A comparison of ancestral state reconstruction methods for quantitative characters

Manuela Royer-Carenzi, Gilles Didier

A comparison of ancestral state reconstruction methods for quantitative characters

Manuela Royer-Carenzi and Gilles Didier
Aix-Marseille Université, CNRS, Centrale Marseille, I2M, UMR 7373,
13453 Marseille, FRANCE

February 25, 2022

Abstract

Choosing an ancestral state reconstruction method among the alternatives available for quantitative characters may be puzzling. We present here a comparison of seven of them, namely the maximum likelihood, restricted maximum likelihood, generalized least squares under Brownian, Brownian-with-trend and Ornstein-Uhlenbeck models, phylogenetic independent contrasts and squared parsimony methods.

A review of the relations between these methods shows that the maximum likelihood, the restricted maximum likelihood and the generalized least squares under Brownian model infer the same ancestral states and can only be distinguished by the distributions accounting for the reconstruction uncertainty which they provide.

The respective accuracy of the methods is assessed over character evolution simulated under a Brownian motion with (and without) directional or stabilizing selection. We give the general form of ancestral state distributions conditioned on leaf states under the simulation models.

Ancestral distributions are used first, to give a theoretical lower bound of the expected reconstruction error, and second, to develop an original evaluation scheme which is more efficient than comparing the reconstructed and the simulated states.

Our simulations show that: (i) the distributions of the reconstruction uncertainty provided by the methods generally make sense (some more than others); (ii) it is essential to detect the presence of an evolutionary trend and to choose a reconstruction method accordingly (iii) all the methods show good performances on characters under stabilizing selection; (iv) without trend or stabilizing selection, the maximum likelihood method is generally the most accurate.

Keywords:

Ancestral state reconstruction, Maximum likelihood, Brownian motion, Energy distance

Introduction

Besides being essential to understand the process of character evolution, ancestral state reconstruction plays an important role in the study of ecological diversification and comparative analysis. We focus here on quantitative characters, i.e. measured as continuous variables such as weight, size etc.

From a methodological point of view, ancestral state reconstruction is a challenging problem which has been addressed by several approaches. The general question can be stated as follows. Taking as inputs the phylogeny of a set of organisms (given as a tree with branch lengths) and their character states, a reconstruction method has to infer - as accurately as possible - the character states of the ancestral

organisms. The reconstruction approaches fall into two major classes: methods based on the parsimony principle (Fitch 1971; Swofford and Maddison 1987; Maddison 1991; Collins et al. 1994), whose general idea is to impute the missing values of the tree by minimizing the sum of distances between ancestors and their direct descendant characters, and methods based on stochastic models of character evolution, mainly Brownian motion for continuous traits (Schluter et al. 1997; Pagel 1999b; Huelsenbeck and Ronquist 2001; Nielsen 2002). Several authors discuss the advantages of stochastic approaches over parsimonious ones (Schluter et al. 1997; Mooers and Schluter 1999; Pagel 1999b; Nielsen 2002; Huelsenbeck et al. 2003). An important point is that stochastic approaches take into account divergence times (branch lengths) while parsimonious methods do not. Moreover, stochastic approaches may provide probability distributions of the reconstructed ancestral states, accounting for their uncertainty and which can be used to develop hypothesis testing and confidence intervals.

In our study, we focus on seven reconstruction methods, namely the maximum likelihood, restricted maximum likelihood, generalized least squares under Brownian, Brownian-with-trend and Ornstein-Uhlenbeck models, phylogenetic independent contrasts and squared parsimony methods. Before comparing their accuracy, we review the methods and their relationship to each other. It turns out that the first three ones reconstruct the same ancestral states. These three methods may still be distinguished, and to some extent compared, since they provide different probability distributions of their uncertainty. There are a few model-based approaches which do not rely on the Brownian assumption. For instance, in Hansen (1997); Martins and Hansen (1997); Pagel (1998, 1999a), authors consider ancestral states reconstructions under the assumption that the character follows either a Brownian motion with trend or an Ornstein-Uhlenbeck model, corresponding to a directional or a stabilizing selection respectively (Hansen and Martins 1996). To our knowledge, the only available reconstruction approaches based on directional or stabilizing model are provided by the java program *COMPARE* which performs general least squares reconstructions according several models (Martins 1995), by the computer package BayesTraits (Pagel et al. 2004; Pagel and Meade 2013), which uses Markov chain Monte Carlo (MCMC) methods to infer ancestral states, and by the R-package *phytools*, which performs, through numerical optimization, maximum likelihood reconstructions under a Brownian motion with trend (Revell 2012).

Evaluating the respective performances of these methods is a natural and important question. Works aiming at answering this question proceed by comparing the reconstructed states with reference "trusted" ones. Such reference values for ancestral states may be obtained either by considering fossil character states or by simulating, via a stochastic model, artificial evolution of the character and by keeping track of the ancestral states observed during simulations (Martins 1999). Webster and Purvis (2002) and Oakley and Cunningham (2000) assess several reconstruction methods with regard to measurements on fossils. They both observe that the methods are confounded by an evolutionary trend toward increasing size.

Our comparison of the seven methods is based on artificial evolution simulated under Brownian motions with and without directional or stabilizing selection. The artificial evolution runs on the phylogenetic tree of Pleistocene planktic Foraminifera (Webster and Purvis 2002). Besides the fact that we consider evolution models with directional or stabilizing selection, a noticeable difference with previous works is that the reconstructed states are compared with regard to the ancestral state distributions conditioned on the simulated leaves, rather than with the simulated ancestral states as it is done usually. Intuitively, in this way, we compare the reconstructed state with all the possible realizations of the evolution process with the given simulated leaf states. Moreover the ancestral distribution conditioned on the leaves does reflect the uncertainty inherent to the stochastic character of evolution as modeled in simulations. In particular, it allows us to determine a lower bound of the expected reconstruction error as well as the reconstructed state achieving this lower bound. This can be seen as a transposition of ideas of (Steel and Szekely 1999) and (Royer-Carenzi et al. 2013).

Another motivation of this work is to assess the relevance of the distributions provided by the methods for the reconstruction uncertainty. These distributions are expected to provide a greater amount of information than single values for ancestral states (Schluter et al. 1997; Polly 2001). Altogether with our new comparison scheme, we compare the conditional ancestral distributions given the leaves with the distributions provided by the methods. A distance between distributions, called the *Energy distance* offers us a consistent framework to compare both reconstructed states and reconstructed probability distributions, with ancestral state distributions conditioned on leaves (Szekely and Rizzo 2013). The Energy distance is strongly related to the absolute bias.

Finally, we provide exact, matrix-based, implementations of Brownian-based methods which were

formerly based on numerical optimization algorithms. Some of our R-scripts have been incorporated into the `reconstruct` function of the `ape` R-package since version 3.2 (Paradis et al. 2004, `https://cran.r-project.org/web/packages/ape/index.html`). We also provide matrix-based implementations of generalized least square (and equivalently maximum likelihood) reconstructions under Brownian motion with trend and Ornstein-Uhlenbeck models. Our R-scripts are available at `https://github.com/gilles-didier/Reconstruction.git`.

The rest of the paper is organized as follows. In Section 1, we present three standard models of quantitative character evolution. Section 2 briefly describes the reconstruction methods and shows how they are related. Section 3 is devoted to our assessment protocol. We provide the form of the ancestral distributions conditioned on the leaf states under the simulation. These ancestral distributions are next used to define our evaluation protocol and to give a lower bound of the expected reconstruction error. In its final version, the protocol is based on the Energy distance between probability distributions, both for assessing the reconstructed states and the distribution provided by the methods. The results of our simulations are finally presented and discussed in Section 4.

1 Models of evolution for quantitative characters

1.1 Phylogenetic trees - Notations

In the standard ancestral character reconstruction problem, one assumes that the evolutionary history of the species is known and given as a rooted phylogenetic tree with branch lengths. Our typical tree contains $n+1$ nodes (including leaves), among which r are internal nodes (excluding the root). By convention, the nodes are indexed in the following way:

- index 0 for the root,
- indices 1 to r for the other internal nodes,
- indices $r+1$ to n for the leaves.

The nodes are numbered in such a way that if a node j descends from a node i then $j > i$. We put $p(j)$ for the index of the direct ancestor of the node j, τ_j for the length of the branch leading to j and T_j for the the sum of the branch lengths between the root and j. Being given two nodes i and j, we put $m(i,j)$ for the index of their most recent common ancestor (mrca).

Let X be a random variable. We write f_X for its density function and $\mathbb{E}(X)$ for its expectation.

1.2 Models of evolution

We make the standard assumptions that character evolves independently along the branches of the phylogenetic tree and that its evolution is homogeneous both through time and lineages. In order to actually model the evolution of a character, we first need to define an initial probability density f_{Z_0} for the state of the root. In the simulation models, f_{Z_0} is the degenerate density at a given value z_0, i.e. our simulations all start from a given root state z_0 which is a parameter of the model. The reconstruction methods assume an improper flat density as initial probability density f_{Z_0} (i.e. $f_{Z_0}(x) = 1$ for all x).

1.2.1 Brownian motion model

The *Brownian motion* (**BM**) model is the simplest stochastic process able to model the evolution of a quantitative character (Felsenstein 1985; Schluter et al. 1997). Under this model, evolution is neutral and governed by a rate parameter σ which accounts for its diffusion. Formally, along a branch of the tree, the stochastic process (X_t) accounting for the character state has the form:

$$\mathrm{d}X_t = \sigma \mathrm{d}B_t, \quad X_0 = x_0,$$

where (B_t) denotes the standard Brownian motion, defined as a centered Gaussian process with stationary and independent increments, and $B_t \sim \mathcal{N}(0,t)$, where $\mathcal{N}(\mu,\sigma^2)$ is the Gaussian distribution of mean μ and variance σ^2. Thus increments $X_{t+s} - X_t$ $(s > 0)$ are independent with law $\mathcal{N}\left(0,\sigma^2 s\right)$.

1.2.2 Arithmetic Brownian motion model

Biological evolution is not always assumed to be neutral. For instance, Cope's rule states that species tend to increase in body size over time (Kingsolver and Pfennig 2004; Van Valkenburgh et al. 2004; Hone and Benton 2005). In Webster and Purvis (2002), fossil evidence suggests that a neutral process cannot model the evolution of Pleistocene planktic Foraminifera size since it tends to increase with time. A similar observation is made by Oakley and Cunningham (2000).

The *Arithmetic Brownian motion* (**ABM**), sometimes called Brownian motion with trend, yields to model a linear deterministic trend μ, which can be either positive or negative. Along a branch of the tree, the stochastic process (X_t) of the character state now has the form:

$$\mathrm{d}X_t = \mu \mathrm{d}t + \sigma \mathrm{d}B_t, \quad X_0 = x_0.$$

Then increments $X_{t+s} - X_t$ $(s > 0)$ are independent with law $\mathcal{N}\left(\mu s, \sigma^2 s\right)$.

Remark 1. *A **BM** model is nothing but an **ABM** model with trend-parameter $\mu = 0$.*

1.2.3 Ornstein-Uhlenbeck model

The Ornstein-Uhlenbeck process (**OU**) was introduced by Vasicek (1977) and models the evolution of a character attracted toward an optimum trait value θ with selection strength $\alpha \geq 0$. It is appropriate for modeling traits that evolve under certain constraints such as stabilizing selection (Lande 1976; Felsenstein 1988; Martins 1994; Hansen and Martins 1996; Hansen 1997; Butler et al. 2000; Harmon et al. 2010). Along a branch of the tree, the stochastic process (X_t) of the character state now has the form:

$$\mathrm{d}X_t = \alpha(\theta - X_t)\mathrm{d}t + \sigma \mathrm{d}B_t, \quad X_0 = x_0.$$

When $\alpha > 0$, the *corrected* increments of an **OU** process $X_{t+s} - e^{-\alpha s} X_t$ $(s > 0)$ are independent with law

$$\mathcal{N}\left(\theta(1 - e^{-\alpha s}), \sigma^2 \frac{1 - e^{-2\alpha s}}{2\alpha}\right).$$

As illustrated in Figure 1, an Ornstein-Uhlenbeck process models evolution with two different stages. Indeed, while the optimum θ is not reached, the phenotype has a directional evolution to this optimum (not in a linear way). And as soon as θ is reached, the phenotype is constrained to remain close to the optimum. The larger α, or the smaller $|x_0 - \theta|$, the faster the optimum is reached (Göing-Jaeschke and Yor 2003). Note that, since the Ornstein-Uhlenbeck process is memoryless, if all the observed values are sampled from the constrained regime, the part of the evolution before the constrained or the (quasi-)stationary regime, including the initial value x_0, cannot be inferred. Thus we will rather consider the particular case where the process directly starts in the constrained regime by assuming the optimum value θ to be equal to the initial value x_0. Such a model will be referred to as an **OU*** model, which is referred to as the single stationary peak model in (Harmon et al. 2010).

Remark 2. *When the time goes to ∞, the character state converges to a stationary distribution, that is $\mathcal{N}\left(\theta, \frac{\sigma^2}{2\alpha}\right)$.*

Remark 3. *A **BM** model is nothing but an **OU*** or an **OU** model with strength of the restraining force $\alpha = 0$.*

1.3 Likelihood of a character evolution

Let us consider a particular realization of the evolution process, which is known only through the vector $(z_0, z_1, \cdots, z_n)$ of the character states at the nodes of the tree (i.e. entry z_i is the character state at node i). The increments of the character state between nodes and their children give us a natural expression of the likelihood of such a vector. Let us put $\phi(., \gamma^2)$ for the density of a centered normal law with variance γ^2:

$$\phi(x, \gamma^2) = \frac{1}{\sqrt{2\pi\gamma^2}} e^{\frac{-x^2}{2\gamma^2}}.$$

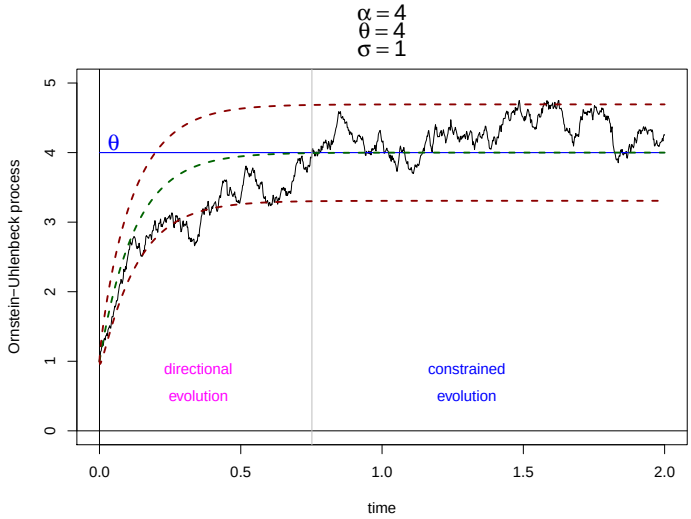

Figure 1: Simulation of an Ornstein-Uhlenbeck process, illustrating the directional and the constrained evolution. Green dashed line shows the expected mean of the Ornstein-Uhlenbeck process, whereas red dashed lines indicate its 95%-fluctuation interval.

Under the **BM** model with parameter σ^2 and root probability density f_{Z_0}, the likelihood of a realization $(z_0, z_1, \cdots, z_n)$ is

$$\mathcal{V}_{\sigma^2}(z_0, z_1, \cdots, z_n) = f_{Z_0}(z_0) \prod_{j=1}^{n} \phi(z_j - z_{p(j)}, \sigma^2 \tau_j).$$

The corresponding log-likelihood is

$$\log\left(\mathcal{V}_{\sigma^2}(z_0, z_1, \cdots, z_n)\right) = \log\left(f_{Z_0}(z_0)\right) - \frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2}\sum_{j=1}^{n}\log(\tau_j) - \sum_{j=1}^{n}\frac{\left(z_j - z_{p(j)}\right)^2}{2\sigma^2\tau_j}. \quad (1)$$

The same log-likelihood under an **ABM** model with parameters σ^2 and μ is:

$$\log\left(\mathcal{V}_{\sigma^2,\mu}(z_0, z_1, \cdots, z_n)\right) = \log\left(f_{Z_0}(z_0)\right) - \frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2}\sum_{j=1}^{n}\log(\tau_j) - \sum_{j=1}^{n}\frac{\left(z_j - z_{p(j)} - \mu\tau_j\right)^2}{2\sigma^2\tau_j}. \quad (2)$$

Under an **OU** model with parameters σ^2, α and θ, this log-likelihood becomes:

$$\begin{aligned}\log\left(\mathcal{V}_{\sigma^2,\alpha,\theta}(z_0, z_1, \cdots, z_n)\right) = \log\left(f_{Z_0}(z_0)\right) &- \frac{n}{2}\log(2\pi) - n\log(\sigma) + \frac{n}{2}\log(2\alpha) \\ &- \frac{1}{2}\sum_{j=1}^{n}\log(1 - e^{-\alpha\tau_j}) - \sum_{j=1}^{n}\frac{\alpha\left(z_j - z_{p(j)} - \theta(1 - e^{-\alpha\tau_j})\right)^2}{\sigma^2(1 - e^{-2\alpha\tau_j})}.\end{aligned} \quad (3)$$

2 Reconstruction methods

2.1 Presentation

2.1.1 Model-based methods

We present here four methods all relying on the assumption that the character evolves following a **BM** model with an improper flat distribution for the root state. Their current implementations return not only a reconstructed state for each internal node j but also a probability distribution of this quantity. Therefore, the reconstructed state may be seen as a random variable Y_j^R, which accounts for the reconstruction uncertainty. Hereafter, we give a brief description of these four methods:

- The Maximum Likelihood method (*ML*) infers the ancestral states which maximize their joint likelihood under a **BM** model with an improper flat distribution for the root state (Schluter et al. 1997). This maximum likelihood estimation is simultaneously performed on the ancestral states and on the variance of the **BM** model. For any internal node j, *ML* returns the reconstructed state y_j^R which is also the mean of Y_j^R and its standard deviation σ_j^R. Schluter et al. (1997) showed that $\frac{Y_j^R - y_j^R}{\sigma_j^R}$ follows a t-distribution with $r+1$ degrees of freedom. Namely, its density is:

$$f_{Y_j^R}(x) = \frac{1}{\sigma_j^R} t_{r+1}\left(\frac{x - y_j^R}{\sigma_j^R}\right),$$

  where t_{r+1} denotes the density of a t-distribution with $r+1$ degrees of freedom:

$$t_{r+1}(x) = \frac{1}{\sqrt{(r+1)\pi}} \frac{\Gamma\left(\frac{r+2}{2}\right)}{\Gamma\left(\frac{r+1}{2}\right)} \left(1 + \frac{x^2}{r+1}\right)^{-\frac{r+2}{2}},$$

  and Γ is the gamma function.

- As it is implemented in the `ape` R-package (Paradis et al. 2004), the Restricted Maximum Likelihood method (*REML*) reconstructs the ancestral states in a very similar way as *ML*. It first estimates the variance of the **BM** model. Next, the reconstructed ancestral states are those maximizing the likelihood under the **BM** model with the estimated variance. The relationship between *ML* and *REML* reconstructions is discussed in more details below. As for *ML*, $\frac{Y_j^R - y_j^R}{\sigma_j^R}$ follows a t-distribution with $r+1$ degrees of freedom.

- *PIC* is based on the Felsenstein's Phylogenetic Independent Contrasts method (Felsenstein 1985). It recursively reconstructs the states of the ancestral nodes by averaging those of their children with weights depending on branch lengths. With *PIC*, the reconstructed state of a node only depends on those of its descendants. The confidence intervals are computed by using the expected variances under the model. They only rely on the tree (not on the leaf states). For any internal node j, *PIC* provides a reconstructed state y_j^R, which also stands for the mean of Y_j^R, and the standard deviation σ_j^R of Y_j^R. The random variable $\frac{Y_j^R - y_j^R}{\sigma_j^R}$ follows a standard Gaussian distribution.

- Generalized least squares approaches reconstruct the ancestral states minimizing the residual sum of squares with regard to the variance-covariance of the states, under a given evolution model (Grafen 1989; Martins and Hansen 1997; Cunningham et al. 1998; Martins 1999). We consider four types of generalized least squares estimation: *GLS_BM*, *GLS_ABM*, *GLS_OU* and *GLS_OU**, which correspond to the **BM**, **ABM**, **OU** and **OU*** models, respectively. Since for all these models, the variance-covariance matrix may be written as a product of the diffusion parameter σ^2 and a matrix which does not depend on σ^2 (see Remark 4 below), σ^2 has no effect on ancestral states inference. Nevertheless, the expected variance of the ancestral states does depend on it (Martins and Hansen 1997). The variance σ^2 is thus needed. It can be explicitly estimated by maximum likelihood (Searle et al. 1992; Hansen 1997, or Equations 1, 2 and 3). For **ABM** models, minimizing the residual sum of squares provides ancestral states as well as the trend μ of the model (Appendix B). In the case of **OU** and **OU*** models, the strength selection α is also involved in the variance-covariance matrix and influences ancestral states inference. Since its maximum likelihood estimation is not straightforward (Hansen 1997), it has to be estimated through numerical optimization. Then, in **OU** and **OU*** models, the residual sum of squares is minimized conditionally to parameter α, for inferring ancestral states. In all the cases and following Martins and Hansen (1997), the confidence intervals of the reconstructed ancestral states are computed from the expected variances under the considered model (see Equation 8 below). For any internal node j, the random quantity $\frac{Y_j^R - y_j^R}{\sigma_j^R}$ follows a standard Gaussian distribution and Y_0^R is assumed to have the degenerate distribution at the reconstructed state of the root.

2.1.2 Parsimony-based methods

There are two main kinds of parsimonious approaches dealing with quantitative characters: linear parsimony (Swofford and Maddison 1987; Maddison and Maddison 1992) and squared parsimony (*SP*) (Maddison 1991). The first one reconstructs the unknown states of the character by minimizing the sum of the absolute differences between the state of a node and that of its direct ancestor. The second one proceeds in the same way but it considers squared differences in place of absolute ones. Since, according to Butler and Losos (1997), linear parsimony often results in many equally parsimonious reconstructions and squared parsimony gives more relevant results, we keep only *SP* for our study. Unlike the methods of Section 2.1.1, *SP* does not provide any probability distribution for the reconstructed states.

2.2 Relations between methods

All these reconstruction methods are strongly related to the Maximum Likelihood reconstruction, thus one with another. These relations were already stated here and there, sometimes without justification. We recall them and give references or elements of proofs.

2.2.1 *SP*

Minimizing the squared parsimony cost is equivalent to maximizing the log-likelihood under a **BM** model with a flat initial distribution for the root state and with all the branch lengths set to any constant value, see (Schluter et al. 1997; Maddison 1991) or Equation 1. It follows that any function computing *ML* may be used to compute *SP*. One just needs to make all the branch lengths equal before calling it.

2.2.2 *REML*

Methods *ML* and *REML* only differ in the fact that the variance of the **BM** model and the ancestor states are simultaneously estimated by maximum likelihood with *ML*, while *REML* first estimates the variance of the **BM** model and next the ancestral states. If the root state follows an improper flat distribution then the term "$\log\left(f_{Z_0}(z_0)\right)$" vanishes from Equation 1. Finding the ancestral states maximizing the log-likelihood just relies on finding the unknown values of the vector z minimizing $\sum_{j=1}^{n} \frac{\left(z_j - z_{p(j)}\right)^2}{2\sigma^2\tau_j}$. This does not depend on σ. Consequently, *ML* and *REML* do provide the same reconstructed states.

2.2.3 *GLS_BM*

Martins and Hansen (1997) state that *GLS_BM* with the Brownian variance-covariance structure reconstructs the same ancestral states as *ML*. We provide a detailed proof of this fact in Appendix A.

2.2.4 *PIC*

Maddison (1991) proved that *PIC* and *ML* reconstruct the same state for the root. Note that this only holds for the root.

2.2.5 Totally equivalent ?

Since *ML*, *REML* and *GLS_BM* all reconstruct the same ancestral states, those methods will be referred to as *ML*/*REML*/*GLS_BM* when studying reconstructed states. Nevertheless, these methods still have to be distinguished since they differ in terms of reconstructed distributions. Both *ML* and *REML* return t-distributions with $r+1$ degrees of freedom and the same mean but with different variances, while *GLS_BM* provides a Gaussian distribution with the same mean as *ML* and *REML* but with another variance.

2.2.6 *GLS_ABM*, *GLS_OU* and *GLS_OU**

Like in the **BM**-model case, the ancestral states reconstructed by the generalized least squares approaches under an **ABM**, an **OU** and an **OU*** model are the same as those maximizing the likelihood under the corresponding model (Equations 2 and 3 and Appendices B and C).

2.3 Implementation

Since former implementations of *ML*, *REML* and *GLS_BM* were based on numerical optimization algorithms, they did not always converge to global optima. We used Equations A5 of Appendix A to reconstruct ancestral states following these methods with exact matrix computations. The resulting R-function `reconstruct` is part of the `ape` R-package since version 3.2.We also implemented the *GLS_ABM*, *GLS_OU* and *GLS_OU** approaches following the equations of Appendices B and C. Our R-routines are available at `https://github.com/gilles-didier/Reconstruction.git`.

3 Assessing the performances of the reconstruction methods

We shall study the performances of the methods when the character is under directional or stabilizing evolution. To this aim, we assume that it evolves following either an **ABM** model with variance σ^2 and trend μ, or an **OU** model with variance σ^2, optimum θ and selection strength α. We consider the degenerate probability density at a given value z_0 for the root state. We start this section by giving the conditional law of an internal state given those of the leaves under such a model. Next we propose an evaluation protocol using this conditional law. The Energy distance is strongly related to this protocol and allows us to compare between distributions and/or single values in a consistent way. Finally, we show that the conditional expectation of an ancestral state given the leaves is, in a sense, the best reconstruction possible and we study its relation with the state inferred by *ML*/*REML*/*GLS_BM*.

3.1 Ancestral distributions conditioned on the leaf states

Let us assume here that the character evolution follows an **ABM** model (z_0, σ^2, μ) or an **OU** model $(z_0, \sigma^2, \alpha, \theta)$, including the particular case of an **OU*** model $(z_0, \sigma^2, \alpha, \theta = z_0)$. We put Z_i for the random variable of the state i and Z, $Z^{|a}$ and $Z^{|l}$ for the random vectors ${}^t(Z_1, \cdots, Z_n)$, ${}^t(Z_1, \ldots, Z_r)$ and ${}^t(Z_{r+1}, \ldots, Z_n)$, corresponding to all the nodes except the root, the internal nodes excluding root, and the leaves, respectively. A set of node states $z_0, \ldots, z_n$ is organized as vectors z, $z^{|a}$ and $z^{|l}$ accordingly.

Let $U = {}^t(U_1, U_2, \cdots, U_n)$ be the random vector of increments, *corrected* for **OU**. Namely under the **ABM** model (z_0, σ^2, μ), we have $U_i = Z_i - Z_{p(i)}$, whereas, under the **OU** model $(z_0, \sigma^2, \alpha, \theta)$, we set $U_i = Z_i - e^{-\alpha\tau_i} Z_{p(i)}$.

Then the vector U is Gaussian with density:

$$f_U(u) = \frac{1}{(2\pi)^{n/2}(\det\Sigma_U)^{1/2}} e^{-\frac{1}{2}{}^t(u-m_U)\Sigma_U^{-1}(u-m_U)},$$

where m_U is the expectation vector of U and Σ_U its variance-covariance matrix which is diagonal since the coordinates of U are independent. For all $i \in \{1, \ldots, n\}$, we have $\mathbb{E}(U_i) = \mu\tau_i$ and $\mathrm{var}(U_i) = \sigma^2\tau_i$ under an **ABM** model, whereas $\mathbb{E}(U_i) = \theta(1 - e^{-\alpha\tau_i})$ and $\mathrm{var}(U_i) = \sigma^2 \frac{1-e^{-2\alpha\tau_i}}{2\alpha}$ under an **OU** model.

In order to compute the joint law of the nodes, i.e. the law of the random vector $Z = {}^t(Z_1, \cdots, Z_n)$, we remark that the vector Z is obtained from a linear transformation of U:

$$Z = CU + z_0\mathbf{w},$$

where C is a non-singular matrix and, for all non-root nodes i, $\mathbf{w}_i = 1$ under an **ABM** model and $\mathbf{w}_i = e^{-\alpha T_i}$ under an **OU** model. Since all its coordinates are affine combinations of the increments U_j, which are independent Gaussian variables, the vector Z is still a Gaussian vector with density:

$$f_Z(z) = \frac{1}{(2\pi)^{n/2}(\det\Sigma_Z)^{1/2}} e^{-\frac{1}{2}{}^t(z-m_Z)\Sigma_Z^{-1}(z-m_Z)},$$

where m_Z is the expectation vector of Z and Σ_Z its variance-covariance matrix, namely

$$\begin{aligned} m_Z &= Cm_U + z_0\mathbf{w} \quad \text{and} \\ \Sigma_Z &= C\Sigma_U{}^tC. \end{aligned}$$

The matrix C is lower triangular (thanks to the nodes numbering) with diagonal entries all equal to 1 and other entries $C_{i,j}$ with $i > j$ either equal to $\gamma_{i,j} \neq 0$ if nodes i and j belong to a same lineage or

equal to 0 otherwise, where $\gamma_{i,j} = 1$ under an **ABM** model, and $\gamma_{i,j} = e^{-\alpha(T_i+T_j-2T_{m(i,j)})}$ under an **OU** model.

Thus, for an **ABM** model we have

$$\begin{aligned} \mathbb{E}(Z_i) &= \mu T_i + z_0 \quad \text{and} \\ (\Sigma_Z)_{i,j} &= \mathrm{cov}(Z_i, Z_j) = \sigma^2 T_{m(i,j)}. \end{aligned} \tag{4}$$

In particular, for a **BM** model, we have

$$\begin{aligned} \mathbb{E}(Z_i) &= z_0 \quad \text{and} \\ (\Sigma_Z)_{i,j} &= \mathrm{cov}(Z_i, Z_j) = \sigma^2 T_{m(i,j)}. \end{aligned} \tag{5}$$

For an **OU** model, we have

$$\begin{aligned} \mathbb{E}(Z_i) &= \theta(1 - e^{-\alpha T_i}) + z_0 e^{-\alpha T_i} \quad \text{and} \\ (\Sigma_Z)_{i,j} &= \mathrm{cov}(Z_i, Z_j) = \sigma^2 \frac{1 - e^{-2\alpha T_{m(i,j)}}}{2\alpha} e^{-\alpha(T_i+T_j-2T_{m(i,j)})}. \end{aligned} \tag{6}$$

In particular, for an **OU*** model, we have

$$\begin{aligned} \mathbb{E}(Z_i) &= z_0 = \theta \quad \text{and} \\ (\Sigma_Z)_{i,j} &= \mathrm{cov}(Z_i, Z_j) = \sigma^2 \frac{1 - e^{-2\alpha T_{m(i,j)}}}{2\alpha} e^{-\alpha(T_i+T_j-2T_{m(i,j)})}. \end{aligned} \tag{7}$$

Since **ABM**, and **OU** models lead to Gaussian processes, the state distributions of all the nodes are multivariate normal. Let us compute the conditional joint law of the internal nodes given the leaf states $z^{|l}$, namely the law of $Y = {}^t(Y_1, \cdots, Y_r) = (Z^{|a} \,|Z^{|l} = z^{|l}\,)$. Let m_a be the expectation vector of $Z^{|a}$ and m_l that of $Z^{|l}$. Since the vector Z is a linear combination of the independent Gaussian increments U_i, then any density $f_Z, f_{Z^{|l}}$ or $f_{(Z^{|a}\,|Z^{|l}=z^{|l}\,)}$ is multivariate Gaussian.

The variance-covariance matrix Σ_Z of Z can be split according to $Z^{|a}$ and $Z^{|l}$:

$$\Sigma_Z = \begin{pmatrix} \Sigma_{a,a} & \Sigma_{a,l} \\ \Sigma_{l,a} & \Sigma_{l,l} \end{pmatrix}$$

where $\Sigma_{a,a}$ is the variance-covariance matrix of $Z^{|a}$, $\Sigma_{a,l}$ is the covariance matrix between $Z^{|a}$ and $Z^{|l}$ and so on.

With the decomposition of matrix Σ_Z, the random vector Y is Gaussian with density $\mathcal{N}(m_Y, \Sigma_Y)$, where

$$\begin{aligned} m_Y &= m_a + \Sigma_{a,l}\Sigma_{l,l}^{-1}(z_l - m_l), \\ \Sigma_Y &= \Sigma_{a,a} - \Sigma_{a,l}\Sigma_{l,l}^{-1}\Sigma_{l,a} \text{ (Schur complement)}. \end{aligned} \tag{8}$$

Remark 4. *There exist two matrices M_Z and M_Y such that*

$$\Sigma_Z = \sigma^2 M_Z \text{ and } \Sigma_Y = \sigma^2 M_Y$$

where M_Z and M_Y depend only

- *on the tree in the **BM** and **ABM** cases,*
- *on the tree and on the parameter α in the **OU** and **OU*** cases.*

Proof. From its definition, the remark holds for the matrix Σ_U. It straightforwardly follows that it holds for Σ_Z, thus for both $\Sigma_{a,a}$, $\Sigma_{a,l}$, $\Sigma_{l,l}$,$\Sigma_{l,a}$ and finally for Σ_Y. □

The expression of the conditional joint law of the internal nodes given the leaf states can be computed for various models such as the "Hansen model", which is an Ornstein-Uhlenbeck process with multiple evolutionary optima (Hansen 1997; Butler and King 2004). Models where parameter σ can depend on time can be handled as well (early-burst processes in Harmon et al. (2010)). The computation of the conditional joint law of the nodes only requires the state vector Z to be Gaussian, its expectation vector m_Z and its variance-covariance matrix Σ_Z. We restrict here the analysis to **ABM** and **OU** models homogeneous both in time and in lineage.

In order to compute the conditional law of an ancestral state i given the leaves, we have to sum over all the other internal states. Since the marginals of a Gaussian vector are still Gaussian, we eventually get that $Y_i = (Z_i | Z_{r+1} = z_{r+1}, \cdots, Z_n = z_n)$ follows an univariate Gaussian distribution $\mathcal{N}(y_i^T, \sigma_i^2)$, where y_i^T is the i^{th} coordinate of vector m_Y and σ_i^2 is the i^{th} diagonal entry of the variance-covariance matrix Σ_Y.

Below, we use the ancestral distribution of the ancestral state i given the leaves as reference when comparing with its reconstructions. This conditional distribution is referred to as the *theoretical distribution* of state i. We put Y_i^T for the corresponding random variable.

3.2 Evaluation protocol

3.2.1 For reconstructed states – Absolute bias

In a simulation context, the relevance of a reconstructed state is generally assessed by measuring its distance from the corresponding simulated ancestral state (Butler and Losos 1997; Martins 1999; Oakley and Cunningham 2000; Webster and Purvis 2002). This distance accounts for the reconstruction error. Remark that, in order to make sense, the distances between the reconstructed and simulated states have to be averaged over a large number of simulations.

The theoretical distributions derived in the previous section may be used to improve the assessment of reconstructed states. Let us consider an evolution z simulated under a given model. A reconstruction method only takes into account the leaf states $z_{r+1}, \cdots, z_n$. On the other hand, Section 3.1 gives us the conditional distribution of an ancestral state i given $z_{r+1}, \cdots, z_n$ under the simulation model. Intuitively, this distribution would be asymptotically observed by running an infinite number of simulations and by keeping only those with leaf states $z_{r+1}, \cdots, z_n$. The distance expectation between Y_i^T and the reconstructed state is exactly the conditional expectation of the reconstruction error on state i given the leaf states.

This suggests to replace the standard evaluation procedure by the following protocol. Being given an evolution model and a distance d,

1. simulate an evolution z under the model;
2. retain only the leaf states $z_{r+1}, \cdots, z_n$;
3. for all nodes i, compute from $z_{r+1}, \cdots, z_n$:
   - the reconstructed state y_i^R,
   - the conditional distribution of state i given $z_{r+1}, \cdots, z_n$ under the simulation model (i.e. the distribution of Y_i^T);
4. for all nodes i, compute $\mathbb{E}\left(\mathrm{d}(y_i^R, Y_i^T)\right)$.

This protocol ensures that the leaf states are well sampled from their probability distribution under the simulation model. It follows that averaging $\mathbb{E}\left(\mathrm{d}(y_i^R, Y_i^T)\right)$, which is conditioned on the leaf states, over all the simulations do converge to the expected reconstruction error on state i under the simulation model.

In the standard evaluation scheme, the distance d between a reconstructed state y_i^R and the corresponding simulated state y_i^S is generally measured in terms of bias $(y_i^S - y_i^R)$, absolute bias $|y_i^S - y_i^R|$ or squared bias $(y_i^S - y_i^R)^2$. Let us compute the expectations of these distances between a reconstructed

state and the random variable Y_i^T following the theoretical distribution $\mathcal{N}(y_i^T, \sigma_i^2)$. They are

$$\begin{aligned}
\mathbb{E}\left(Y_i^T - y_i^R\right) &= y_i^T - y_i^R, \\
\mathbb{E}\left(|Y_i^T - y_i^R|\right) &= \sigma_i \sqrt{\frac{2}{\pi}} e^{-\frac{(y_i^T - y_i^R)^2}{2\sigma_i^2}} + |y_i^T - y_i^R| \mathbb{P}\left(|W| \leq \frac{|y_i^T - y_i^R|}{\sigma_i}\right) \quad \text{and} \\
\mathbb{E}\left((Y_i^T - y_i^R)^2\right) &= (y_i^T - y_i^R)^2 + \sigma_i^2.
\end{aligned} \tag{9}$$

where W stands for the standard Gaussian variable. Although all these measures are suitable to compare the random variable Y_i^T with the reconstructed state y_i^R, they do not take into account the same amount of information from the distribution of Y_i^T. The bias is only based on its mean, the squared bias uses its mean and variance while the absolute bias takes into account both its mean, variance and the normality of the distribution. This point somehow supports the choice of this last distance.

3.2.2 For reconstructed distributions – Energy distance

Assessing the relevance of the uncertainty distributions provided by the reconstruction methods could also be done by considering the simulated ancestral states. But one expects more efficiency by considering the theoretical distributions. Adapting the above protocol to this case could be done by considering, for all nodes i, the expectation $\mathbb{E}\left(\mathrm{d}(Y_i^R, Y_i^T)\right)$, where Y_i^R follows the distribution provided by the method for i, which is also conditioned on the leaf states. A major drawback here is that the above expectation is not a good measure of the similarity between two probability distributions. In particular, it is not equal to zero when Y_i^R and Y_i^T are identically distributed (and not degenerate).

On the other hand, there exist various distances for comparing two probability distributions. Among them, the so-called Energy distance is strongly related to the evaluation protocol when d is the absolute bias. We will see that it offers us a consistent framework to compare states versus states, states versus distributions and distributions versus distributions.

Let A and B be two random variables and F_A and F_B their respective cumulative distribution functions. For convenience reasons, we write the distance between two distributions as the distance between two random variables following them. There are two equivalent ways to define the *Energy distance* (E-distance) between A and B (Szekely and Rizzo 2013):

$$\mathrm{d}_{\mathrm{NRG}}(A, B) = 2\|F_A - F_B\|_{\mathcal{L}^2}^2 = 2\int_{-\infty}^{\infty} |F_A(x) - F_B(x)|^2 \mathrm{d}x \tag{10}$$

and

$$\mathrm{d}_{\mathrm{NRG}}(A, B) = 2\mathbb{E}(|A - B|) - \mathbb{E}(|A - A'|) - \mathbb{E}(|B - B'|), \tag{11}$$

where A' and B' are independent and identically distributed copies of A and B respectively.

A distance between distributions can be used for comparing a single value against a distribution (or even two single values), just by considering the degenerate distribution(s) at the single value(s).

Let us start by checking the behavior of the E-distance when comparing two single values. Assuming that A and B follow degenerate distributions at a and b respectively, we have that

$$\mathrm{d}_{\mathrm{NRG}}(A, B) = 2|a - b|.$$

In plain English, the E-distance between two degenerate distributions is twice the absolute bias between the corresponding values.

Now if A follows the degenerate distribution at a and B follows a Gaussian distribution with variance σ_B^2, we have to compute the expected absolute value of Gaussian variables, whose formula is recalled in Appendix D, Equation D1. Thus the E-distance becomes:

$$\begin{aligned}
\mathrm{d}_{\mathrm{NRG}}(A, B) &= 2\mathbb{E}\left(|a - B|\right) - \mathbb{E}\left(|B - B'|\right) \\
&= 2\mathbb{E}\left(|a - B|\right) - \frac{2\sigma_B}{\sqrt{\pi}}.
\end{aligned} \tag{12}$$

The E-distance between a reconstructed state and the corresponding theoretical distribution is twice the expectation of the absolute bias between the state and a random variable following the theoretical

distribution, minus a correcting term. Up to this term, averaging the E-distances between the reconstructed states and the theoretical distributions is the same as applying the evolution protocol with the absolute bias.

Finally, Equation 11 shows that the Energy distance between two random variables following general distributions is twice the expectation of the absolute bias between them, minus two terms which somehow accounts for their respective dispersion.

In conclusion, the E-distance is strongly related to the protocol of Section 3.2.1 when evaluating reconstructed states or reconstructed distributions with the absolute bias.

The protocol eventually used in our comparisons is that of Section 3.2.1 with the 4^{th} step replaced by

4. for all nodes i, compute $\mathrm{d}_{\mathrm{NRG}}(X, Y_i^T)$.

where X is either the degenerate random variable of the reconstructed state at i or a random variable following the uncertainty distribution provided by the method under evaluation.

We show how to compute the E-distance of pairs of distributions involved in a theoretical vs reconstructed distributions comparison, i.e. Gaussian versus degenerate, Gaussian versus Gaussian and Gaussian versus Student, in Appendix D (R-scripts available on request).

The more usual Kolmogorov-Smirnov distance is actually harder to interpret when degenerate distributions are involved. In particular the Kolmogorov-Smirnov distance between two degenerate distributions is always 1 except if they are equal (Suppementary material). Supplementary Figures S5, S6, S7 and S8 display the Kolmogorov-Smirnov distances. Their general behavior is the same as with the Energy distance (Figures S1, S2, S3 and S4).

3.3 Optimal reconstruction

Let us assume that the character follows an **ABM** model (z_0, σ^2, μ) or an **OU** model $(z_0, \sigma^2, \alpha, \theta)$, including the **OU*** model when $z_0 = \theta$. Being given a set of leaf states and under the model, reconstructing the state of i with the mean y_i^T of its theoretical distribution leads to the smallest expectation error in terms of any standard distance (Equation 9) and of E-distance (Equation 12). The argument is similar as that of (Steel and Szekely 1999). Namely, reconstructing with y_i^T leads to expectations of bias, absolute bias and squared bias equal to 0, $\sigma_i\sqrt{\frac{2}{\pi}}$ and σ_i^2 respectively, and to E-distance $2\sigma_i\frac{\sqrt{2}-1}{\sqrt{\pi}}$. The mean y_i^T will be referred to as the *optimal reconstruction* of the state i. Remark that computing y_i^T requires to know the parameters of the model of evolution. The optimal reconstruction can be determined in a simulation context but unfortunately not in a practical situation.

In the particular case of a **BM** model, the optimal reconstruction is strongly related to the state reconstructed by *ML/REML/GLS_BM*. Indeed, let us consider a **BM** starting at the grand mean $\hat{z}_0$, given by the first formula of Equation A5 in Appendix A. By considering Equation 4 with $\mu = 0$, the partial mean vectors m_a and m_l are equal to $\hat{z}_0\mathbf{1}^a$ and $\hat{z}_0\mathbf{1}^l$ respectively. It follows that the second equation of A5, which gives the ancestral reconstructed states, and the first equation of 8, which gives the conditional means, are identical. In short, if *ML/REML/GLS_BM* infers the "real" state of the root, it reconstructs the whole tree in an optimal way.

4 Results and discussion

4.1 Simulation protocol

In order to assess reconstruction methods performance, we simulate the evolution of a quantitative character along the Pleistocene planktic Foraminifera phylogenetic tree (Webster and Purvis 2002), given in Figure 2. Though we implemented *GLS_OU* reconstruction method, we do not present its performances here. In Cooper et al. (2016), the authors caution against the estimation of the root value under **OU**. Indeed, as shown in Figure 1, under an **OU** model, the root state can be estimated only if the observed states are not all sampled from the constrained regime, which occurs in very peculiar conditions, unlikely to arise in practice. We did observe that the method *GLS_OU* may lead to very inaccurate reconstructions of the ancestral states, notably for deep nodes (Tables S1 and S2 of the Supplementary Material). Under an **OU** model, a small error in estimating the parameter α has a non linear effect, and huge consequences, on the reconstruction. We do not consider neither the **OU** model,

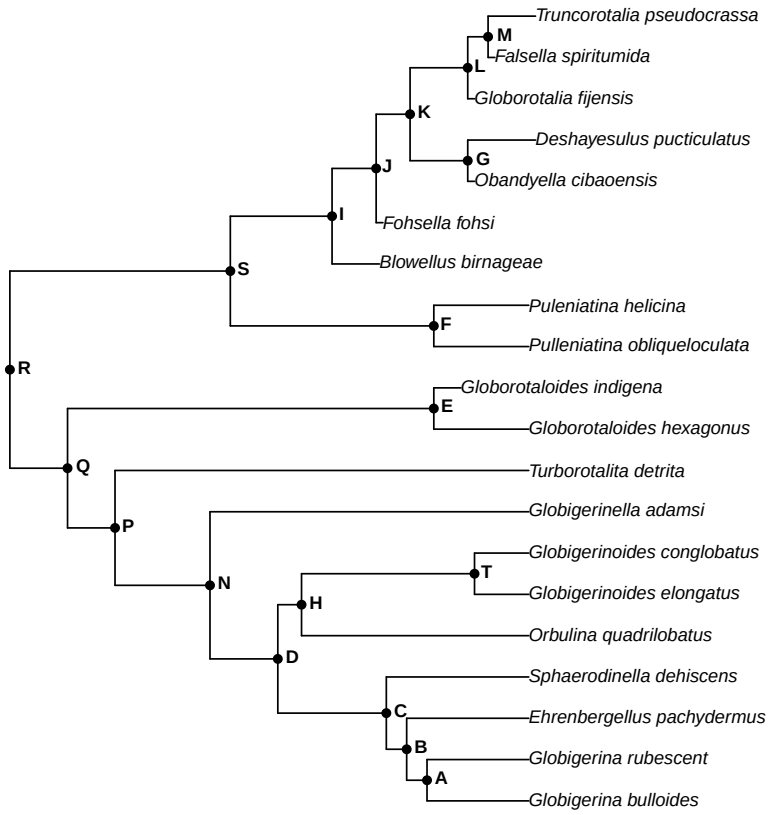

Figure 2: Pleistocene planktic Foraminifera phylogeny (Webster and Purvis 2002) on which the simulations runs.

nor *GLS_OU* in our evaluation. Thus we assess the methods *PIC*, *SP*, *ML*, *REML*, *GLS_BM*, *GLS_ABM* and *GLS_OU**, on data simulated under **ABM** and **OU*** models, including **BM**.

Although the simulation models of evolution depend on the root state z_0, we keep z_0 fixed since, in **ABM** and **OU*** models, it just translates the whole process and does not influence the methods performance. In order to assess their accuracy, we simulate the evolution of a quantitative character along the Pleistocene planktic Foraminifera phylogenetic tree (Webster and Purvis 2002), given in Figure 2, which starts from $z_0 = 100$ at the root and evolves either under **ABM** models with various trends and variances (21 values for parameter μ, ranging from -10 to 10 and 15 values for parameter σ, ranging from 0.01 to 20), or under **OU*** models with an optimum θ set to z_0 and various selection strengths α and variances (20 values for parameter α, ranging from 0 to 0.5 and 2 values for parameter σ equal to 3 and 10.

For each parameter set, we run 500 simulations from which we retain only the leaf states. We apply the evaluation protocol of Section 3.2 on the reconstructed states and on the reconstructed distributions provided by *PIC* from the function `ace`, and by *ML*, *REML* and *SP* from our function `reconstruct` of the `ape` R-package. We use our own R-script to compute *GLS_BM*, *GLS_ABM* and *GLS_OU** ancestral state reconstructions and their distributions. The theoretical distribution of each ancestral state under the simulation model is then compared first, with the reconstructed states and second, with the corresponding reconstructed distributions, in both cases in terms of E-distance. These distances are finally averaged over all the simulations in order to compare the performances of the methods.

4.2 Single reconstructed states

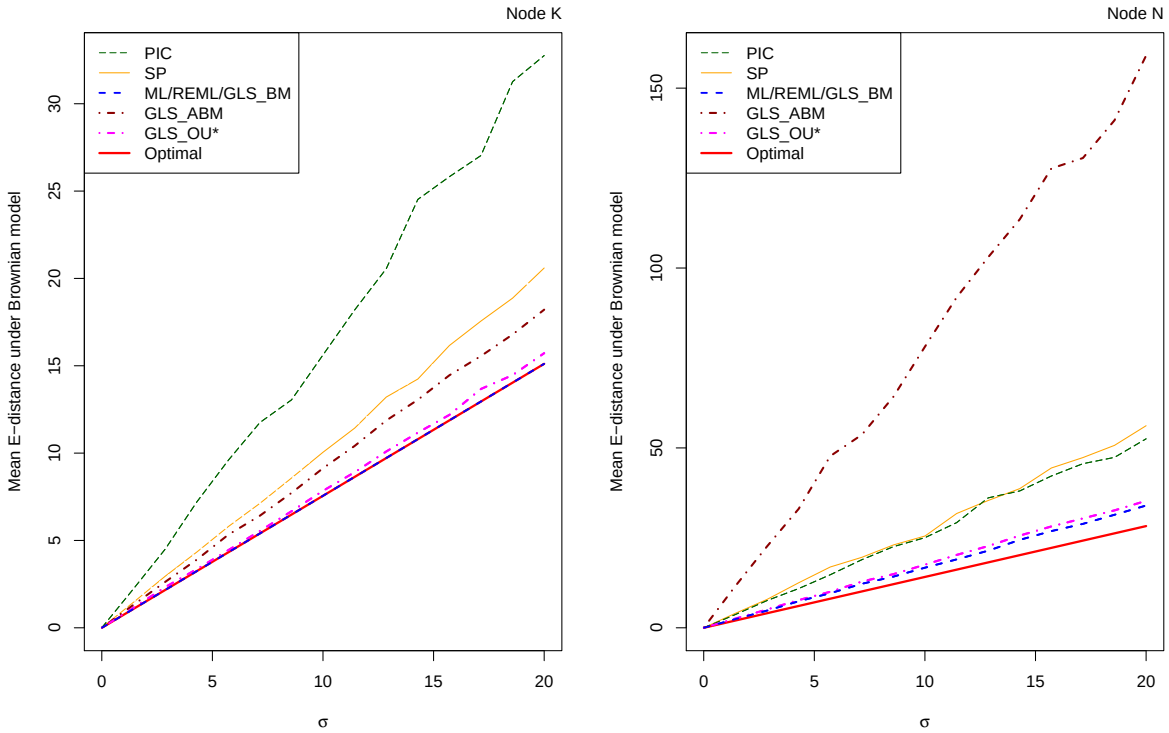

Figure 3: Mean Energy distance between the state reconstructed by *PIC*, *ML*/*REML*/*GLS_BM*, *SP*, *GLS_ABM*, *GLS_OU** and the corresponding theoretical distribution from **BM** models *versus* the parameter σ for the nodes B and N.

We first evaluate the methods accuracy with regard to the ancestral states they provide. Since *ML*, *REML* and *GLS_BM* return the same inferred states, they have the same accuracy which is compared with that of *PIC*, *SP*, *GLS_ABM*, *GLS_OU** and with the optimal one. We recall that, due to its inaccurate

reconstructions for deep nodes (Tables S1 and S2), *GLS_OU* reconstruction method is discarded.

4.2.1 Lower bounds of reconstruction errors

Since – except in degenerate cases – there is no unique ancestral states configuration leading to the leaf states from which we infer, reconstructing an ancestral state with a single value always comes with a certain probability of error. A counterpart of this fact is that the E-distance between a reconstructed state and its theoretical distribution is positive. From Section 3.3, the smallest E-distance which can be obtained with a reconstructed state y_i^R comes by setting y_i^R equal to the mean of the theoretical distribution Y_i^T. It leads to a E-distance $2\sigma_i \frac{\sqrt{2}-1}{\sqrt{\pi}}$ with $\sigma_i = \sigma M_{i,i}$, where $M_{i,i}$ only depends on the tree for **ABM** models, whatever the trend μ and the root state z_0; whereas for **OU*** models, $M_{i,i}$ depends on the tree and on the selective strength α, but not on the root state z_0 (Remark 4). To sum up, being given the parameters of the model, the optimal reconstruction provides a lower bound for the E-distances which does not depend on the initial value z_0. For **ABM** simulations, the optimal reconstruction depends linearly on the variance of the model but not on its trend, whereas for **OU*** simulations, it depends both on the variance and on the selective strength of the model, in a non linear way. It is represented by red lines in Figures 3, 4 and 5.

Supplementary Figures S1 and S3 display the plots of Figures 4 and 5 for all the nodes of the tree.

Asking whether a method achieves the optimal reconstruction, at least when the character follows a **BM** model, is a natural question. *ML/REML/GLS_BM* sounds like a good candidate for that, since it reconstructs the optimal state for any node as soon as the inferred root state matches z_0 under a **BM** model (Section 3.3). Thus in Figures 5 or S3 (it is less obvious on Figures 4 or S1), we clearly observe that, under a **BM** model, the states inferred by *ML/REML/GLS_BM* are almost indistinguishable from the optimal ones for any node except for the deepest nodes (N, P, Q, R and S). The two situations are shown in Figure 3, which displays the results of nodes K and N. The fact that *ML/REML/GLS_BM* are not optimal for some nodes always comes from an inaccurate estimation of the root state. Remark that despite this inaccurate root estimation, *ML/REML/GLS_BM* may still be almost optimal for the nodes close to the tips.

4.2.2 Influence of the simulation parameters

The smaller the parameter σ of a **BM** model, the more accurate the reconstructions of all the methods (Figure 3). Basically, as σ decreases, all the states of the tree (both ancestral and tips) get closer to one another, which makes the reconstruction easier. Still in Figure 3, we observe that *GLS_ABM* is more sensitive to the variance of the data than the other model-based approaches, in particular for the deepest nodes.

Another general observation is that, under an **ABM** model, all the Brownian-based methods perform better as the trend μ is close to 0 (Figures 4). This was expected since this situation is close to a **BM** model which is the assumption underlying all the methods but *SP*, *GLS_ABM* and *GLS_OU**. Only *GLS_ABM*, which is based on the **ABM** model, deals well when the μ is far from 0.

As shown in Figure 5, reconstruction methods behavior is much less intuitive for simulations under an **OU*** model. Above a certain level, the greater the strength selection α, the more accurate the reconstructions of all the methods, including the optimal reconstruction. This is not actually surprising with regard to the properties of the **OU*** model (Equation 7). Thus, for large α, evolving far from the optimum (here z_0) becomes unlikely. All the states stand in a very small range, which makes the ancestral reconstruction easy. When the strength selection is large, *GLS_OU* overcomes the methods which are not based on the Ornstein-Uhlenbeck process but to a lesser extent with regard to what is observed betwwen *GLS_ABM* and the non directional methods under **ABM** models with large trends.

4.2.3 Methods comparison

The methods performances are very close to one another for some of the nodes. This is basically the case for the root for which *ML/REML/GLS_BM* and *PIC* infer the same reconstructed state. But Figures 4 and S1 show that under **ABM** simulations, all the methods, except *GLS_ABM*, have nearly the same E-distance for several nodes. For this reason, we add error bars representing 95%-confidence intervals for the mean E-distance in Figures 4 and 5. Whenever the error bars do not overlap, the

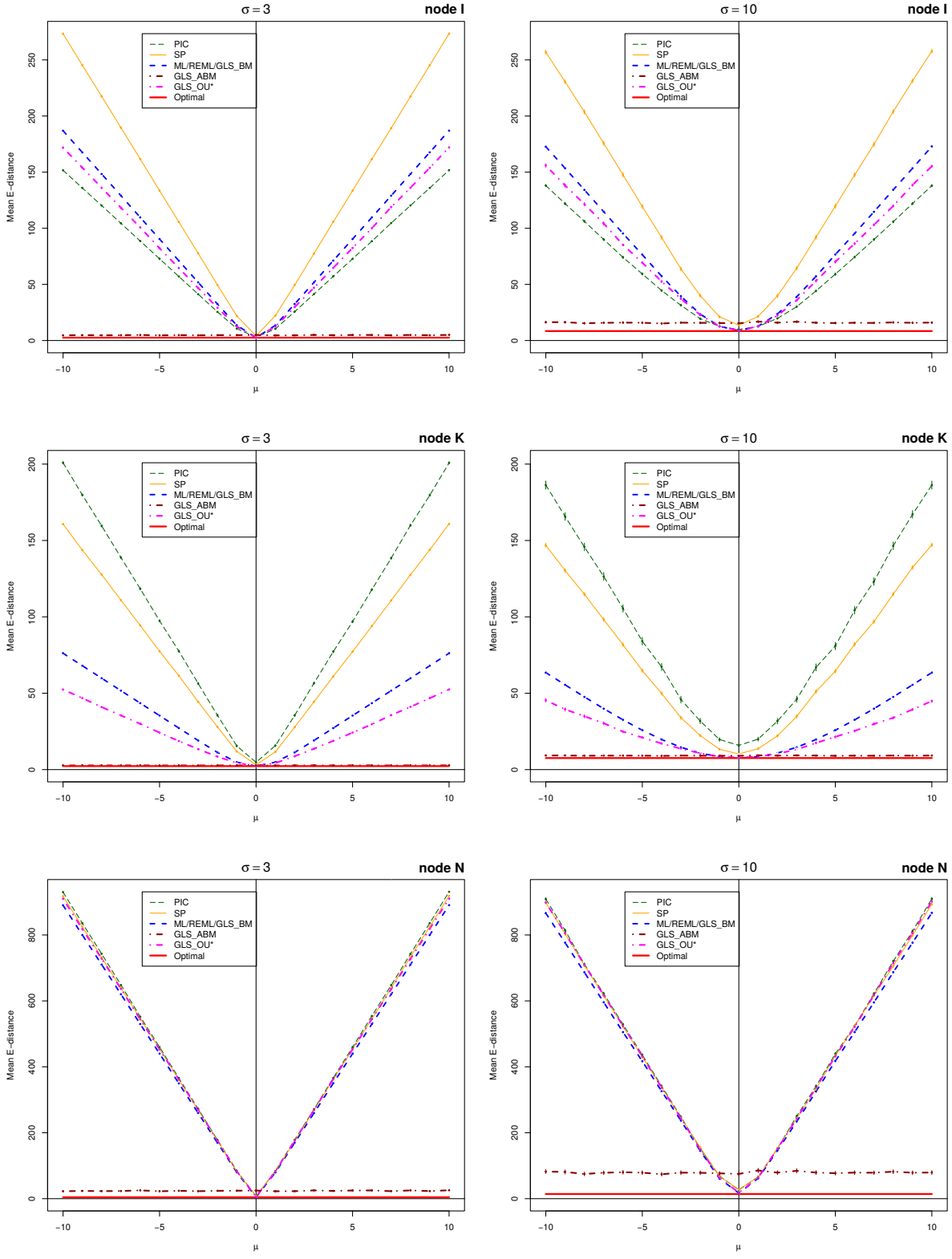

Figure 4: Mean Energy distance between the state reconstructed by *PIC*, *ML/REML/GLS_BM*, *GLS_ABM*, *GLS_OU**, *SP* and the corresponding theoretical distribution from **ABM** models with σ equal to 3 and 10 *versus* the parameter μ for nodes I, K and N.

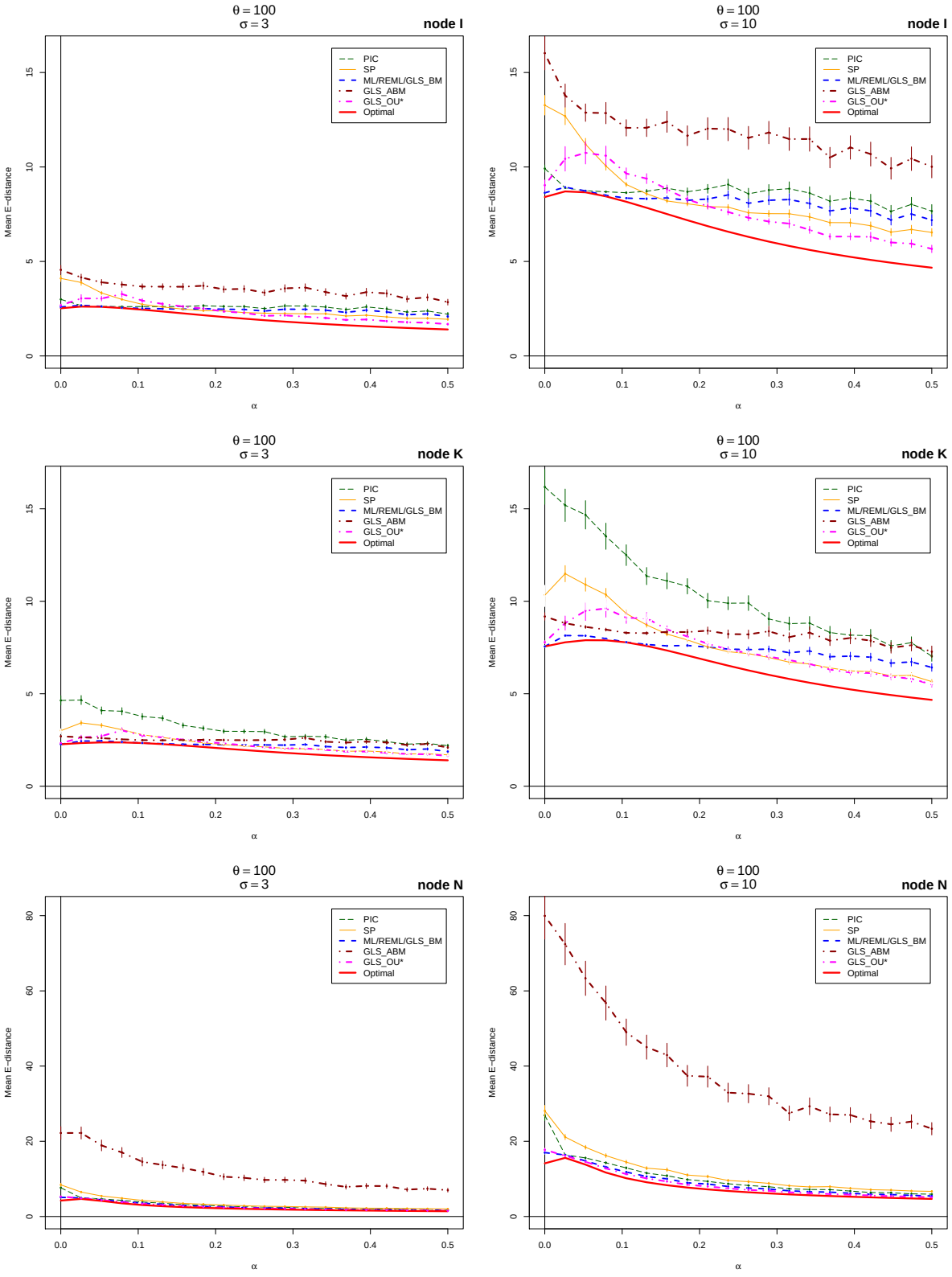

Figure 5: Mean Energy distance between the state reconstructed by *PIC*, *ML/REML/GLS_BM*, *GLS_ABM*, *GLS_OU**, *SP* and the corresponding theoretical distribution from **OU** models with $\theta = 200$, σ equal to 3 and 10 *versus* the parameter α for nodes I, K and N.

method corresponding to the lower curve has a significantly better performance than that of the upper one, according to the Student's t-test for paired samples with $\alpha = 5\%$. Such intervals permit to compare the methods accuracy.

Under a **BM** model, corresponding to data simulated with $\mu = 0$ or with $\alpha = 0$, *ML*/*REML*/*GLS_BM* provides the most accurate reconstruction for the nodes displayed in Figures 3, 4 and 5. This is actually observed for all the nodes of the tree (Figures S1 and S3). Still under a **BM** model, it is surprising to observe that *GLS_OU** is very accurate, quite close to *ML*/*REML*/*GLS_BM* and always better than *PIC* and *SP*. This certainly comes from the fact that when α tends to 0, the variance-covariance under the **OU*** model (Equation 7) converges to that of the **BM** model (Equation 5). Since the reconstructed states are given by the system A5 (Appendix A) for both *GLS_BM* and *GLS_OU**, if the variance-covariance structure of the two models are close then so are the corresponding reconstructions. We did check that *GLS_OU** did estimate small selection strengths α on data simulated under a **BM** model.

In Figure 4 or S1, we compare the methods on data simulated under "real" **ABM** models (i.e. with a significant trend). As soon as the trend is large enough, *GLS_ABM* outperforms all the other methods. It even reaches the optimal bound for nodes G, J, K, L and M. A common feature of these nodes is that they all have descendants at unequal distances from them (Figure 2), which are known to be essential for estimating the trend. When the trend is weak with regard to the variance of the model used for simulating the data, *GLS_ABM* is overcome by the other methods. We observe that its performance does depend on the variance of the **ABM** model used for simulating the data but not on its trend.

Under an **OU*** model, Figures 5 and S3 show that, as expected, *GLS_OU** tends to be the best reconstruction method, when the selective strength is large enough, except for nodes L and M, where it is overcome by *ML*/*REML*/*GLS_BM* and *GLS_ABM*.

4.3 Reconstructed distributions

Let us evaluate the relevance of the distributions provided by the methods for the reconstruction uncertainty still under **ABM** and **OU*** simulations. Due to its inaccurate estimations (Tables S1 and S2), *GLS_OU* reconstruction method is still discarded. The methods are now assessed with the E-distance between the theoretical distribution and the reconstructed one which is either Gaussian (*PIC*, *GLS_BM*, *GLS_ABM* and *GLS_OU**), Student (*ML* and *REML*), or degenerate (*SP*, *GLS_BM*, *GLS_ABM* and *GLS_OU** at the root). As some of the methods could possibly provide a reconstructed distribution matching exactly the theoretical one, we no longer have a lower positive bound of these E-distances (the counterparts of the "optimal" red lines of Figures 4, 5, S1 and S3 are the abscissa axis in Figures 6, 7, S2 and S4). Since they are Student instead of Gaussian, the distributions provided by *ML* and *REML* can not perfectly match the theoretical ones.

Considering reconstructed distributions rather than single values is expected to change the E-distances. The most notable difference is that *ML*, *REML* and *GLS_BM* are no longer equivalent and can now be compared one another as well as with *PIC*, *SP*, *GLS_ABM* and *GLS_OU**. Note that the E-distances are the very same in the case of *SP*, since this method only provides a reconstructed state.

In **ABM** simulations, the distributions provided by *ML*, *REML* and *GLS_BM* do lead to lower E-distances than those obtained from the single reconstructed states. By contrast, there is no observable change for *PIC* and *GLS_OU** performances between Figures S1 and S2. The same holds for *GLS_ABM* except for the nodes G, J, K, L and M, where it is improved, being optimal in the two cases (i.e. it is close to the abscissa axis in Figure S2 and to the plot of the optimal reconstruction in Figure S1). For the other nodes, *GLS_ABM* remains the best reconstruction method as soon as the trend μ is large enough, whereas under a weak trend, *GLS_BM* and *ML* become the most accurate (Figure 6 or S2).

For simulations under **OU*** models, all the reconstruction methods but *SP* are significantly improved between Figures S3 and S4. *GLS_OU**, without being optimal, still has the best performances for all the nodes, as soon as the selective strength is large enough. On the contrary, it is generally overcome by *GLS_BM* when the selection strength is weak.

4.4 Discussion

As expected, we observed that Brownian-based methods perform better when the character evolution follows a **BM** model. In the same way, under an **ABM** model with a significant trend (μ large enough),

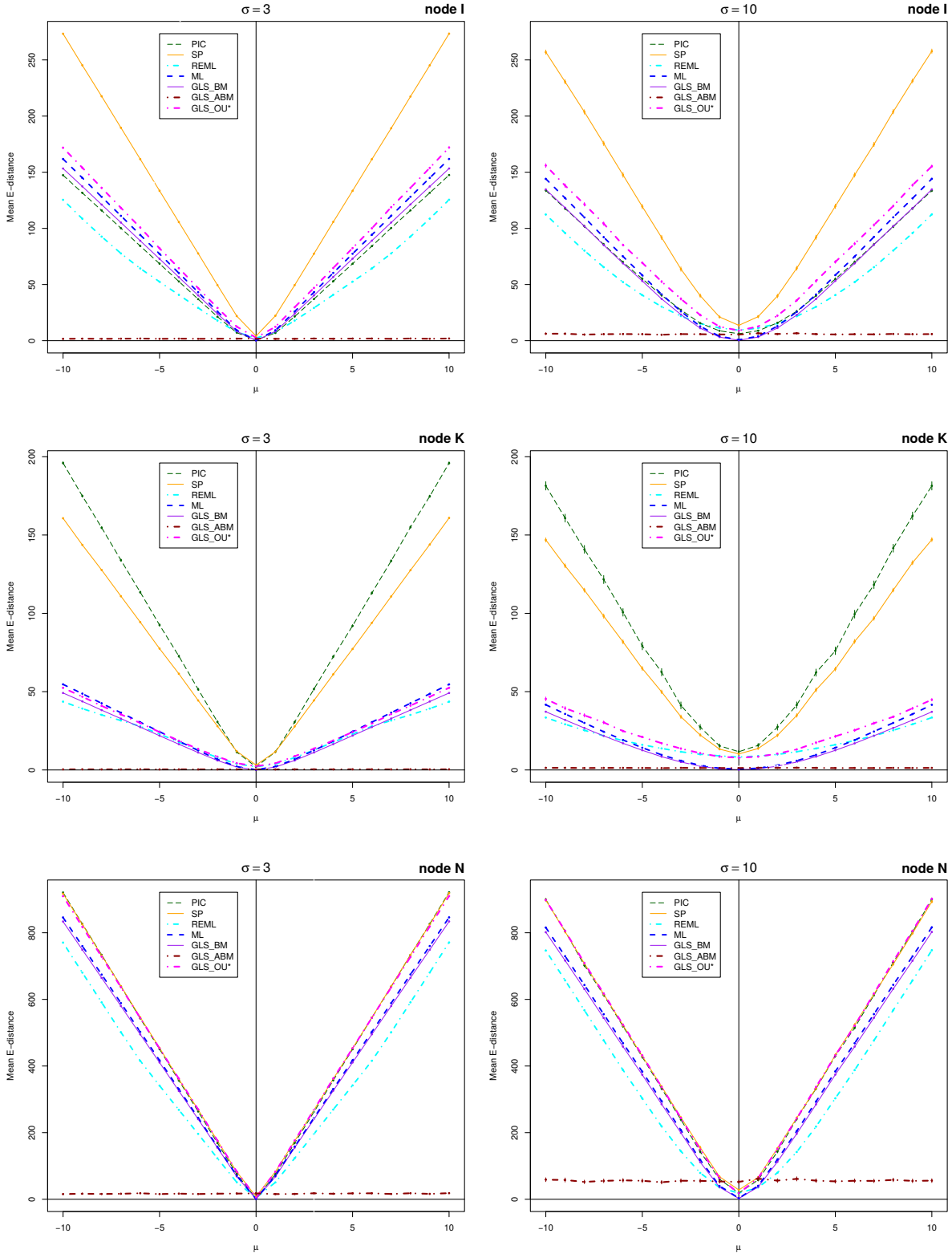

Figure 6: Mean Energy distance between the reconstructed distribution from *PIC*, *ML*, *REML*, *GLS_BM*, *GLS_ABM*, *GLS_OU**, *SP* and the corresponding theoretical distribution from **ABM** models with σ equal to 3 and 10 *versus* the parameter μ for the nodes I, K and N.

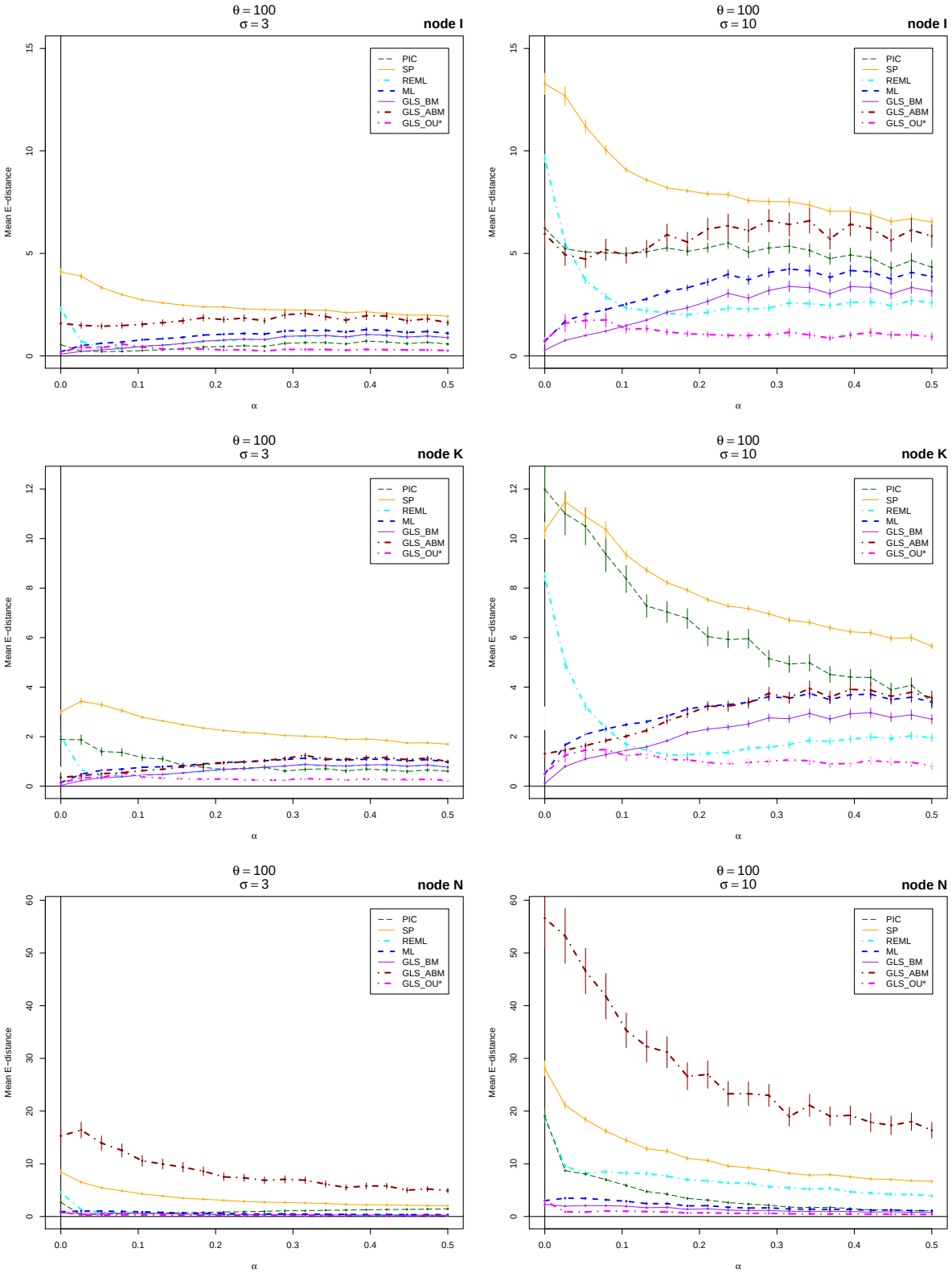

Figure 7: Mean Energy distance between the reconstructed distribution from *PIC*, *ML*, *REML*, *GLS_BM*, *GLS_ABM*, *GLS_OU**, *SP* and the corresponding theoretical distribution from **OU** models with $\theta = 200$, σ equal to 3 and 10 *versus* the parameter α for nodes I, K and N.

GLS_ABM provides the best reconstruction, whereas under an **OU*** model with a significant selection strength (α large enough), *GLS_OU** is generally the most accurate.

Under a **BM** model, the E-distance of all the methods increases linearly with the diffusion σ.Under an **ABM** model, the reconstruction accuracy decreases linearly with the trend, for all the methods except for *GLS_ABM*. Our simulations show that it is essential to use *GLS_ABM*, or at least a method adapted to directional evolution, as it was observed with fossil data in (Webster and Purvis 2002) where standard reconstruction methods led to spurious results. On the contrary, in the absence of trend, *GLS_ABM* may show poor performance. For characters evolving under stabilizing selection, our observations do not allow to draw such categorical conclusions. Though *GLS_OU** is the most accurate for large selective strengths, all the reconstruction methods show very good performances in this situation.

By construction, the theoretical distributions obtained from the simulation model reflect the real uncertainty of the character reconstruction. Thus one expects from the distributions provided by a reconstruction method to approach the theoretical ones or, at least, to be more informative than single reconstructed states. This is actually observed for most of the methods. This indicates that the reconstructed distributions provided by the reconstruction methods, *SP* excluded, may be relevant with regard to the inherent reconstruction uncertainty.

To summarize the results of our comparison, we first observed that *GLS_BM* and/or *ML* are the most accurate – or at least among the most accurate – any time the character evolution is close to the Brownian motion. As expected, we observed that *GLS_ABM* is the most accurate reconstruction method when the character follows an evolution with a trend. In the same way, *GLS_OU** is the most accurate reconstruction method if the character is under stabilizing selection. Overall, all the methods do deal well with stabilizing selection, but not with trend.

An important point is that an essential stage, prior to ancestral reconstruction, is to test the presence or the absence of a trend in the character evolution. This can be performed by model selection approaches such as likelihood ratio test, information criteria or Bayesian methods (Pagel 1998; Harmon et al. 2010). In contrast, detecting if the character is under stabilizing selection (i.e. better fit an **OU*** model than a **BM** model) appears to be less crucial since it rather eases the ancestral reconstruction for all the methods. Cooper et al. (2016) recently highlighted several issues in distinguishing between **OU*** and **BM** models.

Dealing with general **OU** models is more problematic since, from a short amount of data, it is difficult to distinguish between the linear trend of an **ABM** model and the transient "directional evolution" phase arising when an **OU** starts with an initial value far from its optimum. We do recommend to use the **OU** model with caution, notably for ancestral state reconstruction, since it may lead to infer aberrant values.

Acknowledgements

We thank Bastien Bousseau, Pierre Pontarotti and Laurence Reboul for their careful readings and their helpful remarks and suggestions. We also thank an anonymous referee for her/his comments and for indicating several references.

Funding

Centre National de la Recherche Scientifique (PEPS "Mission pour l'interdisciplinarité" *Évolution et génétique des populations* to G. Didier)

References

Butler, M. and King, A. (2004). Phylogenetic Comparative Analysis: A Modeling Approach for Adaptative Evolution. *Am. Nat.*, 164(6):683–695.

Butler, M. and Losos, J. (1997). Testing for unequal amounts of evolution in a continuous character on different branches of a phylogenetic tree using linear and squared-change parsimony: an example using Lesser Antillean Anolis lizards. *Evolution*, 51(5):1623–1635.

Butler, M., Schoener, A., and Losos, J. (2000). The relationship between sexual size dimorphism and habitat use in Greater Antillean *Anolis* lizards. *Evolution*, 54:259–272.

Collins, T., Wimberger, P., and Naylor, G. (1994). Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Syst. Biol.*, 43:482–496.

Cooper, N., Thomas, G. H., Venditti, C., Meade, A., and Freckleton, R. (2016). A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Journal of Finance*, 118:64–77.

Cunningham, C., Omland, K., and T.H., O. (1998). Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology and Evolution*, 13:361–366.

Felsenstein, J. (1973). Maximum-Likelihood Estimation of Evolutionary Trees from Continuous Characters. *Am. J. Hum. Genet.*, 25(5):471–492.

Felsenstein, J. (1985). Phylogenies and the comparative methods. *Am. Nat.*, 125:1–15.

Felsenstein, J. (1988). Phylogenies and quantitative characters. *Annu. Rev. Ecol. Syst.*, 19:445–471.

Fitch, W. (1971). Towards defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.*, 20:406–416.

Göing-Jaeschke, A. and Yor, M. (2003). A clarification note about hitting times densities for ornstein-uhlenbeck processes. *Finance Stoch.*, 8:413–415.

Grafen, A. (1989). The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 326(1233):119–157.

Hansen, T. (1997). Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51:1341–1351.

Hansen, T. and Martins, E. (1996). Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution*, 50:1404–1417.

Harmon, L., Losos, J., Davies, T., Gillepsie, R., Gittleman, J., Jennings, W., Kozak, K., McPeek, M., Moreno-Raork, F., Near, T., Purvis, A., Ricklefs, R., Schluter, D., Schulte II, J., Seehausen, O., Sidlauskas, B., Torres-Carvajal, O., Weir, J., and Mooers, A. (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, 64(8):2385–2396.

Hone, D. and Benton, M. (2005). The evolution of large size: how does Cope's Rule work? *Trends Ecol. Evol.*, 20(1):4–6.

Huelsenbeck, J., Nielsen, R., and Bollback, J. (2003). Stochastic mapping of morphological characters. *Syst. Biol.*, 52:131–158.

Huelsenbeck, J. and Ronquist, F. (2001). MrBayes: Bayesian inference of phylogeny. *Bioinformatics*, 8:754–755.

Kingsolver, J. and Pfennig, D. (2004). Individual-level selection as a cause of Cope's Rule of phylogenetic size increase. *Evolution*, 58:1608–1612.

Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. *Evolution*, 30:314–334.

Maddison, W. (1991). Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. *Syst. Zool.*, 40:304–314.

Maddison, W. and Maddison, D. (1992). MacClade Analysis of Phylogeny and Character Evolution. version 3. Technical report, Sinauer Associates, Inc., New York.

Martins, E. (1994). Estimating rates of character change from comparative data. *Am. Nat.*, 144:193–209.

Martins, E. (1995). COMPARE: statistical analysis for comparative data, version 1.0.

Martins, E. (1999). Estimation of Ancestral States of Continuous Characters: A Computer Simulation Study. *Syst. Biol.*, 48(3):642–650.

Martins, E. and Hansen, T. (1997). Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist*, 149:646–667.

Mooers, A. and Schluter, D. (1999). Reconstructing Ancestor States with Maximum Likelihood: Support for One- and Two-Rate Models. *Syst. Biol.*, 48(3):623–633.

Nielsen, R. (2002). Mapping mutations on phylogenies. *Syst. Biol.*, 51:729–739.

Oakley, T. H. and Cunningham, C. W. (2000). Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. *Evolution*, 54(2):397–405.

Pagel, M. (1998). Inferring evolutionary processes from phylogenies. *Zool. scr.*, 26:331–348.

Pagel, M. (1999a). Inferring the historical pattern of biological evolution. *Nature*, 401:877–884.

Pagel, M. (1999b). The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.*, 48:612–622.

Pagel, M. and Meade, A. (2013). Bayestraits v. 2.0. *Reading : University of Reading*.

Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of character states on phylogenies. *Syst. Biol.*, 57:673–684.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290.

Piessens, R., de Doncker-Kapenga, E., Uberhuber, C. W., and Kahaner, D. K. (1983). *Quadpack : A Subroutine Package for Automatic Integration*. Springer Series in Computational Mathematics. Springer Verlag, Berlin.

Polly, P. (2001). Paleontology and the Comparative Method: Ancestral Node Reconstructions versus Observed Node Values. *Am. Nat.*, 157(6):596–609.

Revell, L. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecol. and Evol.*, 3:217–223.

Royer-Carenzi, M., Pontarotti, P., and Didier, G. (2013). Choosing the best ancestral character state reconstruction method. *Mathematical Biosciences*, 242:95–109.

Schluter, D., Price, T., Mooers, A., and Ludwig, D. (1997). Likelihood of ancestor states in adaptative radiation. *Evolution*, 51(6):1699–1711.

Searle, R., Casella, G., and McCulloch, C. (1992). *Variance Components*. Series in Probability and Mathematical Statistics. Wiley, New York. Applied Probability and Statistics Section.

Steel, M. and Szekely, L. (1999). Inverting random functions. *Annals Combin*, 3:103–113.

Swofford, D. and Maddison, W. (1987). Reconstructing ancestral character state under Wagner parsimony. *Math. Biosci.*, 87:199–229.

Szekely, G. and Rizzo, M. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272.

Van Valkenburgh, B., Wang, X., and Damuth, J. (2004). Cope's Rule, Hypercarnivory, and extinction in North American Canids. 306(5693):101–104.

Vasicek, O. (1977). An equilibrium characterization of the term structure. 5(2):177–188.

Webster, A. and Purvis, A. (2002). Testing the accuracy for reconstructing ancestral states of continuous characters. *Proc. R. Soc. Lond. B*, 269:143–149.

A Equivalence between *GLS_BM* and *ML*

Though it can be shown through more general considerations, we provide here a detailed proof of the equivalence between the methods *GLS_BM* and *ML*. This proof is useful to show the relation between the optimal and the *GLS_BM*/*ML* reconstructions (Section 3.3). Let Z_i be the random variable of the node state i. We put Z, $Z^{|a}$ and $Z^{|l}$ for the random vectors ${}^t(Z_1, \cdots, Z_n)$, ${}^t(Z_1, \ldots, Z_r)$ and ${}^t(Z_{r+1}, \ldots, Z_n)$, corresponding to all the nodes except the root, the internal nodes excluding root, and the leaves, respectively. A set of node states $z_0, \ldots, z_n$ is organized as vectors z, $z^{|a}$ and $z^{|l}$ accordingly.

In order to explain why *GLS_BM* and *ML* are equivalent, we shall consider two different expressions of the likelihood under the assumptions of *ML*. In particular, *ML* assumes that the character evolution follows a **BM** model with variance σ^2. The probability density of a vector $z_0, \ldots, z_n$ can be written either

$$f_{(Z_0,Z)}(z_0, z) = f_{Z_0}(z_0) f_Z(z), \tag{A1}$$

or

$$f_{(Z_0,Z)}(z_0, z) = f_{Z_0}(z_0) f_{(Z^{|a}|Z^{|l}=z^{|l})}(z^{|a}) f_{Z^{|l}}(z^{|l}). \tag{A2}$$

Since the vector Z can be expressed as a linear transformation of the independent Gaussian increments $Z_j - Z_{p(j)}$, both $f_Z, f_{Z^{|l}}$ and $f_{(Z^{|a}|Z^{|l}=z^{|l})}$ are multivariate Gaussian densities. The variance-covariance matrix Σ_Z of Z can be split according to $Z^{|a}$ and $Z^{|l}$:

$$\Sigma_Z = \begin{pmatrix} \Sigma_{a,a} & \Sigma_{a,l} \\ \Sigma_{l,a} & \Sigma_{l,l} \end{pmatrix},$$

where $\Sigma_{a,a}$ is the variance-covariance matrix of $Z^{|a}$, $\Sigma_{a,l}$ is the covariance matrix between $Z^{|a}$ and $Z^{|l}$ and so on. The matrix Σ_Z has the form $\sigma^2 M_Z$ where entry $M_{Z\,i,j}$ is the time between the root and the most recent common ancestor of nodes i and j (Felsenstein 1973). We put $\mathbf{1}$ (resp. $\mathbf{1}^{|a}$ and $\mathbf{1}^{|l}$) for the n-dimensional (resp. r- and $(n-r)$-dimensional) vector with all coordinates equal to 1. Since Z follows the multivariate normal distribution $\mathcal{N}(z_0\mathbf{1}, \Sigma_Z)$, the marginal and conditional random vectors $Z^{|l}$ and $(Z^{|a}|Z^{|l} = z^{|l})$ follow the multivariate normal distributions $\mathcal{N}(z_0\mathbf{1}^{|l}, \Sigma_{l,l})$ and $\mathcal{N}(\tilde{z}, \tilde{\Sigma}_{a,a})$ respectively, where

$$\begin{aligned} \tilde{z} &= z_0\mathbf{1}^{|a} + \Sigma_{a,l}\Sigma_{l,l}^{-1}(z^{|l} - z_0\mathbf{1}^{|l}) \text{ and} \\ \tilde{\Sigma}_{a,a} &= \Sigma_{a,a} - \Sigma_{a,l}\Sigma_{l,l}^{-1}\Sigma_{l,a} \text{ (Schur complement).} \end{aligned}$$

Under the *ML* assumptions, f_{Z_0} is the improper flat density, thus its logarithm just vanishes in the computation of $\log(f_{(Z_0,Z)}(z_0, z))$ with Equations A1 and A2. On the one hand, from Equation A2, the vector of partial derivatives of $\log(f_{(Z_0,Z)}(z_0, z))$ with respect to the internal states $z^{|a}$ is proportional to the vector

$$z^{|a} - z_0\mathbf{1}^{|a} - \Sigma_{a,l}\Sigma_{l,l}^{-1}(z^{|l} - z_0\mathbf{1}^{|l}).$$

On the other hand, from Equation A1, the partial derivative of $\log(f_{(Z_0,Z)}(z_0, z))$ with respect to the root state z_0 is proportional to

$$z_0\,{}^t\mathbf{1}\Sigma_Z^{-1}\mathbf{1} - {}^t\mathbf{1}\Sigma_Z^{-1}z. \tag{A3}$$

The maximum likelihood estimates $\hat{z}_0, \hat{z}_1, \ldots, \hat{z}_r$ of internal states with respect to the vector leaf states $z^{|l}$ may basically be obtained by solving the system of linear equations:

$$\begin{aligned} z_0\,{}^t\mathbf{1}\Sigma_Z^{-1}\mathbf{1} - {}^t\mathbf{1}\Sigma_Z^{-1}\begin{pmatrix} z^{|a} \\ z^{|l} \end{pmatrix} &= 0 \\ z^{|a} - z_0\mathbf{1}^{|a} - \Sigma_{a,l}\Sigma_{l,l}^{-1}(z^{|l} - z_0\mathbf{1}^{|l}) &= \mathbf{1}^{|a}\ . \end{aligned} \tag{A4}$$

Let us get a simpler form for the first equation of A4. The inversion formula for block matrices gives us that

$$\Sigma_Z^{-1} = \begin{pmatrix} \tilde{\Sigma}_{a,a}^{-1} & -\tilde{\Sigma}_{a,a}^{-1}\Sigma_{a,l}\Sigma_{l,l}^{-1} \\ -\Sigma_{l,l}^{-1}\Sigma_{l,a}\tilde{\Sigma}_{a,a}^{-1} & \Sigma_{l,l}^{-1} + \Sigma_{l,l}^{-1}\Sigma_{l,a}\tilde{\Sigma}_{a,a}^{-1}\Sigma_{a,l}\Sigma_{l,l}^{-1} \end{pmatrix}.$$

It follows that Expression A3 can be rewritten as

$$\begin{aligned}
&\left({}^t\mathbf{1}^{|a}\,\tilde{\Sigma}^{-1}_{a,a}\mathbf{1}^{|a} - {}^t\mathbf{1}^{|a}\,\tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l}\Sigma^{-1}_{l,l}\mathbf{1}^{|l} - {}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}\Sigma_{l,a}\tilde{\Sigma}^{-1}_{a,a}\mathbf{1}^{|a}\right) z_0\\
&+\left({}^t\mathbf{1}^{|l}\,(\Sigma^{-1}_{l,l}+\Sigma^{-1}_{l,l}\Sigma_{l,a}\tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l}\Sigma^{-1}_{l,l})\mathbf{1}^{|l}\right) z_0\\
&-({}^t\mathbf{1}^{|a}\,\tilde{\Sigma}^{-1}_{a,a} - {}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}\Sigma_{l,a}\tilde{\Sigma}^{-1}_{a,a})z^{|a}\\
&-\left({}^t\mathbf{1}^{|l}\,(\Sigma^{-1}_{l,l}+\Sigma^{-1}_{l,l}\Sigma_{l,a}\tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l}\Sigma^{-1}_{l,l}) - {}^t\mathbf{1}^{|a}\,\tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l}\Sigma^{-1}_{l,l}\right) z^{|l},
\end{aligned}$$

in which, substituting $z^{|a}$ according to the second equation of A4, leads to

$$z_0\,{}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}\mathbf{1}^{|l} - {}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}z^{|l}.$$

Finally, the estimates $\hat{z}_0$ and $\hat{z}^{|a}$ maximizing the log-likelihood with respect to the vector of leaf states $z^{|l}$ satisfy:

$$\begin{aligned}
\hat{z}_0 &= ({}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}\mathbf{1}^{|l})^{-1}\,{}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}z^{|l},\\
\hat{z}^{|a} &= \hat{z}_0\mathbf{1}^{|a} + \Sigma_{a,l}\Sigma^{-1}_{l,l}(z^{|l} - \hat{z}_0\mathbf{1}^{|l}).
\end{aligned} \tag{A5}$$

These formula are the same as those computing the GLS reconstruction (Martins and Hansen 1997; Cunningham et al. 1998; Martins 1999, or see below) in which $\hat{z}_0$ is called the *grand mean*.

B General least squares estimation under the ABM model

Let us recall that T denotes the vector of distances from the root to all the nodes but the root (i.e. for all nodes $i \neq 0$, T_i is the total length of the path from 0 to i). We put $T^{|a}$ (resp. $T^{|l}$) for the vector of entries of T corresponding to the internal nodes (resp. to the leaves) of the tree.

Under an **ABM** model with trend μ, the residual sum of square (RSS) has the form:

$$\text{RSS} = {}^t(z - z_0\mathbf{1} - \mu T)\Sigma_Z^{-1}(z - z_0\mathbf{1} - \mu T),$$

where Σ_Z is the variance covariance matrix of the states under an **BM** or an **ABM** model, given in Equations 4.

By developing the expression above and by using the block inversion formula, it follows that

$$\begin{aligned}
\text{RSS} &= {}^tz\Sigma_Z^{-1}z - 2z_0({}^t\mathbf{1}\Sigma_Z^{-1}z) - 2\mu({}^tT\Sigma_Z^{-1}z) + z_0^2({}^t\mathbf{1}\Sigma_Z^{-1}\mathbf{1}) + \mu^2({}^tT\Sigma_Z^{-1}T) + 2z_0\mu({}^t\mathbf{1}\Sigma_Z^{-1}T)\\
&= {}^tz^{|a}\,\tilde{\Sigma}^{-1}_{a,a}z^{|a} - 2{}^tz^{|l}\,\Sigma^{-1}_{l,l}\Sigma_{l,a}\tilde{\Sigma}^{-1}_{a,a}z^{|a} + {}^tz^{|l}\,(\Sigma^{-1}_{l,l}+\Sigma^{-1}_{l,l}\Sigma_{l,a}\tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l}\Sigma^{-1}_{l,l})z^{|l}\\
&\quad -2z_0\left[({}^t\mathbf{1}^{|a} - {}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}\Sigma_{l,a})\tilde{\Sigma}^{-1}_{a,a}z^{|a} + ({}^t\mathbf{1}^{|l} + ({}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}\Sigma_{l,a} - {}^t\mathbf{1}^{|a})\tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l})\Sigma^{-1}_{l,l}z^{|l}\right]\\
&\quad -2\mu\left[({}^tT^{|a} - {}^tT^{|l}\,\Sigma^{-1}_{l,l}\Sigma_{l,a})\tilde{\Sigma}^{-1}_{a,a}z^{|a} + ({}^tT^{|l} + ({}^tT^{|l}\,\Sigma^{-1}_{l,l}\Sigma_{l,a} - {}^tT^{|a})\tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l})\Sigma^{-1}_{l,l})z^{|l}\right]\\
&\quad +z_0^2({}^t\mathbf{1}\Sigma_Z^{-1}\mathbf{1}) + \mu^2({}^tT\Sigma_Z^{-1}T) + 2z_0\mu({}^t\mathbf{1}\Sigma_Z^{-1}T).
\end{aligned}$$

Let us set

$$\Delta_{\mathbf{1}} = \mathbf{1}^{|a} - \Sigma_{a,l}\Sigma^{-1}_{l,l}\mathbf{1}^{|l} \text{ and } \Delta_T = T^{|a} - \Sigma_{a,l}\Sigma^{-1}_{l,l}T^{|l}.$$

The partial derivative of RSS with regard to the trend is

$$\frac{\partial\text{RSS}}{\partial\mu} = 2\mu({}^tT\Sigma_Z^{-1}T) + 2z_0({}^t\mathbf{1}\Sigma_Z^{-1}T) - 2({}^t\Delta_T\tilde{\Sigma}^{-1}_{a,a}z^{|a} + ({}^tT^{|l} - {}^t\Delta_T\tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l})\Sigma^{-1}_{l,l}z^{|l}).$$

The partial derivative of RSS with regard to the root state is

$$\frac{\partial\text{RSS}}{\partial z_0} = 2z_0({}^t\mathbf{1}\Sigma_Z^{-1}\mathbf{1}) + 2\mu({}^t\mathbf{1}\Sigma_Z^{-1}T) - 2({}^t\Delta_{\mathbf{1}}\tilde{\Sigma}^{-1}_{a,a}z^{|a} + ({}^t\mathbf{1}^{|l} - {}^t\Delta_{\mathbf{1}}\tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l})\Sigma^{-1}_{l,l}z^{|l}).$$

The vector of partial derivatives of RSS with regard to the internal states is

$$\frac{\partial \text{RSS}}{\partial z^{|a}} = 2\tilde{\Sigma}^{-1}_{a,a} z^{|a} - 2\tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l}\Sigma^{-1}_{l,l} z^{|l} - 2z_0\tilde{\Sigma}^{-1}_{a,a}\Delta_{\mathbf{1}} - 2\mu\tilde{\Sigma}^{-1}_{a,a}\Delta_T.$$

It follows that the vector $\begin{pmatrix} \mu \\ z_0 \\ z^{|a} \end{pmatrix}$ minimizing RSS satisfies:

$$\begin{pmatrix} {}^tT\Sigma_Z^{-1}T & {}^t\mathbf{1}\Sigma_Z^{-1}T & -{}^t\Delta_T\tilde{\Sigma}^{-1}_{a,a} \\ {}^t\mathbf{1}\Sigma_Z^{-1}T & {}^t\mathbf{1}\Sigma_Z^{-1}\mathbf{1} & -{}^t\Delta_{\mathbf{1}}\tilde{\Sigma}^{-1}_{a,a} \\ -\tilde{\Sigma}^{-1}_{a,a}\Delta_T & -\tilde{\Sigma}^{-1}_{a,a}\Delta_{\mathbf{1}} & \tilde{\Sigma}^{-1}_{a,a} \end{pmatrix} \begin{pmatrix} \mu \\ z_0 \\ z^{|a} \end{pmatrix} = \begin{pmatrix} ({}^tT^{|l} - {}^t\Delta_T\tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l})\Sigma^{-1}_{l,l} z^{|l} \\ ({}^t\mathbf{1}^{|l} - {}^t\Delta_{\mathbf{1}}\tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l})\Sigma^{-1}_{l,l} z^{|l} \\ \tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l}\Sigma^{-1}_{l,l} z^{|l} \end{pmatrix}. \quad \text{(B1)}$$

Remark that if one assumes a null trend (i.e. under a **BM** model), the system B1 becomes

$$\begin{pmatrix} {}^t\mathbf{1}\Sigma_Z^{-1}\mathbf{1} & -{}^t\Delta_{\mathbf{1}}\tilde{\Sigma}^{-1}_{a,a} \\ -\tilde{\Sigma}^{-1}_{a,a}\Delta_{\mathbf{1}} & \tilde{\Sigma}^{-1}_{a,a} \end{pmatrix} \begin{pmatrix} z_0 \\ z^{|a} \end{pmatrix} = \begin{pmatrix} ({}^t\mathbf{1}^{|l} - {}^t\Delta_{\mathbf{1}}\tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l})\Sigma^{-1}_{l,l} z^{|l} \\ \tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l}\Sigma^{-1}_{l,l} z^{|l} \end{pmatrix},$$

which leads to Equations A5.

Let us go back to the system B1. By substituting

$$\tilde{\Sigma}^{-1}_{a,a} z^{|a} = \mu\tilde{\Sigma}^{-1}_{a,a}\Delta_T + z_0\tilde{\Sigma}^{-1}_{a,a}\Delta_{\mathbf{1}} + \tilde{\Sigma}^{-1}_{a,a}\Sigma_{a,l}\Sigma^{-1}_{l,l} z^{|l}$$

in the first two equations of B1, we get that

$$\mu({}^tT\Sigma_Z^{-1}T - {}^t\Delta_T\tilde{\Sigma}^{-1}_{a,a}\Delta_T) + z_0({}^t\mathbf{1}\Sigma_Z^{-1}T - {}^t\Delta_T\tilde{\Sigma}^{-1}_{a,a}\Delta_{\mathbf{1}}) = {}^tT^{|l}\,\Sigma^{-1}_{l,l} z^{|l},$$
$$\mu({}^t\mathbf{1}\Sigma_Z^{-1}T - {}^t\Delta_{\mathbf{1}}\tilde{\Sigma}^{-1}_{a,a}\Delta_T) + z_0({}^t\mathbf{1}\Sigma_Z^{-1}\mathbf{1} - {}^t\Delta_{\mathbf{1}}\tilde{\Sigma}^{-1}_{a,a}\Delta_{\mathbf{1}}) = {}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l} z^{|l},$$

which becomes

$$\mu\,{}^tT^{|l}\,\Sigma^{-1}_{l,l}T^{|l} + z_0\,{}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}T^{|l} = {}^tT^{|l}\,\Sigma^{-1}_{l,l} z^{|l},$$
$$\mu\,{}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}T^{|l} + z_0\,{}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}\mathbf{1}^{|l} = {}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l} z^{|l}.$$

Let us remark that if there exists a real number λ such that $T^{|l} = \lambda\mathbf{1}^{|l}$ - in plain English, if all the leaves are at the same distance from the root - then the two equations above are linearly dependent. In this case, neither the trend nor the root state are identifiable from the leaf states.

If there is no such λ, the system can be explicitly solved to get

$$\begin{aligned} \hat{\mu} &= \frac{{}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}\mathbf{1}^{|l}\,.\,{}^tT^{|l}\,\Sigma^{-1}_{l,l} z^{|l} - {}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}T^{|l}\,.\,{}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l} z^{|l}}{{}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}\mathbf{1}^{|l}\,.\,{}^tT^{|l}\,\Sigma^{-1}_{l,l}T^{|l} - ({}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}T^{|l})^2}, \\ \hat{z}_0 &= \frac{{}^tT^{|l}\,\Sigma^{-1}_{l,l}T^{|l}\,.\,{}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l} z^{|l} - {}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}T^{|l}\,.\,{}^tT^{|l}\,\Sigma^{-1}_{l,l} z^{|l}}{{}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}\mathbf{1}^{|l}\,.\,{}^tT^{|l}\,\Sigma^{-1}_{l,l}T^{|l} - ({}^t\mathbf{1}^{|l}\,\Sigma^{-1}_{l,l}T^{|l})^2}, \\ \hat{z}^{|a} &= \hat{\mu}\Delta_T + \hat{z}_0\Delta_{\mathbf{1}} + \Sigma_{a,l}\Sigma^{-1}_{l,l} z^{|l}. \end{aligned} \quad \text{(B2)}$$

It can be shown that, under the **ABM** model, the maximum likelihood estimates of the parameters and of the internal states are the same as those given by Equations B2.

C General least squares estimation under the OU and the OU* models

Let us put $\mathbf{w}$ for the vector such that, for all nodes $i \neq 0$, $\mathbf{w}_i = e^{-\alpha T_i}$. We put $\mathbf{w}^{|a}$ (resp. $\mathbf{w}^{|l}$) for the vector of entries of $\mathbf{w}$ corresponding to the internal nodes (resp. to the leaves).

Under an **OU** model with parameters α and θ, the residual sum of square (RSS) has the form:

$$\text{RSS} = {}^t(z - z_0\mathbf{w} - \theta(\mathbf{1} - \mathbf{w}))\Sigma_Z^{-1}(z - z_0\mathbf{w} - \theta(\mathbf{1} - \mathbf{w})),$$

where Σ_Z is here the variance covariance matrix of the states under an **OU** model, given in Equations 6.

The RSS under an **OU** model has the same general form as under en **ABM** model and, assuming the parameter α known, leads to a linear system which can be explicitly solved to get:

$$\text{(C1)}\qquad \begin{aligned}\hat{\theta} &= \frac{{}^t\mathbf{w}^{|l}\Sigma_{l,l}^{-1}\mathbf{w}^{|l}\,.{}^t(\mathbf{1}^{|l}-\mathbf{w}^{|l})\Sigma_{l,l}^{-1}z^{|l}-{}^t\mathbf{w}^{|l}\Sigma_{l,l}^{-1}(\mathbf{1}^{|l}-\mathbf{w}^{|l}).{}^t\mathbf{w}^{|l}\Sigma_{l,l}^{-1}z^{|l}}{{}^t\mathbf{w}^{|l}\Sigma_{l,l}^{-1}\mathbf{w}^{|l}\,.{}^t(\mathbf{1}^{|l}-\mathbf{w}^{|l})\Sigma_{l,l}^{-1}(\mathbf{1}^{|l}-\mathbf{w}^{|l})-({}^t\mathbf{w}^{|l}\Sigma_{l,l}^{-1}(\mathbf{1}^{|l}-\mathbf{w}^{|l}))^2},\\ \hat{z}_0 &= \frac{{}^t(\mathbf{1}^{|l}-\mathbf{w}^{|l})\Sigma_{l,l}^{-1}(\mathbf{1}^{|l}-\mathbf{w}^{|l}).{}^t\mathbf{w}^{|l}\Sigma_{l,l}^{-1}z^{|l}-{}^t\mathbf{w}^{|l}\Sigma_{l,l}^{-1}(\mathbf{1}^{|l}-\mathbf{w}^{|l}).{}^t(\mathbf{1}^{|l}-\mathbf{w}^{|l})\Sigma_{l,l}^{-1}z^{|l}}{{}^t\mathbf{w}^{|l}\Sigma_{l,l}^{-1}\mathbf{w}^{|l}\,.{}^t(\mathbf{1}^{|l}-\mathbf{w}^{|l})\Sigma_{l,l}^{-1}(\mathbf{1}^{|l}-\mathbf{w}^{|l})-({}^t\mathbf{w}^{|l}\Sigma_{l,l}^{-1}(\mathbf{1}^{|l}-\mathbf{w}^{|l}))^2},\\ \hat{z}^{|a} &= \hat{\theta}(\mathbf{1}^{|a}-\mathbf{w}^{|a}-\Sigma_{a,l}\Sigma_{l,l}^{-1}(\mathbf{1}^{|l}-\mathbf{w}^{|l}))+\hat{z}_0(\mathbf{w}^{|a}-\Sigma_{a,l}\Sigma_{l,l}^{-1}\mathbf{w}^{|l})+\Sigma_{a,l}\Sigma_{l,l}^{-1}z^{|l}\,.\end{aligned}$$

Remark that we don't provide an expression of the parameter α, of which both vector $\mathbf{w}$ and matrix Σ_Z depend. This parameter is estimated by numerical optimization.

If one assumes that $\theta = z_0$ (i.e. under the model **OU***), the RSS becomes

$$\text{RSS} = {}^t(z - z_0\mathbf{1})\Sigma_Z^{-1}(z - z_0\mathbf{1}),$$

and leads to the system A5 with the variance-covariance matrix of the **OU** model.

In both **OU** and **OU*** cases and being given the strength selection α, the ancestral states reconstructed from the general least squares approach are the same as those maximizing the (log-)likelihood (Equation 3).

D Energy Distance

In this section, we will use the two following properties :

Let W be a standard Gaussian random variable. Then for $\mu \in \mathbb{R}$ and $\sigma > 0$,

$$\mathbb{E}(|\sigma W + \mu|) = \sigma\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right) + |\mu|\mathbb{P}\left(|W| \leq \frac{|\mu|}{\sigma}\right). \tag{D1}$$

Let W be a Student random variable with $(r+1)$ degrees of freedom. Then for $\mu \in \mathbb{R}$ and $\sigma > 0$,

$$\mathbb{E}(|\sigma W + \mu|) = \frac{2\sigma}{\sqrt{\pi}}\frac{\sqrt{r+1}}{r}\frac{\Gamma(\frac{r+2}{2})}{\Gamma(\frac{r+1}{2})}\left(1+\frac{\mu^2}{(r+1)\sigma^2}\right)^{-\frac{r}{2}} + |\mu|\mathbb{P}\left(|W| \leq \frac{|\mu|}{\sigma}\right). \tag{D2}$$

Let A and B be two random variables and F_A and F_B their respective cumulative distributions.

- If both A and B follow degenerate distributions at a and b respectively, then

$$\mathrm{d}_{\mathrm{NRG}}(A, B) = 2|a - b|.$$

- If A follows a normal law $\mathcal{N}(\mu_A, \sigma_A^2)$ and B a degenerate distribution at b, then, from Equation D1, we have

$$\mathrm{d}_{\mathrm{NRG}}(A, B) = \sigma_A\frac{2}{\sqrt{\pi}}\left(\sqrt{2}\exp\left(-\frac{(\mu_A - b)^2}{2\sigma_A^2}\right) - 1\right) + 2|\mu_A - b|\mathbb{P}\left(|W| \leq \frac{|\mu_A - b|}{\sigma_A}\right).$$

  If $\mu_A = b$, the Energy distance is equal to $2\sigma_A\frac{\sqrt{2}-1}{\sqrt{\pi}}$ thus increases with σ_A. For any fixed $\sigma_A \neq 0$, the Energy distance goes to infinity as $|\mu_A - b|$ becomes larger while for any fixed distance $|\mu_A - b| \neq 0$, it goes from $2|\mu_A - b|$ to infinity as σ_A goes from 0 to infinity.

- Let us assume that A and B follow a degenerate distribution at a and a Student law $t_{r+1}(\mu_B, \sigma_B^2)$ respectively. From Equation D2, it is possible to compute directly $\mathbb{E}(|A - B|)$ and $\mathbb{E}(|A - A'|)$,

but not $\mathbb{E}(|B - B'|)$ because the difference between two independent Student variables is not a Student variable. We thus rely on Expression 10 of the Energy distance:

$$\mathrm{d}_{\mathrm{NRG}}(A, B) = 2 \int_{-\infty}^{\infty} |F_A(x) - F_B(x)|^2 \mathrm{d}x.$$

with $F_A(x) = \begin{cases} 1 & \text{if } a \leq x, \\ 0 & \text{otherwise.} \end{cases}$ and $F_B(x) = F_{W_S}\left(\frac{x - \mu_B}{\sigma_B}\right)$, where F_{W_S} is the Student cumulative distribution function with $r + 1$ degrees of freedom. The integral of 10 is approximated by $\int_\alpha^\beta |F_A(x) - F_B(x)|^2 \mathrm{d}x$, where $\alpha = \mu_B - q_{r+1}\sigma_B$, $\beta = \max(\mu_B + q_{r+1}\sigma_B, a)$ and q_{r+1} is the $0,9999683$-quantile for a Student variable with $r + 1$ degrees of freedom. The numerical computation of this last integral is performed by the Wynn's Epsilon algorithm (Piessens et al. 1983).

- If both A and B are Gaussian variables, we get from Equation D1 that

$$\mathrm{d}_{\mathrm{NRG}}(A, B) = \frac{2}{\sqrt{\pi}} \left(\sqrt{2(\sigma_A^2 + \sigma_B^2)} \exp\left(-\frac{(\mu_A - \mu_B)^2}{2(\sigma_A^2 + \sigma_B^2)} \right) - (\sigma_A^2 + \sigma_B^2) \right) \quad \text{(D3)}$$

$$+2|\mu_A - \mu_B| \mathbb{P}\left(|W| \leq \frac{|\mu_A - \mu_B|}{\sqrt{\sigma_A^2 + \sigma_B^2}} \right). \quad \text{(D4)}$$

- Let us finally assume that A and B follow a normal distribution $\mathcal{N}(\mu_A, \sigma_A^2)$ and a Student distribution $t_{r+1}(\mu_B, \sigma_B^2)$ respectively. Hence again we have to rely on Expression 10. We approximate the integral with $\int_\alpha^\beta |F_A(x) - F_B(x)|^2 \mathrm{d}x$, where $\alpha = \min(\mu_A - 4\sigma_A, \mu_B - q_{r+1}\sigma_B)$, $\beta = \max(\mu_A + 4\sigma_A, \mu_B + q_{r+1}\sigma_B)$ and q_{r+1} (resp. 4) is the $0,9999683$-quantile for a Student variable with $r + 1$ degrees of freedom (resp. for a standard Gaussian variable). The integral is computed via the Wynn's Epsilon algorithm.