theta} [ \log \{ p_{\theta}(Z|Y) \} | Y ]$. Given an estimate $\theta^{(h)}$ of the parameters, we need to compute some moments of $p_{\theta^{(h)}}(Z|Y)$ (E-step) and then find a new estimate $\theta^{(h+1)} = \arg \max_{\theta} \mathbb{E}_{\theta^{(h)}} [ \log \{ p_{\theta}(Z, Y) \} | Y ]$ (M-step). The parameters are given for BM and OUsun in Section 2.1. We assume here that the number of shifts $K$ is fixed.

We provide only the main steps of the EM algorithm. Additional details can be found in the on-line appendix D.

4.1.2. E-step

As $X$ is Gaussian, the law of the hidden variables $Z$ knowing the observed variables $Y$ is entirely defined by its expectation and variance–covariance matrix and can be computed by using classical formulae for Gaussian conditioning. The moments of $Z|Y$ needed can also be computed by 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).

4.1.3. Complete-likelihood computation

Using the model that was described in Section 2.2, we can use the following decomposition of the complete likelihood:
\[ p_{\theta}(X) = p_{\theta}(X_1) \prod_{j=2}^{m+n} p_{\theta}(X_j | X_{pa(j)}). \]

Each term of this product is then known, and we easily obtain \( \mathbb{E}_{\theta^0}[\log\{p_{\theta}(Z, Y)\}|Y] \).

### 4.1.4. M-step

The difficulty comes here from the discrete variables (location of shifts on the branches). The maximization is exact for the BM process but we only raise the objective function for OUsun: hence computing a generalized EM algorithm (see Dempster et al. (1977)). This stems from the independent increment nature of BM: shifts affect only \( p_{\theta}(X_j | X_{pa(j)}) \) 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 auto-correlated increments: shifts affect \( p_{\theta}(X_j | X_{pa(j)}) \) on the branches where they occur and on all subsequent branches. Maximization is therefore akin to segmentation on a tree, which has complexity \( O(n^K) \).

### 4.1.5. Initialization

Initialization is always a crucial step when using an EM algorithm. Here, we use the linear formulation (2.4) or (2.6) and initialize the vector of shifts by using a lasso regression. The selection strength \( \alpha \) is initialized by using pairs of tips that are likely to be in the same group.

### 4.2. Model selection

#### 4.2.1. Model selection in the independent and identically distributed errors case with unknown variance

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

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

\[ Y' = s' + \gamma E' \quad E' \sim \mathcal{N}(0, I_n) \]

and we define a collection \( S' = \{ S'_{\eta}, \eta \in \mathcal{M} \} \) 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 \( s'_{\eta} = \text{Proj}_{S'_{\eta}}(Y') \) the orthogonal projection of \( Y' \) on \( S'_{\eta} \), i.e. a least square estimator of \( s' \), and \( s_{\eta} = \text{Proj}_{S_{\eta}}(s') \) the projection of \( s' \).

We extract from Baraud et al. (2009) the following theorem, which 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 5** (Baraud et al. (2009), section 4, definitions 2 and 3). Let \( D \) and \( N \) be two positive integers, and \( X_D \) and \( X_N \) be two independent \( \chi^2 \) random variables with degrees of freedom \( D \) and \( N \) respectively. For \( x \leq 0 \), define

\[
\text{Dkhi}[D, N, x] = \frac{1}{\mathbb{E}[X_D]} \mathbb{E} \left[ \left( X_D - x \frac{X_N}{N} \right)^+ \right]
\]
and define $EDkhi[D, N, q]$ as the unique solution of the equation $Dkhi[D, N, EDkhi[D, N, q]] = q$ (for $0 < q \leq 1$).

**Theorem 1** (Baraud et al. (2009), section 4, theorem 2 and corollary 1). In the setting that was defined above, let $D_\eta$ be the dimension of $S_\eta'$, 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 $\Omega^2 = \sum_{\eta \in \mathcal{M}} (D_\eta + 1) \exp(-L_\eta) < \infty$, and assume that, for $A > 1$,

$$pen(\eta) = pen_A, \mathcal{L}(\eta) = A \frac{N_\eta}{N_\eta - 1} EDkhi[D_\eta + 1, N_\eta - 1, \exp(-L_\eta)].$$

Take $\hat{\eta}$ as the minimizer of the criterion $\hat{\eta} = \arg\min_{\eta \in \mathcal{M}} \| Y' - \hat{s}_\eta' \|^2 / \{ 1 + pen(\eta)/N_\eta \}$.

Then, assuming that $N_\eta \geq 7$ and $\max(L_\eta, D_\eta) \leq \kappa n$ for any $\eta \in \mathcal{M}$, with $\kappa < 1$, the following non-asymptotic bound holds:

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

The penalty that is 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. 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.

### 4.2.2. Adaptation to the tree-structured framework

We use the linear formulation that was described in expression (2.3) and assume that we are in OU-sun (this procedure would also work for a BM process with a deterministic root). Then, if $V$ is a matrix of size $n$, with $V_{ij} = \exp(-\alpha i j), V(i, j) \in [1, n^2]$, we have

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

We assume that $\alpha$ is fixed, so that the design matrix $TW(\alpha)$ and the structure matrix $V$ 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 $\Delta$ (with the constraint that the first component, the root, is always included in the model). We denote by $\mathcal{M} = \cup_{K=0}^{p-1} S_\mathcal{P}^K$ 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 6, for $\eta \in \mathcal{M}$, the column vectors $T_\eta$ are linearly independent, and the model $S_\eta = \text{span}(T_\eta, i \in \eta)$ is a linear subspace of $\mathbb{R}^n$ of dimension $D_\eta = |\eta| = K_\eta + 1$, $K_\eta$ being the number of shifts in model $\eta$. As $W(\alpha)$ is diagonal invertible, it does not affect the definition of the linear subspaces. The set of models is then $S = \{ S_\eta, \eta \in \mathcal{M} \}$.

We define the Mahalanobis norm that is associated with $V^{-1}$ by $\| R \|_{V^{-1}} = R^T V^{-1} R, V R \in \mathbb{R}^n$.

The projection on $S_\eta$ according to the metric defined by $V^{-1}$ is then

$$s_\eta = \text{Proj}_{S_\eta}^{-1} (Y) = \arg \min_{a \in S_\eta} \| Y - a \|^2_{V^{-1}}$$

and

$$\hat{s}_\eta = \text{Proj}_{S_\eta}^{-1} (s).$$

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

$$\hat{s}_K = \arg \min_{\eta \in S, |\eta| = K + 1} \| Y - \hat{s}_\eta \|^2_{V^{-1}}.$$
The idea is then to slice the collection of models by the number of shifts \( K \) that they employ. Thanks to the EM algorithm above, we can 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 overfitting, using the model selection procedure that was described above, we select \( K \) by using the penalized criterion

\[
\text{Crit}_{LS}(K) = \parallel Y - \hat{s}_K \parallel_{V^{-1}}^2 \left( 1 + \frac{\text{pen}(K)}{n-K-1} \right).
\]

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

\[
\text{Crit}_{LL}(K) = \frac{n}{2} \log \left( \frac{\parallel Y - \hat{s}_K \parallel_{V^{-1}}^2}{n} \right) + \frac{1}{2} \text{pen}'(K)
\]

with \( \text{pen}'(K) = n \log \{1 + \text{pen}(K)/(n-K-1)\} \). As we use maximum likelihood estimators, we chose this formulation for the implementation. The following proposition then holds.

**Proposition 7** (form of the penalty and guaranties (\( \alpha \) known)). Let \( L = \{ L_K \}_{K \in [0, p-1]} \), with \( p \leq \min \left[ \kappa n/\{2 + \log(2) + \log(n)\}, n-7 \right] \), the maximum dimension of a model, with \( \kappa < 1 \), and

\[
L_K = \log |S_{PI}^K| + 2 \log(K+2), \quad \forall K \in [0, p-1].
\]

Let \( A > 1 \) and assume that \( \hat{K} \) is a minimizer of equation (4.1) or (4.2) with this penalty. Then

\[
E \left[ \frac{\parallel s - \hat{s}_K \parallel_{V^{-1}}^2}{\gamma^2} \right] \leq C(A, \kappa) \inf_{\eta \in \mathcal{M}} \left[ \frac{\parallel s - s_\eta \parallel_{V^{-1}}^2}{\gamma^2} + (K_\eta + 2) \{3 + \log(n)\} \right]
\]

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

The proof of this proposition can be found in the on-line appendix E. It relies on theorem 1, adapting it to our tree-structured observations.

**Remark 4.** With this oracle inequality, we can see that we are missing the oracle by a \( \log(n) \)-term. This term is known to be unavoidable; see Baraud et al. (2009) for further explanations.

**Remark 5.** Note that the chosen penalty may depend on the topology of the tree through the term \( |S_{PI}^K| \) (see Section 3.2).

**Remark 5.** 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 OUsun, with varying parameters. The simulation scheme is inspired by the work of Uyeda and Harmon (2014). We first generated three distinct trees with, 64, 128 and 256 tips, using a pure birth process with birth rate \( \lambda = 0.1 \). The tree heights were scaled to 1, 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 a time. The base scenario was taken to be relatively ‘easy’, with \( \beta_1 = 0 \) (this parameter was fixed for the rest of the simulations), \( \alpha_b = 3 \) (i.e. \( t_{1/2,b} = 23\% \)), \( \gamma_2^2 = 0.5 \) 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 \( \gamma_2^2 = \sigma^2/(2\alpha) \) took nine values in \([0.05, 25]\); the number of shifts \( K \) took nine values in \([0, 16]\) (see Figs 7 and 8 in Section 5.4 for the exact values taken). The problem was all the more difficult because \( \gamma^2, t_{1/2} \) 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 \( N(4, 1) \) and \( N(-4, 1) \) 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 kept only parsimonious allocations.

Each of these configurations was repeated 200 times, leading to 16 200 simulated data sets. An instance of a tree with the generated data is plotted in Fig. 6.

### 5.2. Inference procedures

For each generated data set, 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. We 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 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 \( \alpha \) either known or estimated. The computations took respectively 66 and 570 (cumulated) days of central processor unit time. This amounts to a mean computational time of around 6 min (367 s) for one estimation when \( \alpha \) is fixed, and 52 min (3137 s) when \( \alpha \) 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 the mean number of EM steps needed when \( \alpha \) is fixed or estimated. The quality of the estimates of \( \beta_1, t_{1/2} \) and \( \gamma^2 \) was assessed by using the coefficient of variation. The model selection procedure was evaluated by comparing the true number of shifts with the estimated number, which should be lower. We do not expect to find the exact number as some shifts, which are too small or too close to the tips, cannot be detected. To evaluate the quality of the clustering of the tips, the only quantity that we can observe, we used the adjusted Rand index (ARI) (Hubert and Arabie, 1985) between the true clustering of the tips, and the clustering that was induced by the estimated shifts. The ARI is proportional to the number of concordant pairs in two clusterings and has a 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 strength of selection is notoriously difficult to estimate, with large ranges of values giving similar behaviours (see Thomas et al. (2014)). We hence first analyse the effect of estimating \( \alpha \) in our estimations, showing that the main behaviour of the algorithm stays the same. Then, we study the shifts reconstruction procedure.
Fig. 6. (a) 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 colours of the branches correspond to the true regimes, black being the ancestral state) and (b) one of the three equivalent allocations of shifts for the model inferred from the data, with corresponding vector of mean tip values (shifts that are not recovered are on pendant edges and have low influence on the data; the two other equivalent allocations can be easily deduced from this one)
5.4.1. Convergence and likelihood
For $\alpha$ known, all estimations converged in fewer than 49 iterations, with a median number of 13 iterations. For $\alpha$ estimated, the number of iterations increased greatly, with a median of 69, and a fraction of estimations (around 3.2%) that reached the maximum allowed number (fixed at 1000 iterations) without converging. Unsurprisingly, the more difficult the problem, the more iterations were needed. The log-likelihoods of the estimated parameters are close to the true log-likelihoods even when $\alpha$ is estimated (see the on-line supplementary Fig. S5 in appendix F, first row).

5.4.2. Estimation of continuous parameters
Fig. 7(a) shows that we tend to overestimate $\alpha$ slightly in general. The estimation is particularly bad for large values of $\alpha$ (with a high variance on the result, see the first box of the row), and low values of $\alpha$. In this regime, the model is ‘overconfident’, as it finds a higher selection strength than the real strength and therefore a smaller variance (Fig. 7(b)). 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 number slightly degrades our estimation of these parameters (see the on-line supplementary Fig. S5 in appendix F). The estimation of $\beta_1$ is not affected by knowledge of $\alpha$ or $K$ (see Fig. 8(a)) and only has an increased variance for more difficult configurations. In the remainder of this section we show only results obtained for estimated $\alpha$ as estimating $\alpha$ does not impact ARI, $\hat{K}$ and $\hat{\beta}_0$ (see the on-line supplementary Fig. S6 in appendix F).

5.4.3. Estimation of the number of shifts
The way that shifts were drawn ensures that they are not too small on average, and that they are

\[
\begin{align*}
\Delta t_{1/2} &= \ln(2) / \alpha \\
\gamma^2 &= \chi^2_n - \chi^2
\end{align*}
\]

Fig. 7. Box plots over the 200 repetitions of each set of parameters, for (a) the phylogenetic half-life and (b) root variance with $K$ estimated, and $\alpha$ fixed at its true value (■) or estimated (□), on a tree with 128 taxa: for better legibility, the $y$-axes were rescaled, omitting some outliers (for $t_{1/2}$ and $\gamma^2$, respectively 0.82% and 0.46% of points are omitted); the whisker of the first box for $t_{1/2}$ goes up to 7.5)
located all along the tree. Still, some shifts have a very small influence on the data and are hence difficult to detect (see Fig. 6). The selection model procedure almost always underestimates the number of shifts, except in very favourable cases (Fig. 8(b)). This behaviour is nonetheless expected, as allowing more shifts does not guarantee that the right shifts will be found (see the on-line supplementary Figs S3 and S4 in appendix F).

5.4.4. Clustering of the tips
The ARI tends to be degraded for small values of \(\alpha\) or high variance, but it remains positive (Fig. 8(c)). 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.

5.4.5. Equivalent solutions
When \(\alpha\) and \(K\) are both estimated, only 5.1% of the configurations have two or more equivalent solutions. One inferred configuration with three equivalent solutions is presented in Fig. 6.

5.4.6. Comparison with bayou
As mentioned above, our simulation scheme, although not completely equivalent to the scheme

![Fig. 8. Boxplots over the 200 repetitions of each set of parameters, for (a) \(\beta_1\), (b) the number of shifts and (c) the ARI, with \(\alpha\) estimated, and \(K\) fixed at its true value \(\square\) or estimated \(\blacksquare\): as previously, the y-axis of (a) \((\beta_1)\) was rescaled, omitting some outliers (1.39% of points have been omitted).]
that was used in Uyeda and Harmon (2014), is very similar, so we can compare our results with theirs. The main differences lie in the facts that we took a grid on $\gamma^2 = \sigma^2 / (2\alpha)$ instead of $\sigma^2$, and that we took shifts with higher intensities, making the detection of shifts easier. We can see that we obtain the same qualitative behaviours for our estimators, with the selection strength $\alpha$ overestimated or underestimated in small or large values regions. The main difference lies in the estimation of the number of shifts. Maybe because of the priors that they used ($K \sim \text{conditional Poisson}(\lambda = 9, K_{\text{max}} = n/2)$), they tend to estimate similar numbers of shifts (centred on $9$) for any set of parameters. In particular, although our method seems to be quite good at detecting situations where there are no shifts at all, theirs seems unable to catch these kinds of configuration, despite the fact that their shifts have low intensity, leading to a possible overfitting of the data.

Overall, the behaviour of the algorithm is quite satisfying. Our model selection procedure avoids overfitting, while recovering the correct clustering structure of the data. It furthermore allows a reasonable estimation of the continuous parameters, except for $\alpha$ which is notoriously difficult to estimate.

6. Case-study: Chelonian carapace length evolution

6.1. Description of the data set

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). To test the hypothesis of island and marine gigantism, which could explain the extreme variations observed, Jaffe et al. (2011) compiled a data set containing a measure of the carapace
length for 226 species, along with a phylogenetic tree of these species, spanning 210 million years (Fig. 9). They assigned each species to one of four habitats: mainland terrestrial, freshwater, marine and island terrestrial. Then, testing several fixed regime allocations on the branches of the tree by using the method that is described in Butler and King (2004), they found the best support in favour of an ‘OU’ model that assigned one regime to each habitat. Following Uyeda and Harmon (2014), we shall refer to this model as ‘OU_{habitat}'. Note that this model is ambiguously defined, as it requires us to assign a habitat to each ancestral species. Using proposition 3, we found that there were 48 equivalent parsimonious ways of doing so that respect the habitats that are observed at the tips of the tree. One of these habitat reconstructions is presented Fig. 9.

6.2. Method

We used the version of the data set that is embedded in the package geiger (Harmon et al., 2008), which contains a phylogenetic tree and a vector of log-carapace-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. Rather than estimating $\alpha$ directly within the EM algorithm as we did for the simulations, we took $\alpha$ varying on a grid, taking six values regularly spaced between 0.01 and 0.1, but fixed for each estimation. We found that this approach, although computationally more intensive, gave better results. These $6 \times 20 = 120$ estimations took around 2 h of central processor unit 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 five shifts, and a selection strength of 0.06 (i.e. 5.5% of the total height of the tree). Using a finer grid for $\alpha$ 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 five 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 that were defined in Jaffe et al. (2011) (see Fig. 9). The ancestral optimal value, which 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 (superfamily Chelonioida)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Results for the following methods:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Habitat</td>
</tr>
<tr>
<td>Number of shifts</td>
<td>16</td>
</tr>
<tr>
<td>Number of regimes</td>
<td>4</td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>−133.86</td>
</tr>
<tr>
<td>Marginal log-likelihood</td>
<td>—</td>
</tr>
<tr>
<td>$\alpha (\times h, \text{ per million years})$</td>
<td>9.32</td>
</tr>
<tr>
<td>$\ln(2)/\alpha (\text{million years})$</td>
<td>15.56</td>
</tr>
<tr>
<td>$\sigma^2 (\times h, \text{ per million years})$</td>
<td>6.21</td>
</tr>
<tr>
<td>$\gamma^2$</td>
<td>0.33</td>
</tr>
<tr>
<td>Central processor unit time (min)</td>
<td>65.25</td>
</tr>
</tbody>
</table>

†For bayou, the median of the posterior distributions is given.
are found to have an increased carapace length (with an optimal value of 130 cm), as well as a clade containing soft shell tortoises (family *Trionychidae*; optimal size 110 cm), and a clade containing almost all island tortoises, including several subspecies 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 contains some mainland tortoises of the genus *Geochelone* also, which 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 $1.4 \times 10^{-20}$ cm, for a measured trait of 15 cm. The fact that the shift has a very high negative value ($-49$ on a log-scale) is just an artefact due to the actualization factor on a very small branch (0.18 million years, for an inferred phylogenetic half-life of 11 million years). This is a rather unexpected choice of shift location. When considering the linear model as transformed by the Cholesky matrix of the variance to obtain independent errors (as in the proof of proposition 7), 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

To compare our results with previously published results, we reproduced some of the analysis already conducted on this data set. We hence ran the methods that were described in Jaffe *et al.* (2011) (using the R package **OUwie**, with fixed positions for the shifts), Uyeda and Harmon (2014) (implemented in the package **bayou**), Ingram and Mahler (2013) (package **SURFACE**) and Ho and Ané (2014) (function **OUshifts** in package **phylolm**). See section F.3 in the on-line appendix F for more details on these methods and the parameters that we used.

The shifts that were allocated on the tree by methods **bayou**, **SURFACE** and **OUshifts** are presented in on-line supplementary Fig. S7 (appendix F). We can see that three among the most strongly supported shifts in the posterior distribution given by **bayou**, as well as some among the oldest shifts that were found by **SURFACE** and **OUshifts**, are similar to those 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 quite close to the mode of the prior put on the number of shifts (15). The **SURFACE** and **OUshifts** methods select respectively 33 and eight shifts, including many on pendant edges, which are not easily interpretable. The backward step of **SURFACE** allowed us to merge the regimes that were found for marine turtles and soft shell tortoises that our method found to have very similar optimal values. The results of the five methods are summarized in Table 1. Note that these models are not nested, because of the status assigned to the root, and because of the possible convergences.

Compared with stepwise 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, which are more easily interpretable than the solutions that were found by other methods, and that do not rely on any chosen prior.

### Acknowledgements

We thank Cécile Ané for helpful discussions on an early draft of this paper. We are grateful to the Institut Nationale de Recherche Agronomique MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for providing the computational resources that are needed for the experiments. We also thank the two reviewers whose careful and critical reading greatly helped to improve this paper.
This research was conducted at the two following research units: the Unité Mixte de Recherche Mathématiques et Informatique Appliquées, Paris, AgroParisTech, Institut National de la Recherche Agronomique, Université Paris-Saclay, 75005, Paris, France (PB and SR), and Mathématiques et Informatique Appliquées du Génome à l’Environnement, Institut National de la Recherche Agronomique, Université Paris-Saclay, 78352 Jouy-en-Josas, France (PB and MM).

References


Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

‘Appendices for: “Detection of adaptive shifts on phylogenies using shifted stochastic process on a tree”’. 
