

Version dated: June 5, 2017

1

2 RH: INFERRING ADAPTIVE SHIFTS FOR MULTIVARIATE TRAITS

3

## Inference of Adaptive Shifts for Multivariate Correlated Traits

4

5 PAUL BASTIDE<sup>1,2</sup>, CÉCILE ANÉ<sup>3,4</sup>, STÉPHANE ROBIN<sup>1</sup> & MAHENDRA MARIADASSOU<sup>2</sup>

6

<sup>1</sup> *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France*

7

<sup>2</sup> *MaIAGE, INRA, Université Paris-Saclay, 78352 Jouy-en-Josas, France*

8

<sup>3</sup> *Department of Statistics, University of Wisconsin-Madison, WI, 53706, USA*

9

<sup>4</sup> *Department of Botany, University of Wisconsin-Madison, WI, 53706, USA*

10 **Corresponding author:** Paul Bastide, UMR MIA-Paris, AgroParisTech, INRA,  
11 Université Paris-Saclay, 16 rue Claude Bernard, 75005, Paris, France; E-mail:  
12 paul.bastide@m4x.org.

13

14 *Abstract.*— To study the evolution of several quantitative traits, the classical phylogenetic  
15 comparative framework consists of a multivariate random process running along the  
16 branches of a phylogenetic tree. The Ornstein-Uhlenbeck (OU) process is sometimes  
17 preferred to the simple Brownian Motion (BM) as it models stabilizing selection toward an  
18 optimum. The optimum for each trait is likely to be changing over the long periods of time  
19 spanned by large modern phylogenies. Our goal is to automatically detect the position of  
20 these shifts on a phylogenetic tree, while accounting for correlations between traits, which

21 might exist because of structural or evolutionary constraints. We show that, in the  
22 presence shifts, phylogenetic Principal Component Analysis (pPCA) fails to decorrelate  
23 traits efficiently, so that any method aiming at finding shift needs to deal with correlation  
24 simultaneously. We introduce here a simplification of the full multivariate OU model,  
25 named scalar OU (scOU), which allows for noncausal correlations and is still  
26 computationally tractable. We extend the equivalence between the OU and a BM on a  
27 re-scaled tree to our multivariate framework. We describe an Expectation Maximization  
28 algorithm that allows for a maximum likelihood estimation of the shift positions,  
29 associated with a new model selection criterion, accounting for the identifiability issues for  
30 the shift localization on the tree. The method, freely available as an R-package  
31 (PhylogeneticEM) is fast, and can deal with missing values. We demonstrate its efficiency  
32 and accuracy compared to another state-of-the-art method (*ℓ1ou*) on a wide range of  
33 simulated scenarios, and use this new framework to re-analyze recently gathered datasets  
34 on New World Monkeys and *Anolis* lizards.  
35 (Keywords: Ornstein-Uhlenbeck, Change-point detection, Adaptive evolution, Phylogeny,  
36 Model selection, PhylogeneticEM)

37

38

## *Motivation*

39

40

A major goal of comparative and evolutionary biology is to decipher the past evolutionary mechanisms that shaped the present day diversity. Taking advantage of the

41 increasing amount of molecular data made available by powerful sequencing techniques,  
42 sophisticated mathematical models have made it possible to infer reliable phylogenetic  
43 trees for ever growing groups of taxa (see e.g. Meredith et al. 2011; Jetz et al. 2012).  
44 Models of phenotypic evolution for such large families need to cope with the heterogeneity  
45 of observed traits across the species tree. One source of heterogeneity is the mechanism of  
46 “evolution by jumps” as hypothesized by Simpson (1944). It states that there exists an  
47 adaptive landscape shaping the evolution of functional traits, and that this landscape  
48 might shift, sometimes in a dramatic fashion, in response to environmental changes such as  
49 migration, or colonization of a new ecological niche. Such shifts, like the one observed in  
50 the brain shape and size of New World Monkeys in association with dietary and  
51 locomotion changes (Aristide et al. 2015, 2016), need to be explicitly accounted for in  
52 models of phenotypic evolution.

53 To detect such adaptive shifts, we must cope with two constraints: species do not  
54 evolve independently (Felsenstein 1985) and adaptive evolution is an intrinsically  
55 multivariate phenomenon. The first constraint arises from the shared evolutionary history  
56 of species, usually represented as a phylogenetic tree. It means that traits observed on  
57 closely related taxa are on average more similar than traits observed on distantly related  
58 species. The second constraint results from natural selection acting on many traits at once.  
59 Functional traits are indeed often interdependent, either because they are regulated by the  
60 same portions of the genetic architecture or because they are functionally constrained (e.g.  
61 limb bones lengths in Greater Antillean *Anolis* lizards Mahler et al. (2010)).

62 This work aims to develop a likelihood-based method to detect rapid adaptive  
63 events, referred to as shifts, using a time calibrated phylogenetic tree and potentially  
64 incomplete observations of a multivariate functional trait at the tips of that tree. The  
65 shifts can be used to cluster together species sharing a common adaptive history.

66

## *State of the Art*

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

Phylogenetic comparative methods (PCM) are the *de facto* tools for studying phenotypic evolution. Most of them can be summarized as stochastic processes on a tree. Specifically, given a rooted phylogeny, the traits evolve according to a stochastic process on each branch of the tree. At each speciation event, one independent copy with the same initial conditions is created for each daughter species. A common stochastic process in this setting is the Brownian Motion (BM, Felsenstein 1985). It is well suited to model the random drift of a quantitative, neutral and polygenic trait (see e.g. Felsenstein 2004, chap. 24). Unfortunately, the BM has no stationary distribution and cannot adequately model adaptation to a specific optimum (Hansen and Orzack 2005). The Ornstein-Uhlenbeck (OU) process is therefore preferred to the BM in the context of adaptive evolution (Hansen 1997; Hansen et al. 2008). Note that, as pointed out by Hansen et al. (2008) and Cooper et al. (2016), this model is distinct from the process theoretically derived by Lande (1976) for stabilizing selection toward an optimum on an adaptive landscape at a micro-evolutionary timescale, and is better seen as a heuristic for the macro-evolution of the “secondary optima” themselves in a Simpsonian interpretation of evolution (Hansen et al. 2008). Recently, Levy processes have also been used to capture Simpsonian evolution (Landis et al. 2013; Duchon et al. 2017).

Extensions to multivariate traits have been proposed for both BM (Felsenstein 1985) and OU processes (Bartoszek et al. 2012). Cybis et al. (2015) considered even more complex models, with a mix of both quantitative and discrete characters modeled with an underlying multivariate BM and a threshold model (Felsenstein 2005, 2012) for drawing discrete characters from the underlying continuous BM.

The work on adaptive shifts also enjoyed a growing interest in the last decade. In their seminal work, Butler and King (2004) considered a univariate trait with known shift locations on the tree and estimated shift amplitudes in the trait optimal value using a

92 maximum-likelihood framework. Beaulieu et al. (2012) extended the work by estimating  
93 shift amplitudes not only in the optimal value but also in the evolutionary rate. The focus  
94 then moved to estimating the number and locations of shifts. Eastman et al. (2011, 2013)  
95 detected shifts, respectively, in the evolutionary rate or the trait expectations, for traits  
96 evolving as BM, in a Bayesian setting using reversible jump Markov Chain Monte Carlo  
97 (rjMCMC). Ingram and Mahler (2013); Uyeda and Harmon (2014); Bastide et al. (2016)  
98 detected shifts in the optimal value of a trait evolving as an OU. Uyeda and Harmon  
99 (2014) and Bastide et al. (2016) detect all shifts for a given number of shifts and use either  
100 rjMCMC or penalized likelihood to select the number of shifts. By contrast, Ingram and  
101 Mahler (2013) uses a stepwise procedure, based on AIC, to detect shifts sequentially,  
102 stopping when adding a shift does not improve the criteria anymore.

103         Extensions from univariate to multivariate shifts are more recent. It should be  
104 noted that all methods assume that shifts affect all traits simultaneously. Given known  
105 shift locations and a multivariate OU process, Bartoszek et al. (2012) was the first to  
106 develop a likelihood-based method (package `mvSLOUCH`) to estimate both matrices of  
107 multivariate evolutionary rates and selection strengths. Clavel et al. (2015) soon followed  
108 with `mvmorph`, a comprehensive package covering a wide range of multivariate processes.  
109 Detection of shifts in multivariate traits is more involved and both Ingram and Mahler  
110 (2013) and Khabbazian et al. (2016) make the simplifying assumption that all traits are  
111 independent, conditional on their shared shifts. Ingram and Mahler (2013) then proceed  
112 with the same stepwise procedure as in the univariate case whereas Khabbazian et al.  
113 (2016) uses a lasso-regression to detect the shifts and a phylogenetic BIC (pBIC) criterion  
114 to select the number of shifts.

### 115 *Scope of the Article*

116         In this work, we present a new likelihood-based method to detect evolutionary shifts

117 in multivariate OU models. We make the simplifying assumptions that all traits have the  
118 same selection strength but, unlike in Khabbazian et al. (2016) and Ingram and Mahler  
119 (2013), traits can be correlated. Our contribution is multifaceted. We show that the scalar  
120 assumption that we make (see Section Model) and the independence assumption share a  
121 similar feature in their structure that make the shift detection problem tractable. Building  
122 upon a formal analysis made in the univariate case (Bastide et al. 2016), we show that the  
123 problem suffers from identifiability issues as two or more distinct shift configurations may  
124 be indistinguishable. We propose a latent variable model combined with an OU to BM  
125 reparametrization trick to estimate the unknown number of shifts and their locations. Our  
126 method is fast and can handle missing data. It also proved accurate in a large scale  
127 simulation study and was able to find back known shift locations in re-analysis of public  
128 datasets. Finally, we show that the standard practice of decorrelating traits using  
129 phylogenetic principal component analysis (pPCA) before using a method designed for  
130 independent traits can be misleading in the presence of shifts.

131 The article is organized as followed. We present the model and inference procedure  
132 in Section Model, the theoretical bias of pPCA in the presence of shifts in Section pPCA  
133 and Shifts, the simulation study in Section Simulations Studies, the re-analysis of the New  
134 World Monkeys and Greater Antillean *Anolis* lizards datasets in Section Examples and  
135 discuss the results and limitations of our method in Section Discussion.

## 136 MODEL

### 137 *Trait Evolution on a Tree*

138 *Tree.*— We consider a fixed and time-calibrated phylogenetic tree linking the present-day  
139 species studied. The tree is assumed ultrametric with height  $h$ , but with possible

140 polytomies. We denote by  $n$  the number of tips and by  $m$  the number of internal nodes,  
141 such that  $N = n + m$  is the total number of nodes. For a fully bifurcating tree,  $m = n - 1$ ,  
142 and  $N = 2n - 1$ .

143 *Traits.*— We note  $\mathbf{Y}$  the matrix of size  $n \times p$  of measured traits at the tips of the tree. For  
144 each tip  $i$ , the row-vector  $\mathbf{Y}^i$  represents the  $p$  measured traits at tip  $i$ . Some of the data  
145 might be missing, as discussed later (see Section Statistical Inference).

146 *Brownian Motion (BM).*— The multivariate BM has  $p + p(p + 1)/2$  parameters:  $p$  for the  
147 ancestral mean value vector  $\boldsymbol{\mu}$ , and  $p(p + 1)/2$  for the drift rate (in the genetic sense)  
148 matrix  $\mathbf{R}$ . The variance of a given trait grows linearly in time, and the covariance between  
149 two traits  $k$  and  $l$  at nodes  $i$  and  $j$  is given by  $t_{ij}R_{kl}$ , where  $t_{ij}$  is the time elapsed between  
150 the root and the most recent common ancestor (MRCA) of  $i$  and  $j$  (see e.g. Felsenstein  
151 2004, chap. 24). Using the vectorized version of matrix  $\mathbf{Y}$  (where  $\text{vec}(\mathbf{Y})$  is the vector  
152 obtained by “stacking” all the columns of  $\mathbf{Y}$ ), we get:  $\text{Var}[\text{vec}(\mathbf{Y})] = \mathbf{R} \otimes \mathbf{C}$ , where  $\otimes$  is  
153 the Kronecker product, and  $\mathbf{C} = [t_{ij}]_{1 \leq i, j \leq n}$ .

154 *Ornstein-Uhlenbeck (OU).*— The Ornstein-Uhlenbeck process has  $p^2$  extra parameters in  
155 the form of a selection strength matrix  $\mathbf{A}$ . The traits evolve according to the stochastic  
156 differential equation  $d\mathbf{X}_t = \mathbf{A}(\boldsymbol{\beta} - \mathbf{X}_t)dt + \mathbf{R}d\mathbf{W}_t$ , where  $\mathbf{W}_t$  stands for the standard  
157  $p$ -variate Brownian motion. The first part represents the attraction to a “primary  
158 optimum”  $\boldsymbol{\beta}$ , with a dynamic controlled by  $\mathbf{A}$ . This matrix is not necessarily symmetric in  
159 general, but it must have positive eigenvalues for the traits to indeed be attracted to their  
160 optima. This assumption also ensures the existence of a stationary state, with mean  $\boldsymbol{\beta}$  and  
161 variance  $\boldsymbol{\Gamma}$  (see Bartoszek et al. 2012; Clavel et al. 2015, for further details and general  
162 expression of  $\boldsymbol{\Gamma}$ ).

163 *Shifts.*— We assume that some environmental changes affected the traits evolution in the

164 past. In the BM model, we take those changes into account by allowing the process to be  
165 discontinuous, with shifts occurring in its mean value vector (as e.g. Eastman et al. 2013).  
166 This is reasonable if the adaptive response to a change in the environment is fast enough  
167 compared to the evolutionary time scale. For the OU, we assume that environmental  
168 changes result in a shift in the primary optimum  $\beta$  (as e.g. Butler and King 2004). The  
169 process is hence continuous, and goes to a new optimum, with a dynamic controlled by  $\mathbf{A}$ .  
170 In both cases, we make the standard assumptions that all traits shift at the same time (but  
171 with different magnitudes), that each shift occurs at the beginning of its branch, and that  
172 all other parameters ( $\mathbf{A}$ ,  $\mathbf{R}$ ) of the process remain unchanged. We further assume that each  
173 jump induces a specific optimum, which implies that there is no homoplasy for the  
174 optimum, that is, no convergent evolution.

### 175 *Simplifying Assumptions*

176 *Trait Independence Assumption.*— The general OU as described above is computationally  
177 hard to fit (Clavel et al. 2015), even when the shifts are fixed *a priori*. For automatic  
178 detection to be tractable in practice, several assumptions can be made. The two methods  
179 that (to our knowledge) tackle this problem in the multivariate setting assume that all the  
180 traits are independent, i.e. that matrices  $\mathbf{A}$  and  $\mathbf{R}$  are *diagonal* (Ingram and Mahler 2013;  
181 Khabbazian et al. 2016). This is often justified by assuming that *a priori* preprocessing  
182 with phylogenetic Principal Component Analysis (pPCA, Revell 2009) leads to  
183 independent traits. However, pPCA assumes a no-shift BM evolution of the traits, and it  
184 can introduce a bias in the downstream analysis conducted on the scores, as shown by  
185 Uyeda et al. (2015). The choice of the number of PC axes to keep is also crucial, and can  
186 qualitatively change the results obtained, leading to the detection of artificial shifts near  
187 the root when not enough PC axes are kept for the analysis, as observed by Khabbazian



188 et al. (2016). Finally, we show theoretically (Section pPCA and Shifts) and numerically  
 189 (Section Simulations Studies, last paragraph) that pPCA fails to decorrelate the data in  
 190 the presence of shifts and may even hamper shift detection accuracy.

191 *Scalar OU (scOU)*.— We offer here an alternative to the independence assumption.  
 192 Computations are greatly simplified when matrices  $\mathbf{A}$  and  $\mathbf{R}$  commute. This happens when  
 193 both of these matrices are diagonal for example, or when  $\mathbf{R}$  is unconstrained and  $\mathbf{A}$  is  
 194 *scalar*, i.e. of the form  $\mathbf{A} = \alpha \mathbf{I}_p$ , where  $\mathbf{I}_p$  is the identity matrix. We call a process  
 195 satisfying the latter assumptions a *scalar OU* (scOU), as it behaves essentially as a  
 196 univariate OU. In particular, its stationary variance is simply given by  $\mathbf{\Gamma} = \mathbf{R}/(2\alpha)$   
 197 (analogous to the formula  $\gamma^2 = \sigma^2/(2\alpha)$  in the univariate case, see e.g. Hansen 1997).

198 We define the scOU model as follows: at the root  $\rho$ , the traits are either drawn from  
 199 the stationary normal distribution with mean  $\boldsymbol{\mu}$  and variance  $\mathbf{\Gamma}$  ( $\mathbf{X}^\rho \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Gamma})$ ), or fixed  
 200 and equal to  $\boldsymbol{\mu}$ . The initial optimum vector is  $\boldsymbol{\beta}_0$  and the conditional distribution of trait  
 201  $\mathbf{X}^i$  at node  $i$  given trait  $\mathbf{X}^{\text{pa}(i)}$  at its parent node  $\text{pa}(i)$  is

$$\mathbf{X}^i \mid \mathbf{X}^{\text{pa}(i)} \sim \mathcal{N} \left( e^{-\alpha \ell_i} \mathbf{X}^{\text{pa}(i)} + (1 - e^{-\alpha \ell_i}) \boldsymbol{\beta}_i, \frac{1}{2\alpha} (1 - e^{-\alpha \ell_i}) \mathbf{R} \right) \quad (1)$$

202 where  $\boldsymbol{\beta}_i = \boldsymbol{\beta}_{\text{pa}(i)} + \boldsymbol{\Delta}^i$  is the optimal value of the process on the branch with length  $\ell_i$   
 203 going from  $\text{pa}(i)$  to  $i$  and  $\boldsymbol{\Delta}$  is the  $N \times p$  matrix of shifts on the branches of the tree: for  
 204 any node  $i$  and any trait  $l$ ,  $\Delta_{il}$  is 0 if there are no shift on the branch going from  $\text{pa}(i)$  to  $i$ ,  
 205 and the value of the shift on trait  $l$  otherwise. At the root, we define  $\boldsymbol{\beta}_\rho = \boldsymbol{\beta}_0$  and, for each  
 206 trait  $l$ :  $\Delta_{\rho l} = e^{-\alpha h} \mu_l + (1 - e^{-\alpha h}) \beta_{0l}$ , where  $h$  is the age of the root (or tree height).

207 The scOU model can also be expressed under a linear form. Let  $\mathbf{U}$  be the  $N \times N$   
 208 matrix where  $U_{ij}$  is 1 if node  $j$  is an ancestor of node  $i$  and 0 otherwise. Let  $\mathbf{T}$  be the  
 209  $n \times N$  matrix made of the  $n$  rows of  $\mathbf{U}$  corresponding to tip taxa. For a given  $\alpha$ , we further  
 210 define the diagonal  $N$  matrix  $\mathbf{W}(\alpha)$  with diagonal term  $W_{ii}(\alpha) = 1 - e^{-\alpha a_{\text{pa}(i)}}$  for any

211 non-root node  $i$ , where  $a_{\text{pa}(i)}$  is the age of node  $\text{pa}(i)$ , and  $W_{\rho\rho}(\alpha) = 1$  for the root node  $\rho$ .  
212 Then the joint distribution of the observed traits  $\mathbf{Y}$  is normal

$$\text{vec}(\mathbf{Y}) \sim \mathcal{N}(\text{vec}(\mathbf{T}\mathbf{W}(\alpha)\mathbf{\Delta}), \mathbf{R} \otimes \mathbf{F}(\alpha)) \quad (2)$$

213 where  $\mathbf{F}(\alpha)$  is the symmetric scaled correlation matrix between the  $n$  tips, with entries  
214  $F_{ij} = \frac{1}{2\alpha}e^{-\alpha d_{ij}}$  if the root is drawn from the stationary distribution, and  
215  $F_{ij} = \frac{1}{2\alpha}e^{-2\alpha d_{ij}}(1 - e^{-2\alpha t_{ij}})$  if the root is fixed, where  $d_{ij}$  is the tree distance between nodes  
216  $i$  and  $j$ . In the next section, this will allow us to rewrite scOU as a BM on a tree with  
217 rescaled branch lengths. This observation is at the core of our statistical inference strategy.

218 The scOU process allows us to handle the correlations that might exist between  
219 traits, and spares us from doing a preliminary pPCA. This however comes at the cost of  
220 assuming that all the traits evolve at the same rate toward their respective optima, with  
221 the same selection strength  $\alpha$ . See the Discussion for further analysis of these assumptions.

## 222 *Identifiability Issues*

223 *Root State.*— It can be easily checked that the parameters  $\boldsymbol{\mu}$  and  $\boldsymbol{\beta}_0$  at the root are not  
224 jointly identifiable from observations at the tips of an ultrametric tree, only the  
225 combination  $\boldsymbol{\lambda} = e^{-\alpha h}\boldsymbol{\mu} + (1 - e^{-\alpha h})\boldsymbol{\beta}_0$  is. See Ho and Ané (2014) for a derivation in the  
226 univariate case. Note that  $\boldsymbol{\lambda}$  corresponds to the first row of the shift matrix  $\mathbf{\Delta}$ . As we  
227 cannot decide from the data, we assume by default  $\boldsymbol{\beta}_0 = \boldsymbol{\mu} = \boldsymbol{\lambda}$ .

228 *Shift Position.*— The location of the shifts may not always be uniquely determined, as  
229 several sets of locations (and magnitudes) may yield the same joint marginal distribution of  
230 the traits at the tips. These identifiability issues have been carefully studied in Bastide  
231 et al. (2016) for the univariate case. Because we assume that all traits shift at the same

232 time, the sets of equivalent shift locations are the same in the multivariate case as in the  
233 univariate case; only the number of parameters involved is different. So, the problem of  
234 counting the total number of parsimonious, non-equivalent shift allocations remains the  
235 same, as well as the problem of listing the allocations that are equivalent to a given one.  
236 As a consequence, all the combinatorial results and algorithms used in Bastide et al. (2016)  
237 are still valid here; only the model selection criterion needs be adapted (see Section  
238 Statistical Inference).

### 239 *Re-scaling of the Tree*

240 *Equivalency scOU / rBM.*— As recalled above, the inference of OU models raises specific  
241 issues, mostly because some maximum likelihood estimates do not have a closed form  
242 expression. Many of these issues can be circumvented using the equivalence between the  
243 univariate BM and OU models described in Blomberg et al. (2003); Ho and Ané (2013);  
244 Pennell et al. (2015), for ultrametric trees, when  $\alpha$  is known. Thanks to the scalar  
245 assumption, this equivalence extends to the multivariate case. Indeed, the marginal  
246 distribution of the traits at the observed tips  $\mathbf{Y}$  given in (2) is the same as the one arising  
247 from a BM model on a re-scaled tree defined by:

$$\mathbf{X}^p \sim \mathcal{N}(\boldsymbol{\beta}_0, \ell_\rho(\alpha)\mathbf{R}) \text{ or } \mathbf{X}^p = \boldsymbol{\beta}_0 \text{ (fixed)}$$

$$\mathbf{X}^i \mid \mathbf{X}^{\text{pa}(i)} \sim \mathcal{N}(\mathbf{X}^{\text{pa}(i)} + \boldsymbol{\Delta}^i(\alpha), \ell_i(\alpha)\mathbf{R}), \quad \text{for non-root node } i.$$

248 where  $\ell_\rho(\alpha) = \frac{1}{2\alpha}e^{-2\alpha h}$ ,  $\ell_i(\alpha) = \frac{1}{2\alpha}e^{-2\alpha h}(e^{2\alpha t_i} - e^{2\alpha t_{\text{pa}(i)}})$ , and  
249  $\boldsymbol{\Delta}^i(\alpha) = (\mathbf{W}(\alpha)\boldsymbol{\Delta})^i = (1 - e^{-\alpha(h-t_{\text{pa}(i)})})\boldsymbol{\Delta}^i$ . Note that, when the root is taken random,  
250 everything happens as if we added a fictive branch above the root with length  $\ell_\rho(\alpha)$ . The  
251 length of this branch increases when  $\alpha$  goes to zero.

252 We emphasize that only the distribution of the observed traits  $\mathbf{Y}$  is preserved and  
253 not the distribution of the complete dataset  $\mathbf{X}$ . As a consequence, ancestral traits at  
254 internal nodes cannot be directly inferred using this representation. Still, the equivalence  
255 recasts inference of  $\mathbf{R}$  and  $\mathbf{W}(\alpha)\mathbf{\Delta}$  in the scOU model into inference of the same  
256 parameters in a much simpler BM model, albeit on a tree with rescaled branch lengths  
257  $\ell_i(\alpha)$ . Note that the rescaling depends on  $\alpha$ , which needs to be inferred separately. See the  
258 discussion (Section Interpretation Issues) for further analysis of this re-scaling.

### 259 *Statistical Inference*

260 *Incomplete Data Model.*— We now discuss how to infer the set of parameters  $\boldsymbol{\theta} = (\mathbf{\Delta}, \mathbf{R})$ .  
261 We adopt a maximum likelihood strategy, which consists in maximizing the log-likelihood  
262 of the observed tip data  $\log p_{\boldsymbol{\theta}}(\mathbf{Y})$  with respect to  $\boldsymbol{\theta}$  to get the estimate  $\hat{\boldsymbol{\theta}}$ . The maximum  
263 likelihood estimate  $\hat{\boldsymbol{\theta}}$  is difficult to derive directly as the computation of  $\log p_{\boldsymbol{\theta}}(\mathbf{Y})$  requires  
264 to integrate over the unobserved values of the traits at the internal nodes. We denote by  $\mathbf{Z}$   
265 the unobserved matrix of size  $m \times p$  of these ancestral traits at internal nodes of the tree:  
266 for each internal node  $j$ ,  $\mathbf{Z}^j$  is the row-vector of the  $p$  ancestral traits at node  $j$ . Following  
267 Bastide et al. (2016), we use the expectation-maximization (EM) algorithm (Dempster  
268 et al. 1977) that relies on an incomplete data representation of the model and takes  
269 advantage of the decomposition of  $\log p_{\boldsymbol{\theta}}(\mathbf{Y})$  as  $\mathbb{E}[\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y}] - \mathbb{E}[\log p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y}) \mid \mathbf{Y}]$ .

270 *EM.*— The M step of the EM algorithm consists in maximizing  $\mathbb{E}[\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y}]$  with  
271 respect to  $\boldsymbol{\theta}$ . For a given value of  $\alpha$ , thanks to the rescaling described in Section Model,  
272 the formulas to update  $\mathbf{\Delta}$  and  $\mathbf{R}$  are explicit (see Appendix EM Inference). The  
273 optimization of  $\alpha$  is achieved over a grid of values, at each point of which a complete EM  
274 algorithm is run.

275 At the M step, we need the mean and variance of the unobserved traits  $\mathbf{Z}^j$  at each internal

276 node  $j$  conditional on the observed traits  $\mathbf{Y}$  at the tips. The E step is dedicated to the  
277 computation of these values, which can be achieved via an upward-downward recursion  
278 (Felsenstein 2004). The upward path goes from the leaves to the root, computing the  
279 conditional means and variances at each internal node given the values of its offspring in a  
280 recursive way. The downward recursion then goes from the root to the leaves, updating the  
281 values at each internal node to condition on the full taxon set. Thanks to the joint  
282 normality of the tip and internal node data, all update formulas have closed form matrix  
283 expressions, even when there are some missing values (see Appendix EM Inference).

284 *Initialization.*— The EM algorithm is known to be very sensitive to the initialization.  
285 Following Bastide et al. (2016), we take advantage of the linear formulation (2) to initialize  
286 the shifts position using a lasso penalization (Tibshirani 1996). This initialization method  
287 is similar to the procedure used in *ℓ1ou* (Khabbazian et al. 2016). See Appendix EM  
288 Inference for more details.

289 *Missing Data.*— EM was originally designed to handle missing data. As a consequence, the  
290 algorithm described above also applies when some traits are unobserved for some taxa.  
291 Indeed, the conditional distribution of the missing traits given the observed ones can be  
292 derived in the same way as in the E step. However, missing data break down the factorized  
293 structure of the dataset and some computational tricks are needed to handle the missing  
294 data efficiently (see Appendix EM Inference).

295 *Model Selection.*— For each value of the number of shifts  $K$ , the EM algorithm described  
296 above provides us with the maximum likelihood estimate  $\hat{\theta}_K$ .  $K$  needs to be estimated to  
297 complete the inference procedure. We do so using a penalized likelihood approach. The  
298 model selection criterion relies on a reformulation of the model in terms of multivariate  
299 linear regression, where we remove the phylogenetic correlation, like independent contrasts

300 and PGLS do. We can re-write (2), for a given  $\alpha$ , as

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{T}}\mathbf{\Delta} + \mathbf{E} \quad \text{where } \tilde{\mathbf{Y}} = \mathbf{F}(\alpha)^{-1/2}\mathbf{Y}, \quad \tilde{\mathbf{T}} = \mathbf{F}(\alpha)^{-1/2}\mathbf{TW}(\alpha),$$

301 where  $\mathbf{E}$  is a  $n \times p$  matrix with independent and identically distributed rows, each row  
 302 being a (transposed) centered Gaussian vector with variance  $\mathbf{R}$ . In the univariate case  
 303 (Bastide et al. 2016), this representation allowed us to cast the problem in the setting  
 304 considered by Baraud et al. (2009), and hence to derive a penalty on the log-likelihood, or,  
 305 equivalently, on the least squares. Taking advantage of the well known fact that the  
 306 maximum likelihood estimators of the coefficients are also the least square ones, and do not  
 307 depend on the variance matrix  $\mathbf{R}$  (see, e.g. Mardia et al. 1979, Section 6), we propose to  
 308 estimate  $K$  using the penalized least squares:

$$\hat{K} = \arg \min_K \left( 1 + \frac{\text{pen}(K)}{n - K} \right) \sum_{j=1}^p \|\tilde{\mathbf{Y}}_j - \hat{\tilde{\mathbf{Y}}}_j^K\|^2$$

309 where  $\tilde{\mathbf{Y}}_j$  is the column of  $\tilde{\mathbf{Y}}$  for the  $j$ -th trait, and  $\hat{\tilde{\mathbf{Y}}}_j^K$  the predicted means for trait  $j$   
 310 from the best model with  $K$  shifts. Using the EM results, this can be written as:

$$\hat{K} = \arg \min_K \left( 1 + \frac{\text{pen}(K)}{n - K} \right) \text{tr} \left[ \hat{\mathbf{R}}(K, \hat{\alpha}) \right]$$

311 where  $\hat{\mathbf{R}}(K, \hat{\alpha})$  is the ML estimate of the variance parameter obtained by the EM for a  
 312 fixed number  $K$  of shifts. The penalty is the same as in the univariate case:

$$\text{pen}(K) = A \frac{n - K - 1}{n - K - 2} \text{EDkhi} \left[ K, n - K - 2, (K + 1)^2 / |\mathcal{S}_K^{\text{PI}}| \right]$$

313 where EDkhi is the function from Definition 3 from Baraud et al. (2009) and  $|\mathcal{S}_K^{\text{PI}}|$  is the  
 314 number of parsimonious identifiable sets of locations for  $K$  shifts, as defined in Bastide

315 et al. (2016). It hence might depends on the topology of the tree, for a tree with  
316 polytomies. For a fully resolved tree,  $|\mathcal{S}_K^{\text{PI}}| = \binom{2n-2-K}{K}$ .  $A$  is a normalizing constant, that  
317 must be greater than 1. In Baraud et al. (2009), the authors showed that it had little  
318 influence in the univariate case, and advised for a value around  $A = 1.1$ . We took this  
319 value as a default.

320 The criterion is directly inspired from the univariate case and inherits its theoretical  
321 properties in the special case  $\mathbf{R} = \sigma^2 \mathbf{I}_p$ . In general however, the criterion should be seen as  
322 a heuristic, although with good empirical properties (see Section Simulations Studies).

### 323 *Implementation*

324 We implemented the method presented above in the PhylogeneticEM R package (R Core  
325 Team 2017), available on the Comprehensive R Archive Network (CRAN). A thorough  
326 documentation of its functions, along with a brief tutorial, is available from the GitHub  
327 repository of the project ([pbastide.github.io/PhylogeneticEM](https://github.com/pbastide/PhylogeneticEM)). Thanks to a  
328 comprehensive suite of unitary tests, that cover approximately 79% of the code  
329 ([codecov.io/gh/pbastide/PhylogeneticEM](https://codecov.io/gh/pbastide/PhylogeneticEM)), and that are run automatically on an  
330 independent Ubuntu server using the continuous integration tool Travis CI  
331 ([travis-ci.org](https://travis-ci.org)), the package was made as robust as possible. The computationally  
332 intensive parts of the analysis, such that the upward-downward algorithm of the M step,  
333 have been coded in C++ to improve performance (see Section Simulations Studies for a  
334 study of the computation times needed to solve problems of typical size). Because the  
335 inference on each  $\alpha$  value on the grid used is independent, they can be easily be done in  
336 parallel, and a built in option allows the user to choose the number of cores to be allocated  
337 to the computations.

## 338 PPCA AND SHIFTS

339 Shift detection in multivariate settings is usually done by first decorrelating traits  
 340 with pPCA before feeding phylogenetic PCs to detection procedures that assume  
 341 independent traits. We show hereafter that even in the simple BM setting, phylogenetic  
 342 PC may still be correlated in the presence of shifts. The problem is only exacerbated in the  
 343 OU setting.

344 *pPCA is biased in the presence of shifts*

345 Assume that the traits evolve as a shifted BM process on the tree, so that  
 346  $\text{vec}(\mathbf{Y}) \sim \mathcal{N}(\text{vec}(\mathbf{a}), \mathbf{R} \otimes \mathbf{C})$ , with  $\mathbf{a}$  being the  $n \times p$  matrix of trait means at the tips.  
 347 Decomposing  $\mathbf{R}$  as  $\mathbf{R} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$ , pPCA relies on the fact that the columns of the matrix  
 348  $\mathbf{Y}\mathbf{V}$  are independent. Therefore, its efficiency relies on an accurate estimation of  $\mathbf{R}$ .

349 The estimate of  $\mathbf{R}$  used in pPCA is  $\hat{\mathbf{R}} = (n - 1)^{-1}(\mathbf{Y} - \mathbf{1}_n \bar{\mathbf{Y}}^T)^T \mathbf{C}^{-1}(\mathbf{Y} - \mathbf{1}_n \bar{\mathbf{Y}}^T)$ ,  
 350 where  $\bar{\mathbf{Y}}^T = (\mathbf{1}_n^T \mathbf{C}^{-1} \mathbf{1}_n)^{-1} \mathbf{1}_n^T \mathbf{C}^{-1} \mathbf{Y}$ , which is known as the estimated phylogenetic mean  
 351 vector (Revell 2009). Decomposing the estimate of  $\mathbf{R}$  as  $\hat{\mathbf{R}} = \hat{\mathbf{V}}\hat{\mathbf{D}}^2\hat{\mathbf{V}}^T$ , pPCA then  
 352 computes the scores as  $\mathbf{S} = (\mathbf{Y} - \mathbf{1}_n \bar{\mathbf{Y}}^T)\hat{\mathbf{V}}$ .

353 In the absence of shift, all species have the same mean vector  $\boldsymbol{\mu}$  so  $\mathbf{a} = \mathbf{1}_n \boldsymbol{\mu}^T$  and  
 354  $\mathbb{E}[\bar{\mathbf{Y}}] = \boldsymbol{\mu}$ . In the presence of shifts, species do not all share the same mean vector so the  
 355 uniform centering is not valid anymore. As a consequence, the estimate of  $\mathbf{R}$  is biased (see  
 356 appendix PCA: Mathematical Derivations):

$$\mathbb{E}[\hat{\mathbf{R}}] = \mathbf{R} + \mathbf{B} \quad \text{where} \quad \mathbf{B} = \frac{1}{n-1} \mathbf{G}^T \mathbf{C}^{-1} \mathbf{G}, \quad \mathbf{G} = \mathbf{a} - \mathbf{1}_n \bar{\mathbf{a}}^T \quad (3)$$

357 The extra term  $\mathbf{B}$  is analogous to the between-group variance in the context of linear  
 358 discriminant analysis and cancels out in the absence of shifts (note that  $\mathbf{R}$  is analogous to  
 359 the within-group variance, see Mardia et al. 1979). Because  $\hat{\mathbf{R}}$  is biased, the columns of the  
 360 score matrix  $\mathbf{S}$  resulting from pPCA are still correlated. We illustrate this phenomenon

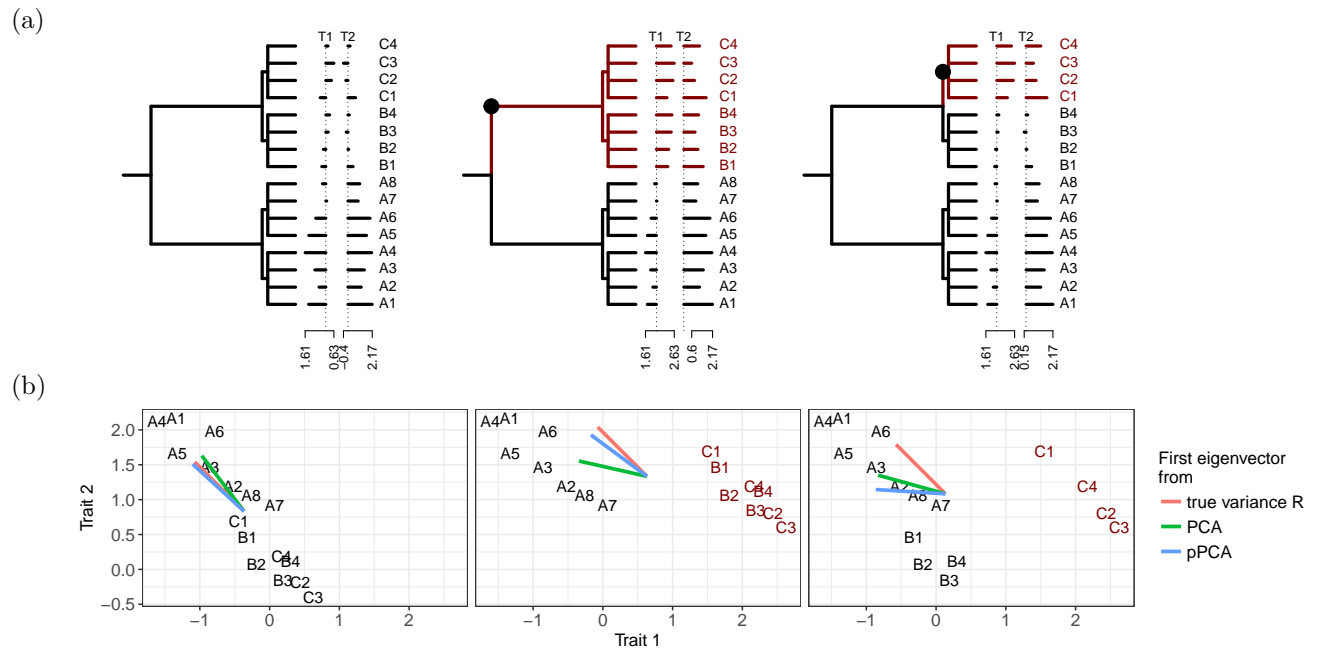


361 below using toy examples.

362 *Illustration: a simple example*

363 To illustrate the impact of shifts on the decorrelation performed by (p)PCA, we used the  
364 simple tree presented in Figure 1a and considered three scenarios. In all scenarios, we  
365 simulated two highly correlated traits under a BM starting from  $(0, 0)$  at the root and with  
366 covariance matrix  $\mathbf{R} = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}$ . The tree has two clearly marked clades, designed  
367 to highlight the differences between pPCA and PCA.  $\mathbf{R}$  is identical in all scenarios; any  
368 preprocessing aiming at decorrelating the traits should retrieve the eigenvectors of  $\mathbf{R}$  as  
369 PCs. In the first scenario, there are no trait shifts on the tree, corresponding to the pPCA  
370 assumptions, and pPCA is indeed quite efficient in finding the PCs (see Fig. 1b, left panel).  
371 In the second scenario, we added a shift on a long branch. This shift induces a species  
372 structure in the trait space that misleads standard PCA. The same structure can however  
373 be achieved by a large increment of the BM on that branch and large increments are likely  
374 on long branches. pPCA therefore copes with the shift quite well and is able to recover  
375 accurate PCs. More quantitatively, the bias induced by the shift on  $\hat{\mathbf{R}}$  is quite small,  
376  $\mathbf{B} = \begin{pmatrix} 0.16 & 0.08 \\ 0.08 & 0.04 \end{pmatrix}$ , around one tenth of the values of  $\mathbf{R}$ . In the third scenario, we put a  
377 shift on a small branch. The structure induced by the shift “breaks down” the upper clade  
378 and is unlikely to arise from the increment of a BM on that branch. It is therefore  
379 antagonistic to pPCA and results in a large bias for  $\hat{\mathbf{R}}$ : the extra term  $\mathbf{B}$  is equal to  
380  $\begin{pmatrix} 1.58 & 0.79 \\ 0.79 & 0.4 \end{pmatrix}$  and comparable to  $\mathbf{R}$ . In that scenario, both PCA and pPCA find axes that  
381 are far away from the eigenvectors of  $\mathbf{R}$  (Figure 1b, right panel). The first eigenvector of  $\mathbf{R}$   
382 captures the evolutionary drift correlation between traits, whereas the PCs of both PCA

383 and pPCA capture a mix of evolutionary drift correlation and correlation resulting from  
 384 shifts along the tree.



## SIMULATIONS STUDIES

385

### *Experimental Design*

386

387 *General Setting.*— We studied the performance of our method using a “star-like”  
 388 experimental design, as opposed to a full-factorial design. We first considered a base  
 389 scenario, corresponding to a base parameter set, and then varied each parameter in turn to

390 assess its impact as in Khabbazian et al. (2016). The base scenario was chosen to be only  
391 moderately difficult, so that our method would find shifts most but not all of the time.

392 For the base scenario, we generated one 160-taxon tree according to a pure birth  
393 process, using the R package `TreeSim` (Stadler 2011), with unit height and birth rate  
394  $\lambda = 0.1$ . We then generated 4 traits on the phylogeny according to the scOU model, with a  
395 rather low selection strength  $\alpha_b = 1$  ( $t_{1/2} = 69\%$  of the tree height), and with a root taken  
396 with a stationary variance of  $\gamma_b^2 = \sigma_b^2 / (2\alpha_b) = 1$ . Diagonal entries of the rate matrix  $\mathbf{R}$  are  
397  $\sigma_b^2$  and off-diagonal entries were set to  $\sigma_b^2 r_d$  with a base correlation of  $r_d = 0.4$  (correlated  
398 traits) when testing the effect of shift number and amplitude, or  $r_d = 0$  (independent  
399 traits) otherwise.

400 Finally, we added three shifts on this phylogeny, with fixed positions (see Figure 2).  
401 Shift amplitudes were calibrated so that the means at the tips differ by about 1 standard  
402 deviation, which constitute a reasonable shift signal (Khabbazian et al. 2016). Each  
403 configuration was replicated 100 times. We then used both our `PhylogeneticEM` and `ℓ1ou`  
404 package (Khabbazian et al. 2016) to study the simulated data. We excluded `SURFACE`  
405 (Ingram and Mahler 2013) from the comparison as is (i) quite slow, (ii) assumes the same  
406 evolutionary model as `ℓ1ou` and (iii) was found to achieve worse accuracy than `ℓ1ou`  
407 (Khabbazian et al. 2016). We used default setting for both methods. For `PhylogeneticEM`  
408 this implies an inference on an automatically chosen grid with 10  $\alpha$  values, on a log scale,  
409 and a maximum number of shifts of  $\sqrt{n} + 5$  (See Bastide et al. 2016 and Appendix EM  
410 Inference for a justification of these default parameters).

411 *Number and Amplitude of Shifts.*— We explored the effect of shifts by varying both their  
412 number and amplitude. We considered successively 0, 3, 7, 11, 15 shifts on the topology,  
413 with positions and values fixed as in Figure 2. Shifts values were chosen to form well  
414 separated tip groups; adjacent (in the tree) group means differ by about 1 standard

415 deviation  $\gamma_b$ . To mimic adaptive events having different consequences on different traits, all  
416 shifts on a trait were then randomly multiplied by  $-1$  or  $+1$ . Finally and to assess the  
417 effect of shift amplitude, we rescaled all shifts by a common factor taking values in  $[0.5, 3]$ .  
418 Low scaling values correspond to smaller, harder to detect, shifts and high values to larger  
419 and easier to detect shifts.

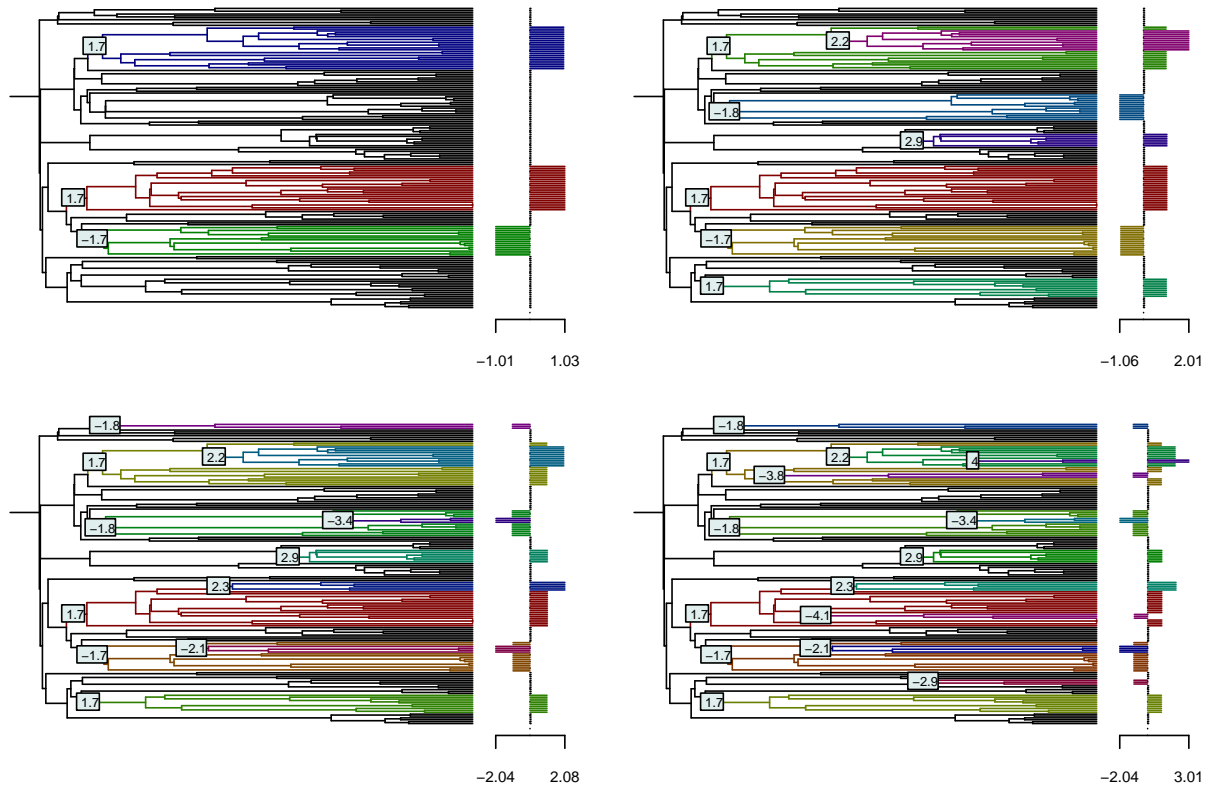


Figure 2: Shifts locations and magnitudes used in the base scenario. Mean trait values are identical for the 4 traits, up to a multiplicative  $\pm 1$  factor and shown at the tips. Colors correspond to the different regimes. The bar plots on the right represent the expected traits values under the base model.

420 *Selection Strength.*— When exploring parameters not related to the shifts, we considered a  
421 base number of 3 shifts and a base scaling factor of 1.25, empirically found to correspond  
422 to a moderately difficult scenario. We also assumed independent traits with the same

423 variance and selection strength (i.e. scalar  $\mathbf{A}$  and  $\mathbf{R}$ , see *model A* in  
424 appendix Kullback-Leibler Divergences). We first varied  $\alpha$  from 1 to 3 (i.e.  $t_{1/2}$  varied  
425 between 35% and 23% of the tree height). The variance  $\sigma^2$  varied with  $\alpha$  to ensure that the  
426 stationary variance  $\gamma_b^2$  remained fixed at  $\gamma_b^2 = 1$ .

427 *Model Mis-specification.*— The two current frameworks ( $\ell 1_{ou}$  and scOU) for multivariate  
428 shift detection assume independent traits (diagonal  $\mathbf{A}$  and  $\mathbf{R}$ ) or correlated traits with  
429 equal selection strengths (scalar  $\mathbf{A}$  and arbitrary  $\mathbf{R}$ ). To assess robustness to model  
430 mis-specification, we simulated data under four classes of models, referred to as A, B, C, D.  
431 Model A is correctly specified for both scOU and  $\ell 1_{ou}$  whereas B, C, D correspond  
432 respectively to mis-specifications for  $\ell 1_{ou}$ , scOU and both. We used the Kullback-Leibler  
433 divergence between models A and B (resp. C, D) to choose parameters that attain  
434 comparable “levels” of mis-specification (see appendix Kullback-Leibler Divergences for  
435 details).

- 436 • *Model A* assumes scalar  $\mathbf{A}$  and  $\mathbf{R}$  (independent traits, same selection strength and  
437 variance) and meets the assumptions of both scOU and  $\ell 1_{ou}$ .
- 438 • *Model B* assumes scalar  $\mathbf{A}$  and arbitrary  $\mathbf{R}$  (correlated traits, same selection  
439 strength) and corresponds to the scOU model. The level of correlation is controlled  
440 by setting all off-diagonal terms to  $\sigma_b^2 r_d$  in  $\mathbf{R}$ . Following Khabbazian et al. (2016),  $r_d$   
441 varies from 0.2 to 0.8, leading to Kullback divergences of up to 288.36 units.
- 442 • *Model C* assumes diagonal, but not scalar,  $\mathbf{A}$ , and diagonal  $\mathbf{R}$  (independent traits,  
443 different selection strengths), which matches the assumptions of  $\ell 1_{ou}$  only. We  
444 considered  $\mathbf{A} = \alpha \text{Diag}(s^{-1.5}, s^{-0.5}, s^{0.5}, s^{1.5})$  with  $s$  varying from 2 to 8. We  
445 accordingly set  $\mathbf{R} = 2\gamma_b^2 \mathbf{A}$  to ensure that all traits have stationary variance  $\gamma_b^2 = 1$ .  
446 This led to Kullback divergences of up to 286.78 units.

447 • *Model D* assumes non-diagonal  $\mathbf{A}$  and diagonal  $\mathbf{R}$  (uncorrelated drift, but correlated  
448 traits selection) and violates both models. Following Khabbazian et al. (2016), all  
449 off-diagonal elements of  $\mathbf{A}$  were set to  $\alpha_b r_s$ , varying from 0.2 to 0.8. In this case, the  
450 stationary variance is not diagonal but has diagonal entries equal to  $\frac{\sigma^2}{2} \frac{1+(p-2)r_s}{(1-r_s)(1+(p-1)r_s)}$ .  
451 We thus rescaled  $\sigma^2$  appropriately to ensure that each trait has marginal stationary  
452 variance  $\gamma_b^2 = 1$  as previously. This led to Kullback divergences of up to 112.98 units.

453 We expected  $\ell 1ou$  to outperform scOU in model C and vice versa in model B. To be  
454 fair to both methods, we selected parameter ranges leading to similar Kullback divergences,  
455 to achieve similar levels of mis-specifications. However, both deviations produce datasets  
456 with groups that are also theoretically easier to discriminate compared to model A (see  
457 Figure 3). Indeed, we can quantify the difficulty of a dataset in terms of group separation  
458 by the Mahalanobis distance between the observed data and their expected mean,  
459 (phylogenetically) estimated in the absence of shifts:

$$D = \left\| \mathbf{Y}_{\text{vec}} - (\mathbf{1}^T \boldsymbol{\Sigma}_d \mathbf{1})^{-1} \mathbf{1}^T \boldsymbol{\Sigma}_d \mathbf{Y}_{\text{vec}} \right\|_{\boldsymbol{\Sigma}_d^{-1}}^2 - (np - N_{\text{NA}}) \quad (4)$$

460 where  $\mathbf{Y}_{\text{vec}}$  is the vector of observed data at the tips (omitting missing values),  $\boldsymbol{\Sigma}_d$  is the  
461 true variance of  $\mathbf{Y}_{\text{vec}}$  and  $N_{\text{NA}}$  is the number of missing values. In the absence of shifts  
462  $\mathbb{E}[D] = 0$  and  $\mathbb{E}[D]$  increases when groups are well separated.

463 *Number of Observations.*— We varied the number of observations by (i) varying the  
464 number of taxa and (ii) adding missing values. To change the number of taxa, we  
465 generated 6 extra trees with the same parameters as before but with 32 to 256 taxa. The  
466 three shifts were fixed as in Figure 4. To test the ability of our method to handle missing  
467 data, we removed observations at random in our base scenario, taking care to keep at least  
468 one observed trait per species, so as not to change the number of taxa. The fraction of

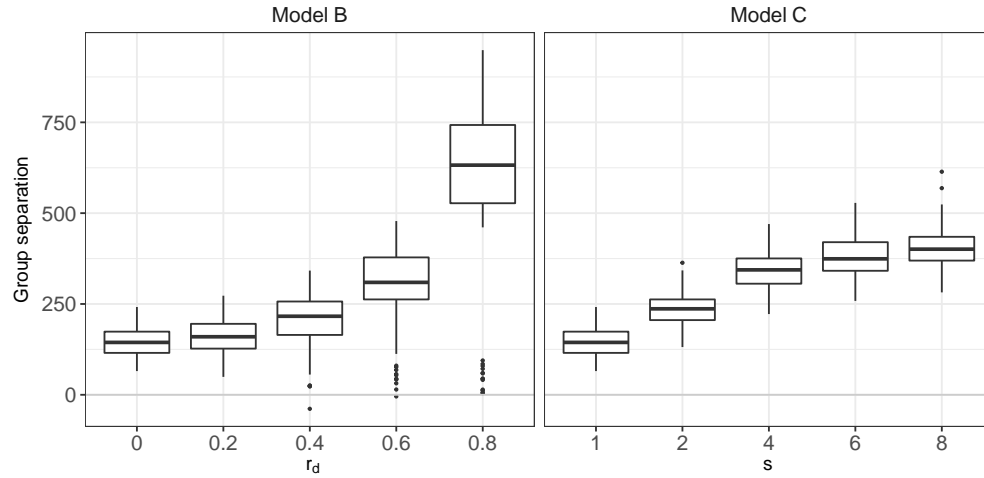


Figure 3: Impact of trait correlation  $r_d$  (left) and unequal selection strengths  $s$  (right) on group separation, as defined in Eq. (4). Unequal selection strengths ( $s > 1$ ) and trait correlations ( $r_d > 0$ ) both increase group separation and make it easier to detect shifts.

469 missing data varied from 5% to 50%.

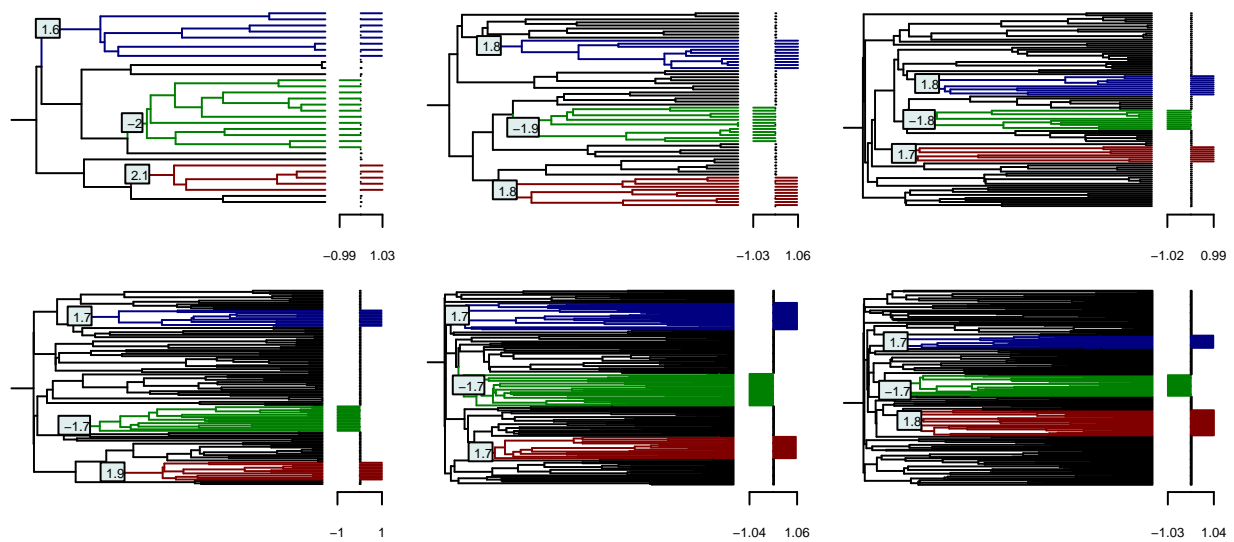


Figure 4: Shifts locations and magnitudes used for the test trees with, respectively, 32, 64, 96, 128, 192, 256 taxa.

471 *Number and Amplitude of Shifts.*— We assessed shifts detection accuracy with the Adjusted  
472 Rand Index (ARI, Hubert and Arabie 1985) between the true clustering of the tips, and the  
473 clustering induced by the inferred shifts (Fig. 5, top). Before adjustment, the Rand index  
474 is proportional to the number of pairs of species correctly classified in the same group or  
475 correctly classified in different groups. The ARI has maximum value of 1 (for a perfectly  
476 inferred clustering) and has expected value of 0, conditional on the inferred number and  
477 size of clusters. We use this measure rather than the classical precision/sensitivity graphs  
478 as only the clustering can be recovered unambiguously (see Section Model). Note also that  
479 when there is no shift ( $K = 0$ ), there is only one true cluster, and the ARI is either 1 if no  
480 shift is found, or 0 otherwise (see appendix Note on the ARI).

481 Figure 5 (top panel) shows that, unsurprisingly, both methods detect the number  
482 and positions of shifts more accurately when the shifts have higher amplitudes.  
483 PhylogeneticEM is also consistently better than  $\ell 1_{ou}$  when there is a base correlation (here,  
484  $r_b = 0.4$ , see section Simulations Studies), which is expected as the independence  
485 assumption of  $\ell 1_{ou}$  is then violated. The case  $K = 0$  (no shift) shows that  $\ell 1_{ou}$   
486 systematically finds shifts when there are none, leading to an ARI of 0. More generally,  
487  $\ell 1_{ou}$  is prone to over-estimating the number of shifts, even when they have a high  
488 magnitude (Fig. 5, bottom) whereas PhylogeneticEM is more conservative and  
489 underestimates the number of shifts when they are difficult to detect.

490 *Selection Strength and Model Mis-specifications.*— Our method is relatively robust to  
491 model mis-specification (Fig. 6, top). The first panel confirms that, under model A, high  
492 values of  $\alpha$  reduce the stationary variance and lead to higher ARI values and lower RMSEs  
493 for continuous parameters (Fig. 6, bottom, leftmost panel). Similarly, scOU (resp.  $\ell 1_{ou}$ )  
494 achieves high ARI values under well specified models A and B (resp. A and C). The  
495 mis-specification of model C (different selection strengths) does not affect scOU much: it



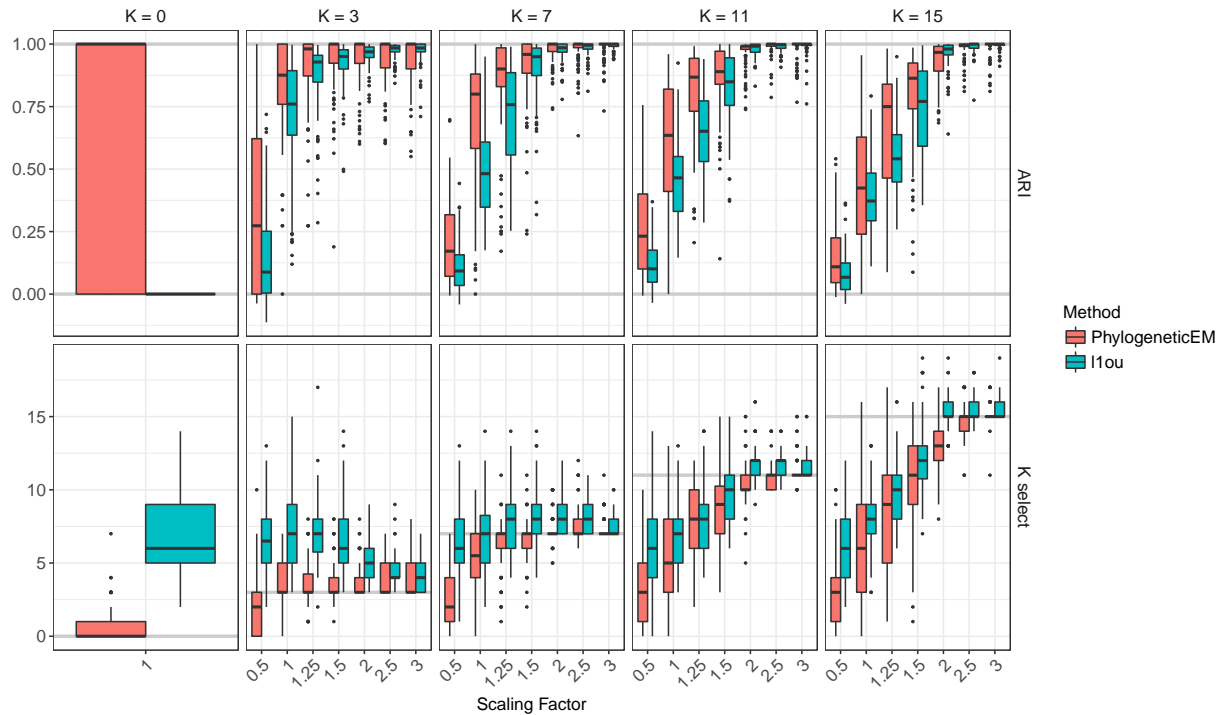


Figure 5: ARI (top) and number of shifts selected (bottom) for the solutions found by PhylogeneticEM (red) and  $\ell_{1ou}$  (blue). Each box corresponds to one of the configuration shown in Figure 2, with a scaling factor varying between 0.5 and 3, and a true number of shift between 0 and 15 (solid lines, bottom). For the ARI, the two lines represent the maximum (1) and expected (0, for a random solution) ARI values.

496 has higher ARI dispersion than  $\ell_{1ou}$  but their median ARI are comparable. By contrast,  
 497  $\ell_{1ou}$  is severely affected by correlated evolution (model C) and higher levels of correlations  
 498 lead to significantly lower accuracy, even though group separation is increased (Fig. 3,  
 499 right). Finally, both methods are negatively affected by correlated selection strengths  
 500 (Model D), although  $\ell_{1ou}$  seems more robust to this type of mis-specification.

501 Although shift detection is relatively unaffected by model mis-specification,  
 502 parameter estimations suffers from it (Fig. 6, bottom, center and right panels). Both  $\ell_{1ou}$   
 503 and scOU behave better for model A than for model D and as expected, scOU is not  
 504 affected by trait correlation (model B) whereas  $\ell_{1ou}$  is. Unequal selection strengths (model  
 505 C) degrades parameter estimation for both PhylogeneticEM and, surprisingly,  $\ell_{1ou}$ , that

506 should in principle remain unaffected. Overall, features of trait evolution not properly  
 507 accounted for by the inference methods (e.g. correlated selection strengths) are turned into  
 508 overestimated variances. Note that the quality of the estimation of  $\Gamma$  is depends strongly  
 509 on the estimation of  $\alpha$ , and could be improved by taking a finer grid for this parameter.

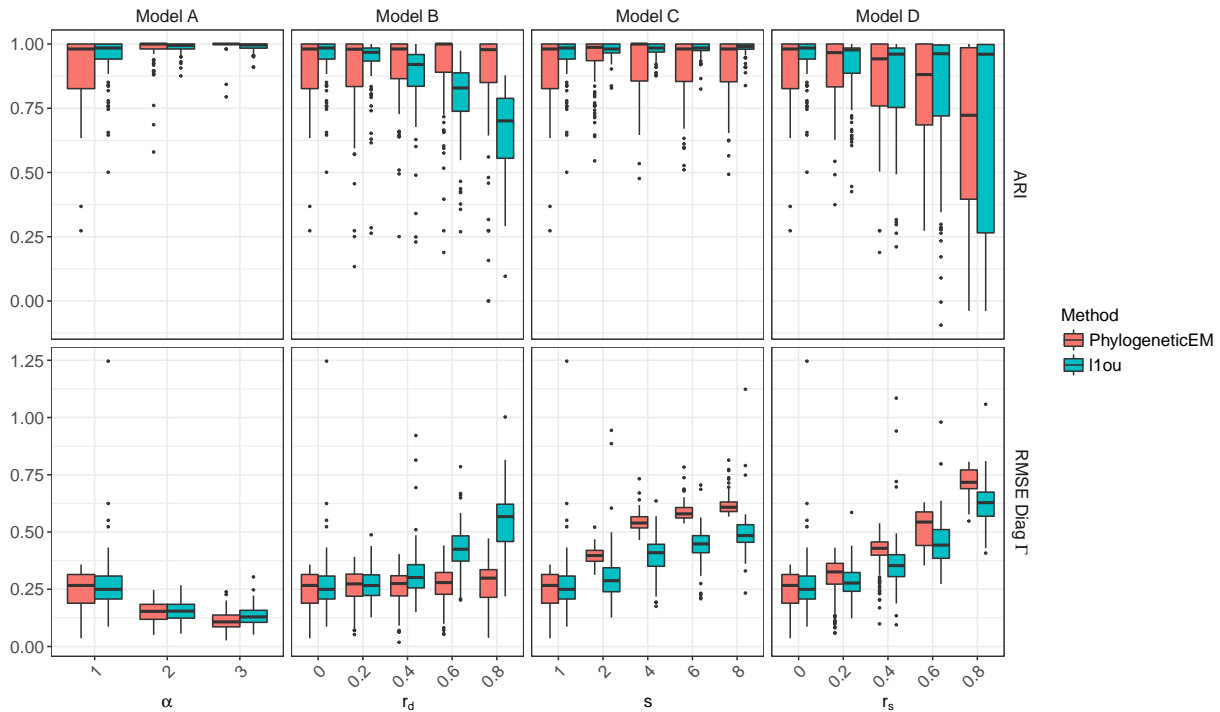


Figure 6: ARI (top) and root mean squared error (RMSE) of the diagonal values of the estimated stationary variance  $\Gamma$  (bottom) for the solutions found by PhylogeneticEM (red) and l1ou (blue). Each panel corresponds to a different type of mis-specification (except Model A) and the parameters  $r_d$ ,  $s$  and  $r_s$  control the level of mis-specification, with leftmost values corresponding to no mis-specification. For the ARI, the solid lines represent the maximum (1) and expected (0, for a random solution with the same number and size of clusters) ARI values.

510 *Number of observations and Computation Time.*— For a given number of shifts, shift  
 511 detection becomes easier as the number of taxa increases (Fig. 7, left). Furthermore, our  
 512 method is robust against missing data with detection accuracy only slightly decreased  
 513 when up to 50% of the observations are missing (Fig. 7, right). Finally, our implementation

514 of the EM algorithm, using only two tree traversals (see appendix EM Inference) and coded  
515 in C++, is reasonably fast. Inference takes roughly 15 minutes on a single core on the base  
516 160 taxa tree and less than 45 minutes on the largest simulated trees (256 taxa).  $\ell 1ou$   
517 scales less efficiently: it is faster for very small trees (32 taxa) but median running times go  
518 up to 20 hours for the large 256-taxon tree. Those long running times were unexpected and  
519 higher than the ones reported in Khabbazian et al. (2016). This discrepancy is partly due  
520 to the maximum number of shifts allowed, which strongly impacts the running time of  
521  $\ell 1ou$ . Khabbazian et al. (2016) capped it at twice the true number of shifts (6 shifts in our  
522 base scenario), while we used the default setting, which is half the number of tips (i.e. from  
523 16 to 128 shifts).

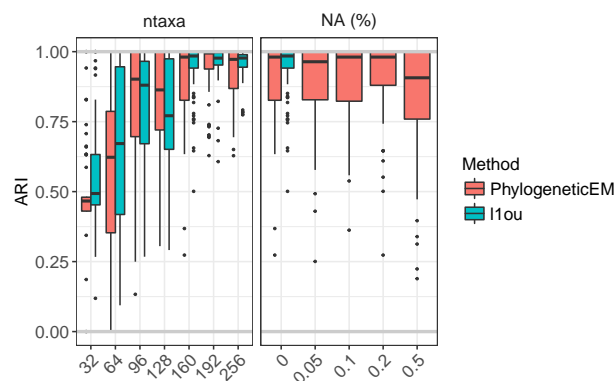


Figure 7: ARI of the solutions found by PhylogeneticEM (red) and  $\ell 1ou$  (blue) when the number of taxa (left) or the number of missing values (right) increases. No ARI is available for  $\ell 1ou$  when there are missing values as it does not accept them in the version used here, v1.21.

524 *Impact of pPCA on shift detection accuracy.*— To illustrate how pPCA can both improve  
525 and hamper shift detection, we compared PhylogeneticEM on raw traits to  $\ell 1ou$  on both  
526 raw traits and phylogenetic PCs. Figure 9a shows that in our base scenario, with three  
527 moderate shifts, pPCA preprocessing slightly decreases performance for low levels of  
528 correlations ( $r_d \leq 0.2$ ) but drastically improves them for moderate to high correlations

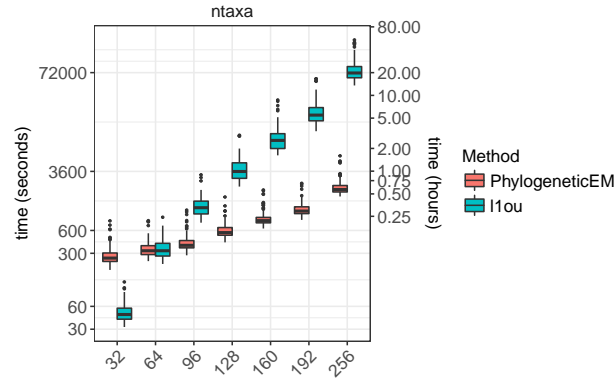


Figure 8: Inference running times (in log-scale) of scOU and  $\ell 1_{ou}$ . All tests were run on a high-performance computing facility with CPU speeds ranging from 2.2 to 2.8Ghz.

529 levels ( $r_d \geq 0.6$ ). Although pre-processing is neutral at moderate correlation levels  
 530 ( $r_d = 0.4$ ) with three “easy” shifts, it becomes harmful and degrades the performances of  
 531  $\ell 1_{ou}$  when the number or magnitude of the shifts increases (Fig. 9b). As expected,  
 532 PhylogeneticEM is unaffected by the pPCA preprocessing, up to numerical issues.

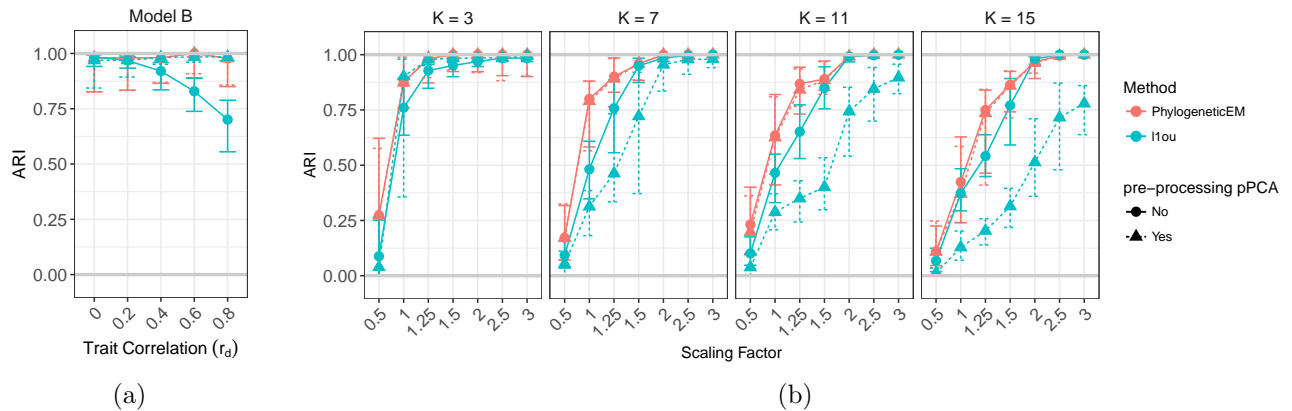


Figure 9: ARI of the solutions found by PhylogeneticEM (red) and  $\ell 1_{ou}$  (blue), without (solid lines) or with (dotted lines) pPCA preprocessing. (a) Trait correlation ( $r_d$ ) increases from 0 to 0.8. (b) Each box corresponds to one of the configuration shown in Figure 2, and shifts are increasingly large with a scaling factor varying between 0.5 and 3.

534 We used PhylogeneticEM to re-analyse two publicly available datasets.

### 535 *New World Monkeys*

536 We first considered the evolution of brain shape in New World Monkeys studied by Aristide  
537 et al. (2016). The dataset consists of 49 species on a time-calibrated maximum-likelihood  
538 tree. The traits under study are the first two principal components (PC1, PC2) resulting  
539 from a PCA on 399 landmarks describing brain shape. We ran PhylogeneticEM on a grid of  
540 30 values for the  $\alpha$  parameter. To make this parameter easily interpretable, we report the  
541 *phylogenetic half-life*  $t_{1/2} = \ln(2)/\alpha$  (Hansen 1997), expressed in percentage of total tree  
542 height. Here,  $t_{1/2}$  took values between 0.46 % and 277.26 %. We allowed for a maximum of  
543 20 shifts. The inference took 17.56 minutes, parallelized on 5 cores.

544 The model selection criterion suggests an optimal value of  $\hat{K} = 4$  shifts (Fig. 10,  
545 inset graph). The criterion does not show a very sharp minimum, however, and a value of  
546  $\hat{K} = 5$  shifts also seems to be a good candidate. In order to compare our results with that  
547 presented in Aristide et al. (2016), we present the solution with 5 shifts (see Fig. 10, left).  
548 The solution with 4 shifts is very similar, except that the group with *Aotus* species is  
549 absent (in red, see Fig. 10, and supplementary Fig. 14 in Appendix Case Study). Note  
550 that, because of this added group, the solution with  $\hat{K} = 5$  has 3 equivalent parsimonious  
551 allocations of the shifts (see supplementary Fig. 15 in Appendix Case Study). The groups  
552 found by PhylogeneticEM (Fig. 10) are in close agreement with the ecological niches defined  
553 in Aristide et al. (2016). There are three main differences. First, there is no jump  
554 associated with the *Pithecia* species who, although having their own ecological niche, seem  
555 to have quite similar brain shapes as closely related species. Second, *Callicebus* and *Aotus*  
556 are marked as convergent in Aristide et al. (2016) (in red, right), but form two distinct  
557 groups in our model (in pink and red, left). This is due to our assumption of no homoplasy.  
558 Finally, the group with *Chiropotes*, *Ateles* and *Cebus* species (in black) was found as

559 having the “ancestral” trait optimum, while it is marked as “convergent” in Aristide et al.  
 560 (2016). This is because we did not include any information from the fossil record (not  
 561 available for brain shape), but instead used a parsimonious solution. Note that the coloring  
 562 displayed in Aristide et al. (2016) is *not* parsimonious. The two models have the same  
 563 number of distinct groups.

564 The selected  $\alpha$  value was found to be reasonably high, with  $t_{1/2} = 12.58\%$ . The  
 565 estimated correlation between the two PCs was  $-0.13$ , confirming that PCA does not  
 566 result in independent traits.

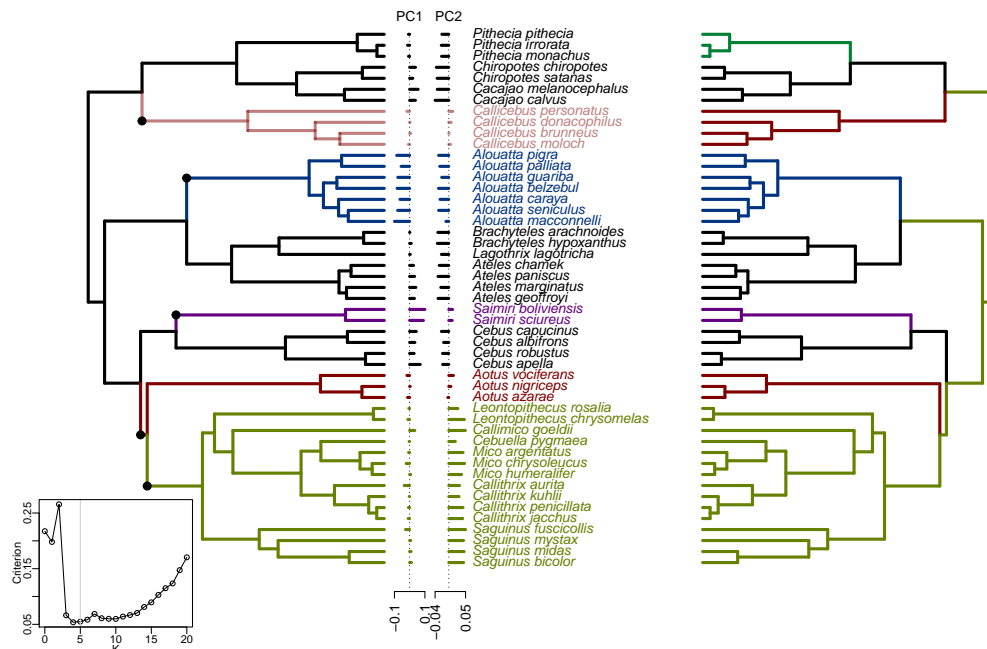


Figure 10: Solution given by PhylogeneticEM for  $K = 5$  (left) against groups defined in Aristide et al. (2016, Fig. 3) (right), based on ecological criteria including *locomotion* (arboreal quadrupedal walk, clamber and suspensory locomotion or clawed locomotion), *diet* (leaves, fruits, seeds or insects) and *group size* (smaller or larger than 15 individuals). The inset graph shows the model selection criterion. The minimum is for  $K = 4$ , but  $K = 5$  is also a good candidate.

568 We then considered the dataset from Mahler et al. (2013), which consists in 100  
569 lizard species on a time-calibrated maximum likelihood tree and 11 morphological traits.  
570 We chose this example because of the large number of traits and the high correlation  
571 between traits, as all traits are highly correlated ( $0.82 < \rho < 0.97$ ) with snout-to-vent  
572 length (SVL).

573 To deal with the correlation between traits, Mahler et al. (2010, 2013) first  
574 performed a phylogenetic regression of all the traits against SVL, retrieved the residuals  
575 and then applied a phylogenetic PCA on SVL and the previous residuals, from which they  
576 used the first four components (pPC1 to pPC4) for their shift analysis. We first explored  
577 how the number of pPCs used can impact the shift detection. Hence we ran  
578 PhylogeneticEM 11 times, including 1 to 11 pPCs in the input dataset. Each run was done  
579 on a grid of 100 values of  $\alpha$ , with  $t_{1/2} = \ln(2)/\alpha \in [0.99, 693.15]$  % of tree height, and  
580 allowing for a maximum of 20 shifts. It appears that the result is quite sensitive to the  
581 number of pPCs included: the selected number of shifts varies from 20, the maximum  
582 allowed, to 5 (Fig. 11). When 4 pPCs were used, as in the original study, the estimated  
583 covariance matrix  $\mathbf{R}$  contains many high correlations, showing that the pPCs are not  
584 phylogenetically independent (Fig. 11).

585 To avoid the difficult choice of the number of pPCs, we considered the direct  
586 analysis of the raw traits without any pre-processing, and found no shift when running  
587 PhylogeneticEM. Although the likelihood was found to increase with  $K$ , the model selection  
588 criterion profile was found erratic, suggesting numerical instability. A natural suspect for  
589 such instability is the extreme correlation between some traits (0.996 for tibia and  
590 metatarsal lengths), which results in bad conditioning of several matrices that must be  
591 inverted. To circumvent this problem, we used the two pseudo-orthogonalization strategies  
592 described above, running PhylogeneticEM on the SVL plus residuals dataset, and on the 11  
593 pPCs, with the same parameters as above. Note that all these transformations use a

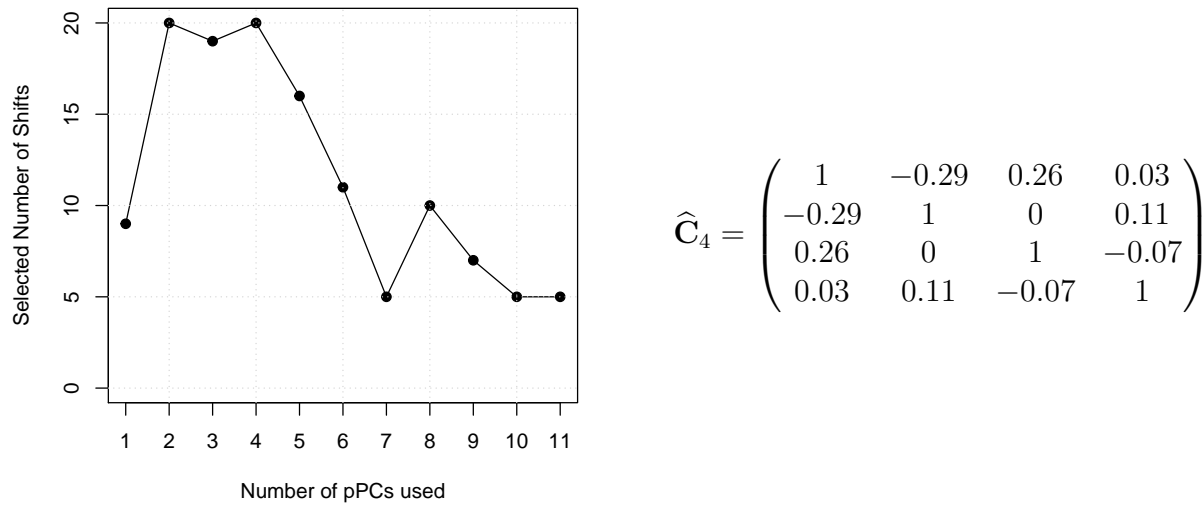


Figure 11: Lizard dataset: selected number of shifts  $\hat{K}$  given the number of pPCs included in the analysis (left) and estimated correlation matrix between the first four pPCs (right).

594 rotation matrix, so that the likelihood and the least squares of the original or of any of the  
595 two transformed datasets are the same. Hence, the objective function, as well as the model  
596 selection criterion, should remain unchanged. Still, slight differences were found between  
597 the maximized likelihood for each pseudo-orthogonalized datasets. For each value of  $K$ , we  
598 therefore retained the solution with the highest likelihood.

599 Using the model selection criterion given in Section Statistical Inference, we found  
600  $\hat{K} = 5$  shifts, which are displayed in Figure 12, along with the ecomorphs as described in  
601 Mahler et al. (2013).

602 Three of those shifts seem to single out grass-bush *Anolis*, that appear to have a  
603 rather small body size, with longer than expected lower limbs and tail, and shorter upper  
604 limbs. The two others might be associated with twig *Anolis*, that have smaller than  
605 expected limbs and tails. Because of our no-homoplasy assumption, one of those shifts  
606 encompasses some species living in other ecomorphs (namely, trunk, trunk-crown and  
607 un-classified). The shift, designed to be coherent with the phylogeny, is located on the



608 stem lineage of the smallest clade encompassing the bulk of twig lizards.

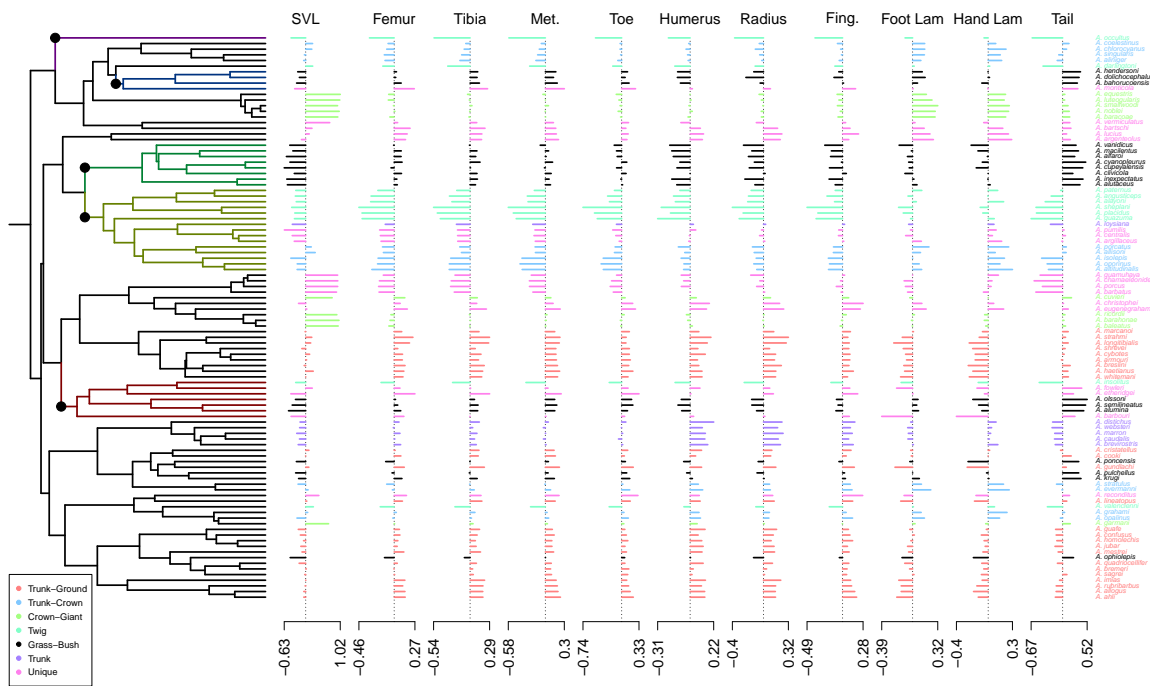


Figure 12: Lizard dataset: solution found by PhylogeneticEM. Groups produced by the shifts are colored on the edges of the tree. The species are colored according to ecomorphs defined in Mahler et al. (2013). The traits are the snout-to-vent length (SVL), and the phylogenetic residuals of the regression against SVL of the following traits: femur length, tibia length, metatarsal IV length, toe IV length, humerus length, radius length, finger IV length, lamina number (toe and finger IV), and tail length. The same transformations were used as in Mahler et al. (2010, 2013)

609

### Comments

610 On both examples (p)PCA does not correct *a priori* for the correlation between the  
611 traits in the presence of shifts. In Section pPCA and Shifts we formally proved that it  
612 cannot correct for it, actually. As a consequence, any shift detection methods has to

613 account for the correlation between traits.

614 Still, high correlations between traits may raise strong numerical issues, so PCA can be  
615 used as a *pseudo-orthogonalization* of traits, as well as any other linear distance-preserving  
616 transformation that would reduce the correlation between them. This does not dispense of  
617 considering the correlation between the transformed traits in the model.

618 The other interest of PCA is to reduce the dimension of the data, which may be  
619 desirable when dealing with a large number of traits, such as the original dataset from  
620 Aristide et al. (2016). Since PCA does not correct for the right correlation, we have no clue  
621 whether or not the dimension reduction performed by PCA is relevant for shift detection,  
622 or if it may remove precisely the direction along which the shifts occur. The relevant  
623 dimension reduction would consist in approximating the correlation matrix  $\mathbf{R}$  with a matrix  
624 of lower rank  $q < p$ . This can obviously not be done before the shifts are known, which  
625 suggests that shift detection and dimension reduction should be performed simultaneously.

## 626 DISCUSSION

627 Many phenotypic traits appear to evolve relatively smoothly over time and across  
628 many taxa. However, changes in evolutionary pressures (dispersal to new geographic zones,  
629 diet change, etc) or key innovations (bipedal locomotion) may cause bursts of rapid trait  
630 evolution, coined evolutionary jumps by Simpson (1944). Phenotypic traits typically evolve  
631 in a coordinated way (Mahler et al. 2013; Aristide et al. 2015) and a multivariate  
632 framework is thus best suited to detect evolutionary jumps. We introduced here an  
633 Expectation Maximization algorithm embedded in a maximum-likelihood multivariate  
634 framework to infer shifts strength, location and number. Importantly, our method uses  
635 Gaussian elimination, just like Fitzjohn (2012), to avoid computing inverses of large  
636 variance-covariance matrices and can cope with missing data, an especially important

637 problem in the multivariate setting where some traits are bound to be missing for some  
638 taxa. We demonstrated the applicability and accuracy of our method on simulated datasets  
639 and by identifying jumps for body size evolution in *Anolis* lizards and brain shapes of New  
640 World Monkeys. In both systems, the well-supported jumps occurred on stem lineages of  
641 clades that differ in terms of diet, locomotion, group size or foraging strategy (see Aristide  
642 et al. 2016 for a detailed discussion) supporting the Simpsonian hypothesis.

### 643 *Interpretation Issues*

644 We emphasize that the interpretation of  $\alpha$  is a matter of discussion. We introduced  
645 the scOU in terms of adaptive evolution with a selection strength  $\alpha$  on the tree. However,  
646 the equivalency between OU and BM on a distorted tree suggests that  $\alpha$  can also be seen  
647 as a “phylogenetic signal” parameter, like Pagel’s  $\lambda$  (Pagel 1999). When  $\alpha$  is small,  
648  $\ell_i(\alpha) \simeq \ell_i$  so that branch lengths are unchanged and the phylogenetic variance is preserved.  
649 At the other end of the spectrum, when  $\alpha$  is large,  $\ell_i(\alpha) \simeq 0$  for inner branches and the  
650 rescaled tree behaves almost like a star tree. However and unlike Pagel’s  $\lambda$ ,  $\alpha$  also dictates  
651 how shifts in the optima in the original OU ( $\Delta^{OU}$ ) are transformed into shifts in the traits  
652 values in the rescaled BM ( $\Delta^{BM}(\alpha)$ ). For small  $\alpha$ , recall to the optima is weak and shifts  
653 on the optima affect the traits values minimally ( $\Delta^{BM}(\alpha) \simeq 0$ ). By contrast, for large  $\alpha$ ,  
654 the recall is strong and shifts on the optima are instantaneously passed on to the traits  
655 values ( $\Delta^{BM}(\alpha) \simeq \Delta^{OU}$ ). Note however that in both cases, the topology is never lost: a  
656 shift, no matter how small its amplitude or how short the branch it occurs on, always  
657 affects the same species.

658 Note that if we observed traits values at some ancestral nodes (e.g. from the fossil  
659 record), the equivalency between BM and OU would break down:  $\alpha$  would recover its strict  
660 interpretation as selection strength. On non-ultrametric trees, our inference strategy does  
661 not benefit from the computational trick to speed up the M step. Similarly to the

662 univariate case, we could write a *generalized* EM algorithm to handle this situation. In  
663 Bastide et al. (2016), we used a lasso-based heuristic to raise, if not maximize, the  
664 objective function at the M step. It worked quite well, but was much slower. This  
665 approach could be extended to the multivariate setting, although with impaired  
666 computational burden. Note also that some shifts configuration that are not identifiable in  
667 the absence of fossil data become distinguishable with the addition of fossil data. This  
668 affects our model selection criterion, which relies on the number of distinct identifiable  
669 solutions. Computing this number on a non-ultrametric tree for an OU remains an open  
670 problem, and is probably highly dependent on the topology of the tree.

### 671 *Noncausal Correlations*

672  $\ell_{1ou}$ , SURFACE and PhylogeneticEM make many simplifying assumptions to achieve  
673 tractable models. Chief among them is the assumption that  $\mathbf{A}$  is diagonal. While  $\ell_{1ou}$  and  
674 SURFACE both assume independent traits, PhylogeneticEM can handle correlated traits  
675 through non-diagonal variance matrix  $\mathbf{R}$ . We warn the reader that correlations encoded by  
676  $\mathbf{R}$  are not causal and only capture *coordinated* and non selective traits evolution: i.e. when  
677 arm length increases, so does leg length. In order to capture evolution of trait  $i$  *in response*  
678 to changes in trait  $j$  (i.e. when arm length strays away from its optimal value, does leg  
679 length move away or toward its own optimum) one should rather look at the value of  $A_{ij}$ ,  
680 as was recently pointed out (Reitan et al. 2012; Liow et al. 2015; Manceau et al. 2016).  
681 Our simplifying assumptions are justified by various considerations: our focus on inference  
682 of shifts rather than proper estimation of  $\mathbf{A}$  and  $\mathbf{R}$ , simulations showing that shift  
683 detection is robust to moderate values of off-diagonal terms in  $\mathbf{A}$ , difficulties to  
684 simultaneously estimate  $\alpha$  and shifts even in the univariate case (Butler and King 2004),  
685 and computational gain achieved by considering scalar or diagonal  $\mathbf{A}$ . They also suggest  
686 that if the focus is on causal correlation in the presence of shifts, a two-step strategy that

687 first detects shifts using a crude but robust model, then includes those shifts in a more  
688 complex model, may achieve good performance.

689         The other simplifying assumption we made is that all traits shift at the same time.  
690 It makes formal analysis of identifiability issues and selection of the number of shifts  
691 similar to the univariate case, previously studied in Bastide et al. (2016). The assumption  
692 is likely to be false in practice, however. Asynchronous shifts are an interesting extension of  
693 the model. An ambitious framework would be to build from the ground up a model that  
694 allows for different shifts on different traits. It would have to deal with the combinatorial  
695 complexity induced by asynchronous shifts, and to use a different selection criterion for the  
696 number of shifts. A less ambitious but more pragmatic approach would be a postprocessing  
697 of the shifts to select, for each shift, the traits that actually jumped. This would require  
698 derivation of confidence intervals for the shift values.

699 Finally, and unlike SURFACE and new version v1.40 of *ℓ1ou*, our model excludes convergent  
700 evolution. This limitation is shared with other shift detection methods such as *bayou*  
701 (Uyeda and Harmon 2014) in the univariate case. This exclusion simplifies formal analysis  
702 and allows us to borrow from the framework of convex characters on a tree developed in  
703 Semple and Steel (2003) but is also likely to be false in practice. A straightforward  
704 extension of our method to detect convergence relies again on postprocessing of the shifts:  
705 the inferred optimal value of a trait after a shift can be tested to assess whether or not it is  
706 different from previously inferred optimal values and warrants a regime of its own.

### 707 *Nature of the jumps*

708         We model shifts as instantaneous and immediately following speciation events, like  
709 in the punctuated equilibrium theory of Eldredge and Gould (1972). We don't argue that  
710 this is necessary the case. Selection and drift can reasonably be seen as instantaneous over  
711 macroevolutionary timescales but by no means over microevolutionary timescales. There is

712 very strong evidence, for example in peppered moths (Cook et al. 2012), that rapid  
713 adaptation can happen even in the absence of speciation. However our model does not  
714 allow us to distinguish between many small jumps distributed across a branch, one big  
715 jump anywhere on that branch and one big jump immediately following speciation, and  
716 therefore between punctuated or Simpsonian evolution.

## 717 ACKNOWLEDGMENTS

718 We are grateful to the INRA MIGALE bioinformatics platform  
719 (<http://migale.jouy.inra.fr>) for providing the computational resources needed for the  
720 experiments.

## 721 FUNDINGS

722 The visit of PB to the University of Wisconsin-Madison during the fall of 2015 was  
723 funded by a grant from the Franco-American Fulbright Commission.

724 \*

## 725 References

- 726 Aristide L, dos Reis SF, Machado AC, Lima I, Lopes RT, Perez SI. 2016. Brain shape  
727 convergence in the adaptive radiation of New World monkeys. *Proceedings of the*  
728 *National Academy of Sciences*. 113:2158–2163.
- 729 Aristide L, Rosenberger AL, Tejedor MF, Perez SI. 2015. Modeling lineage and phenotypic  
730 diversification in the New World monkey (Platyrrhini, Primates) radiation. *Molecular*  
731 *Phylogenetics and Evolution*. 82:375–385.

- 732 Baraud Y, Giraud C, Huet S. 2009. Gaussian model selection with an unknown variance.  
733 *Annals of Statistics*. 37:630–672.
- 734 Bartoszek K, Pienaar J, Mostad P, Andersson S, Hansen TF. 2012. A phylogenetic  
735 comparative method for studying multivariate adaptation. *Journal of Theoretical*  
736 *Biology*. 314:204–215.
- 737 Bastide P, Mariadassou M, Robin S. 2016. Detection of adaptive shifts on phylogenies by  
738 using shifted stochastic processes on a tree. *Journal of the Royal Statistical Society:*  
739 *Series B (Statistical Methodology)*. .
- 740 Beaulieu JM, Jhwueng DC, Boettiger C, O’Meara BC. 2012. Modeling stabilizing  
741 selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution*.  
742 66:2369–2383.
- 743 Blomberg SP, Garland T, Ives AR. 2003. Testing for Phylogenetic Signal in Comparative  
744 Data: Behavioral Traits Are More Labile. *Evolution*. 57:717–745.
- 745 Butler MA, King AA. 2004. Phylogenetic Comparative Analysis: A Modeling Approach for  
746 Adaptive Evolution. *The American Naturalist*. 164:683–695.
- 747 Clavel J, Escarguel G, Merceron G. 2015. mvmorph : an r package for fitting multivariate  
748 evolutionary models to morphometric data. *Methods in Ecology and Evolution*.  
749 6:1311–1319.
- 750 Cook LM, Grant BS, Saccheri IJ, Mallet J. 2012. Selective bird predation on the peppered  
751 moth: the last experiment of Michael Majerus. *Biology Letters*. 8:609–612.
- 752 Cooper N, Thomas GH, Venditti C, Meade A, Freckleton RP. 2016. A cautionary note on  
753 the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biological Journal*  
754 *of the Linnean Society*. 118:64–77.

- 755 Cybis GB, Sinsheimer JS, Bedford T, Mather AE, Lemey P, Suchard MA. 2015. Assessing  
756 phenotypic correlation through the multivariate phylogenetic latent liability model. *The*  
757 *Annals of Applied Statistics*. 9:969–991.
- 758 Dempster A, Laird N, Rubin DB. 1977. Maximum likelihood from incomplete data via the  
759 EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*. 39:1–38.
- 760 Duchen P, Leuenberger C, Szilágyi SM, Harmon LJ, Eastman JM, Schweizer M, Wegmann  
761 D. 2017. Inference of Evolutionary Jumps in Large Phylogenies using Lévy Processes.  
762 *Systematic Biology*. 00:1–14.
- 763 Eastman JM, Alfaro ME, Joyce P, Hipp AL, Harmon LJ. 2011. A Novel comparative  
764 method for identifying shifts in the rate of character evolution on trees. *Evolution*.  
765 65:3578–3589.
- 766 Eastman JM, Wegmann D, Leuenberger C, Harmon LJ. 2013. Simpsonian 'Evolution by  
767 Jumps' in an Adaptive Radiation of Anolis Lizards. *ArXiv*. p. 1305.4216.
- 768 Eldredge N, Gould SJ. 1972. Punctuated equilibria: an alternative to phyletic gradualism.
- 769 Felsenstein J. 1985. Phylogenies and the Comparative Method. *The American Naturalist*.  
770 125:1–15.
- 771 Felsenstein J. 2004. Inferring Phylogenies.
- 772 Felsenstein J. 2005. Using the quantitative genetic threshold model for inferences between  
773 and within species. *Philosophical Transactions of the Royal Society B: Biological*  
774 *Sciences*. 360:1427–1434.
- 775 Felsenstein J. 2012. A Comparative Method for Both Discrete and Continuous Characters  
776 Using the Threshold Model. *The American Naturalist*. 179:145–156.



- 777 Fitzjohn RG. 2012. Diversitree: Comparative phylogenetic analyses of diversification in R.  
778 *Methods in Ecology and Evolution*. 3:1084–1092.
- 779 Hansen TF. 1997. Stabilizing Selection and the Comparative Analysis of Adaptation.  
780 *Evolution*. 51:1341.
- 781 Hansen TF, Orzack SH. 2005. Assessing Current Adaptation and Phylogenetic Inertia as  
782 Explanations of Trait Evolution: The Need for Controlled Comparisons. *Evolution*.  
783 59:2063–2072.
- 784 Hansen TF, Pienaar J, Orzack SH. 2008. A Comparative Method for Studying Adaptation  
785 to a Randomly Evolving Environment. *Evolution*. 62:1965–1977.
- 786 Ho LST, Ané C. 2013. A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait  
787 Evolution Models. *Systematic Biology*. 63:397–408.
- 788 Ho LST, Ané C. 2014. Intrinsic inference difficulties for trait evolution with  
789 Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution*. 5:1133–1146.
- 790 Hubert L, Arabie P. 1985. Comparing partitions. *Journal of Classification*. 2:193–218.
- 791 Ingram T, Mahler DL. 2013. SURFACE: Detecting convergent evolution from comparative  
792 data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion.  
793 *Methods in Ecology and Evolution*. 4:416–425.
- 794 Jetz W, Thomas G, Joy J, Hartmann K, Mooers A. 2012. The global diversity of birds in  
795 space and time. *Nature*. 491:444–448.
- 796 Khabbazian M, Kriebel R, Rohe K, Ané C. 2016. Fast and accurate detection of  
797 evolutionary shifts in Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution*.  
798 7:811–824.

- 799 Lande R. 1976. Natural Selection and Random Genetic Drift in Phenotypic Evolution.  
800 *Evolution*. 30:314.
- 801 Landis MJ, Schraiber JG, Liang M. 2013. Phylogenetic analysis using Lévy processes:  
802 Finding jumps in the evolution of continuous traits. *Systematic Biology*. 62:193–204.
- 803 Liow LH, Reitan T, Harnik PG. 2015. Ecological interactions on macroevolutionary time  
804 scales: clams and brachiopods are more than ships that pass in the night. *Ecology*  
805 *Letters*. 18:1030–1039.
- 806 Mahler DL, Ingram T, Revell LJ, Losos JB. 2013. Exceptional Convergence on the  
807 Macroevolutionary Landscape in Island Lizard Radiations. *Science*. 341:292–295.
- 808 Mahler DL, Revell LJ, Glor RE, Losos JB. 2010. Ecological opportunity and the rate of  
809 morphological evolution in the diversification of greater Antillean anoles. *Evolution*.  
810 64:2731–2745.
- 811 Manceau M, Lambert A, Morlon H. 2016. A unifying comparative phylogenetic framework  
812 including traits coevolving across interacting lineages. *Systematic Biology*. p. syw115.
- 813 Mardia KV, Kent JT, Bibby JM. 1979. Multivariate analysis. Probability and  
814 mathematical statistics. Academic Press.
- 815 Meredith RW, Janecka JE, Gatesy J, et al. (22 co-authors). 2011. Impacts of the  
816 Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification.  
817 *Science*. 334:521–524.
- 818 Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature*.  
819 401:877–884.

- 820 Pennell MW, FitzJohn RG, Cornwell WK, Harmon LJ. 2015. Model Adequacy and the  
821 Macroevolution of Angiosperm Functional Traits. *The American Naturalist*.  
822 186:E33–E50.
- 823 R Core Team. 2017. R: A Language and Environment for Statistical Computing. R  
824 Foundation for Statistical Computing, Vienna, Austria.
- 825 Reitan T, Schweder T, Henderiks J. 2012. Phenotypic evolution studied by layered  
826 stochastic differential equations. *The Annals of Applied Statistics*. 6:1531–1551.
- 827 Revell LJ. 2009. Size-correction and principal components for interspecific comparative  
828 studies. *Evolution*. 63:3258–3268.
- 829 Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller  
830 M, Maechler M. 2014. robustbase: Basic Robust Statistics.
- 831 Semple C, Steel M. 2003. Phylogenetics. Oxford University Press, oxford lec edition.
- 832 Simpson GG. 1944. Tempo and Mode in Evolution. A Wartime book. Columbia University  
833 Press.
- 834 Stadler T. 2011. Simulating Trees with a Fixed Number of Extant Species. *Systematic  
835 Biology*. 60:676–684.
- 836 Tibshirani R. 1996. Regression Selection and Shrinkage via the Lasso. *Journal of the Royal  
837 Statistical Society. Series B (Methodological)*. 58:267–288.
- 838 Uyeda JC, Caetano DS, Pennell MW. 2015. Comparative Analysis of Principal  
839 Components Can be Misleading. *Systematic Biology*. 64:677–689.
- 840 Uyeda JC, Harmon LJ. 2014. A Novel Bayesian Method for Inferring and Interpreting the

841 Dynamics of Adaptive Landscapes from Phylogenetic Comparative Data. *Systematic*  
842 *Biology*. 63:902–918.

843

## PCA: MATHEMATICAL DERIVATIONS

844 *Expectation of the estimated Variance-Covariance Matrix.*— Taking

845  $\tilde{\mathbf{C}} = (\mathbf{1}_n^T \mathbf{C}^{-1} \mathbf{1}_n)^{-1} \mathbf{1}_n^T \mathbf{C}^{-1}$ , we have that  $\bar{\mathbf{Y}}^T = \tilde{\mathbf{C}} \mathbf{Y}$ , and  $\bar{\mathbf{a}}^T = \mathbb{E} [\bar{\mathbf{Y}}^T] = \tilde{\mathbf{C}} \mathbf{a}$ . Denote by

846  $\mathbf{N}_{\mathbf{C}^{-1}} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^{p^2}$  the function that to a  $n \times p$  matrix  $\mathbf{A}$  associates the  $p \times p$  matrix

847  $\mathbf{A}^T \mathbf{C}^{-1} \mathbf{A}$ . We get:

$$\begin{aligned} (n-1) \mathbb{E} [\hat{\mathbf{R}}] &= \mathbb{E} [\mathbf{N}_{\mathbf{C}^{-1}} (\mathbf{Y} - \mathbf{1}_n \bar{\mathbf{Y}}^T)] = \mathbb{E} [\mathbf{N}_{\mathbf{C}^{-1}} ((\mathbf{Y} - \mathbf{a}) + (\mathbf{a} - \mathbf{1}_n \bar{\mathbf{a}}^T) + (\mathbf{1}_n \bar{\mathbf{a}}^T - \mathbf{1}_n \bar{\mathbf{Y}}^T))] \\ &= \mathbb{E} [\mathbf{N}_{\mathbf{C}^{-1}} ((\mathbf{I} - \mathbf{1}_n \tilde{\mathbf{C}})(\mathbf{Y} - \mathbf{a}) + (\mathbf{a} - \mathbf{1}_n \bar{\mathbf{a}}^T))] \\ &= \mathbb{E} [\mathbf{N}_{\mathbf{C}^{-1}} ((\mathbf{I} - \mathbf{1}_n \tilde{\mathbf{C}})(\mathbf{Y} - \mathbf{a}))] + \mathbf{N}_{\mathbf{C}^{-1}} (\mathbf{a} - \mathbf{1}_n \bar{\mathbf{a}}^T) \end{aligned}$$

848 where the two double products cancel out, as  $\mathbb{E} [\mathbf{Y}] = \mathbf{a}$ . But, for any non-singular

849 symmetric matrix  $\mathbf{H}$ , we have:

$$\begin{aligned} \mathbb{E} [(\mathbf{Y} - \mathbf{a})^T \mathbf{H}^{-1} (\mathbf{Y} - \mathbf{a})] &= \sum_{1 \leq i, j \leq n} [\mathbf{H}^{-1}]_{ij} \mathbb{E} [(\mathbf{Y}^i - \mathbf{a}^i)(\mathbf{Y}^j - \mathbf{a}^j)^T] \\ &= \sum_{1 \leq i, j \leq n} [\mathbf{H}^{-1}]_{ij} C_{ij} \mathbf{R} = \text{tr}(\mathbf{H}^{-1} \mathbf{C}) \mathbf{R} \end{aligned}$$

850 Hence, applying this formula with  $\mathbf{H}^{-1} = (\mathbf{I} - \mathbf{1}_n \tilde{\mathbf{C}})^T \mathbf{C}^{-1} (\mathbf{I} - \mathbf{1}_n \tilde{\mathbf{C}}) = \mathbf{C}^{-1} - \mathbf{C}^{-1} \mathbf{1}_n \tilde{\mathbf{C}}$ ,

851 some straightforward matrix algebra manipulations give us:

$$(n-1) \mathbb{E} [\hat{\mathbf{R}}] = (n-1) \mathbf{R} + (\mathbf{a} - \mathbf{1}_n \bar{\mathbf{a}}^T)^T \mathbf{C}^{-1} (\mathbf{a} - \mathbf{1}_n \bar{\mathbf{a}}^T)$$

852 which is the result stated in the text, with  $\mathbf{G} = \mathbf{a} - \mathbf{1}_n \bar{\mathbf{a}}^T = (\mathbf{I}_n - (\mathbf{1}_n^T \mathbf{C}^{-1} \mathbf{1}_n)^{-1} \mathbf{1}_n \mathbf{1}_n^T \mathbf{C}^{-1}) \mathbf{a}$ .

## PhylogeneticEM PACKAGE CASE STUDY: NEW WORLD MONKEYS

In this section, we demonstrate the basic use of the R package PhylogeneticEM for the analysis of the New World Monkeys dataset (Aristide et al. 2016).

### *Loading and Plotting the data*

The data have been embedded in the R package PhylogeneticEM, to be loaded easily. The traits can be plotted on the tree thanks to the function `plot` applied to a void `params_process` object with dimension 2 (Fig. 13).

```
library(PhylogeneticEM)
data(monkeys)

plot(params_BM(p=2), data = monkeys$dat,
      phylo = monkeys$phy, show.tip.label = TRUE)
```

This `plot` function inherits from most of the optional arguments of the popular `ape` `plot` function (here for instance, the optional argument `show.tip.label` is used). Many other graphical parameters can be set by the user, so as to control the output of the function. All the results showed in the main text were produced by the package's plotting function. The two traits are represented on the right, each with its own scale. Plotting the data on the tree before analyzing it allows us to spot potential errors or outliers.

### *Analyzing the data*

The automatic shift detection is done using function `PhyloEM`. We show below how the function can be called, using an `scOU` process (with stationary root, the default), for a maximum number of shifts equal to 10, on an automatically chosen grid with 4 values for

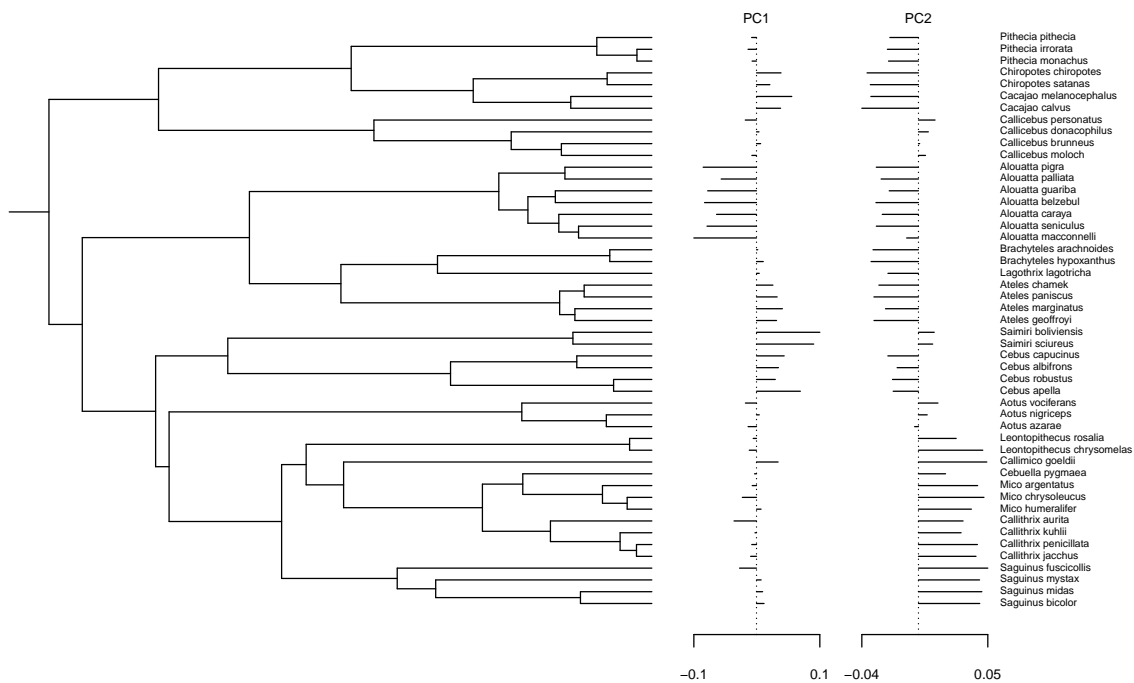


Figure 13: New World Monkey dataset as plotted in PhylogeneticEM

871 the selection strength  $\alpha$ , and parallelized on 2 cores. These parameters were chosen only to  
 872 demonstrate the function, for this example analysis would run in about one minute.  
 873 Different parameters were used to obtain the results below and in the main text. There are  
 874 many more options available to guide the analysis, all described in the manual entry of the  
 875 function.

```
res <- PhyloEM(Y_data = monkeys$dat,      ## data
               phylo = monkeys$phy,      ## phylogeny
               process = "scOU",        ## scalar OU
               K_max = 10,               ## maximal number of shifts
               nbr_alpha = 4,            ## number of alpha values
               parallel_alpha = TRUE,    ## parallelize on alpha values
               Ncores = 2)              ## number of computing cores
```

876 The result is stored in an object of class PhyloEM, which has several extractors

877 available (see manual). By default, the plot function draws the maximum likelihood  
878 function selected by the method (Fig. 14). The same optional parameters can be used as  
879 before to control how the figure should look like.

```
plot(res, edge.width = 2, show.tip.label = TRUE)
```

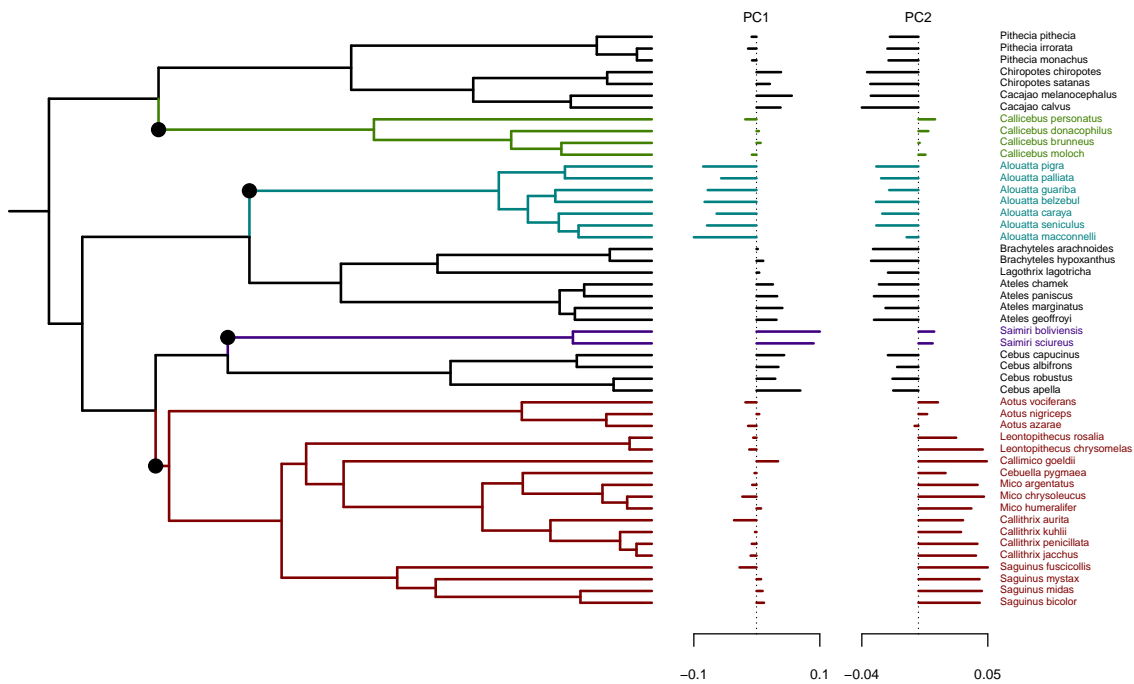


Figure 14: Maximum likelihood solution with 4 shifts selected by the method.

880 The solution showed in the main text (Fig. 10) has 5 shifts, instead of 4. It can be  
881 plotted using the extractor `params_process`, which extracts some inferred parameters from  
882 an object of class `PhyloEM`.

```
params_5 <- params_process(res, K = 5)  
plot(res, params = params_5)
```



883

## Plotting Equivalent Solutions

884

885

886

887

888

889

The previous call actually results in a warning being issued: “Warning in `params_process.PhyloEM(res, K = 5)`: There are several equivalent solutions for this shift position.” Indeed, as mentioned in the main text, the solution with 5 shifts has three equivalent shift allocations on the branches. These solutions can be found and plotted thanks to the function `equivalent_shifts`, that returns an object that can be visualized (Fig. 15).

```
eq_shifts <- equivalent_shifts(monkeys$phy, params_5)
plot(eq_shifts, show_shifts_values = FALSE, shifts_cex = 0.5)
```

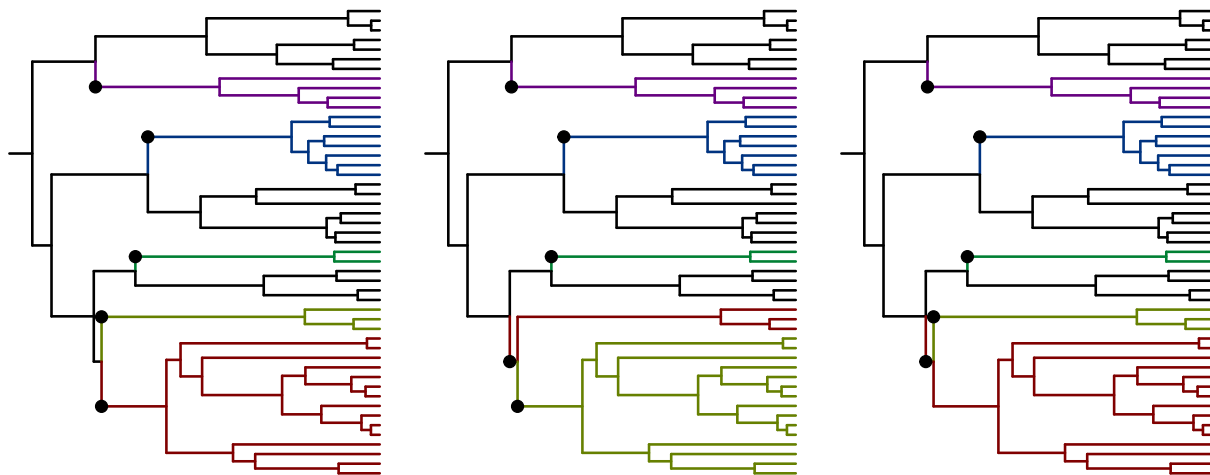


Figure 15: The three equivalent maximum likelihood shift allocations for the solution with 5 shifts.

890

891

892

893

894

By default, the shifts values for the first trait is showed for all equivalent solutions. Black is always reserved to the “ancestral state”, and the value  $\lambda = \beta_0 = \mu$  of the ancestral optimal value is shown at the root. Here, the three equivalent solutions are quite straightforward, as one configuration has two shifts on sister edges. Note that the clustering of the species at the tips of the tree remains unchanged, while the historic

895 scenario of the adaptive shifts is slightly altered. This ambiguity is inherent to the data.  
896 More information to resolve this ambiguity can only come from a prior distribution on shift  
897 values, or ideally from fossil data sampled in the right region of the tree.

898

## EM INFERENCE

899

This section provides the update formulas for the EM algorithm in Section

900

Statistical Inference. Throughout this section, the superscript  $h$  refers to the current

901

iteration index, e.g.  $\boldsymbol{\theta}^{(h)}$  stands for the vector of parameters estimate at iteration  $h$ :

902

$\boldsymbol{\theta}^{(h)} = (\boldsymbol{\mu}^{(h)}, \boldsymbol{\Delta}^{(h)}, \mathbf{R}^{(h)}, \boldsymbol{\Gamma}^{(h)})$ . We denote further by  $\mathbf{X}$  the  $N \times p$  matrix of the traits at all

903

the nodes of the tree, that contains both  $\mathbf{Z}$  and  $\mathbf{Y}$ . In these derivations, nodes are

904

numbered in a preorder, such that the root comes first:  $\rho = 1$ , the internal nodes are

905

numbered from 1 to  $m$ , and the tips from  $m + 1$  to  $N = m + n$ .

906

*Conditional expectation of the complete likelihood.*— The EM algorithm mainly deals with

907

$\mathbb{E}[\log p_{\boldsymbol{\theta}}(\mathbf{X}) \mid \mathbf{Y}^d]$ , where  $\mathbf{Y}^d$  is the vector of the observed tips data (that might be missing

908

some values). In our case we have that

$$\begin{aligned}
 -2\mathbb{E}[\log p_{\boldsymbol{\theta}}(\mathbf{X}) \mid \mathbf{Y}^d] &= p(m+n) \log 2\pi + p \sum_{j=2}^{m+n} \log \ell_j \\
 &\quad + \log |\boldsymbol{\Gamma}| + \text{tr} \{ \boldsymbol{\Gamma}^{-1} \text{Var} [\mathbf{X}^1 \mid \mathbf{Y}^d] \} + \| \mathbb{E} [\mathbf{X}^1 \mid \mathbf{Y}^d] - \boldsymbol{\mu} \|_{\boldsymbol{\Gamma}^{-1}}^2 \\
 &\quad + (m+n-1) \log |\mathbf{R}| + \sum_{j=2}^{m+n} \ell_j^{-1} \text{tr} \{ \mathbf{R}^{-1} \text{Var} [\mathbf{X}^j - \mathbf{X}^{\text{pa}(j)} \mid \mathbf{Y}^d] \} \\
 &\quad + \sum_{j=2}^{m+n} \ell_j^{-1} \| \mathbb{E} [\mathbf{X}^j - \mathbf{X}^{\text{pa}(j)} \mid \mathbf{Y}^d] - \boldsymbol{\Delta}^j \|_{\mathbf{R}^{-1}}^2. \tag{5}
 \end{aligned}$$

909

### *M step*

910

At the M step, the parameters are updated as the minimizers of (5) evaluated with

911

the conditional moments of the hidden variables given  $\mathbf{Y}^d$ . We get the following updates.

912

*Root Parameters.*—

$$\boldsymbol{\mu}^{(h+1)} = \mathbb{E}^{(h)} [\mathbf{X}^1 \mid \mathbf{Y}^d], \quad \boldsymbol{\Gamma}^{(h+1)} = \text{Var}^{(h)} [\mathbf{X}^1 \mid \mathbf{Y}^d]. \tag{6}$$

913 where the conditional moments are obtained as part of the E step, see Equation (8).  
 914 Notations  $\mathbb{E}^{(h)}$  and  $\mathbb{V}\text{ar}^{(h)}$  denote the moments taken with the law defined by current  
 915 parameters  $\boldsymbol{\theta}^{(h)}$ .

916 *Rate Matrix.*—

$$\begin{aligned} (m+n-1)\mathbf{R}^{(h+1)} &= \sum_{j=2}^{m+n} \ell_j^{-1} \mathbb{V}\text{ar}^{(h)} [\mathbf{X}^j - \mathbf{X}^{\text{pa}(j)} \mid \mathbf{Y}^d] \\ &\quad + \ell_j^{-1} (\mathbb{E}^{(h)} [\mathbf{X}^j - \mathbf{X}^{\text{pa}(j)} \mid \mathbf{Y}^d] - \boldsymbol{\Delta}^{(h+1)j}) \\ &\quad \cdot (\mathbb{E}^{(h)} [\mathbf{X}^j - \mathbf{X}^{\text{pa}(j)} \mid \mathbf{Y}^d] - \boldsymbol{\Delta}^{(h+1)j})^T. \end{aligned} \quad (7)$$

917 *Optimal Shift Location.*— Only the last term of (5) depends on the shifts so we have to  
 918 minimize the sum of costs to find  $\boldsymbol{\Delta}^{(h+1)}$ :

$$\begin{aligned} C^{(h)}(\boldsymbol{\Delta}) &= \sum_{j=2}^{m+n} C_j^{(h)}(\boldsymbol{\Delta}) \\ \text{with } C_j^{(h)}(\boldsymbol{\Delta}) &= \ell_j^{-1} \left\| \mathbb{E}^{(h)} [\mathbf{X}^j - \mathbf{X}^{\text{pa}(j)} \mid \mathbf{Y}^d] - \boldsymbol{\Delta}^j \right\|_{(\mathbf{R}^{(h)})^{-1}}^2. \end{aligned}$$

919 This minimization can be achieved using the same algorithm as in the univariate case  
 920 (Bastide et al. 2016) to get the optimal shifts allocations and values. Said algorithm  
 921 essentially sorts the branches in decreasing order of  $C_j^{(h)}(\boldsymbol{\Delta})$  and assigns shifts to the first  
 922  $K$  branches.

### 923 *E step*

924 The aim of the E step is to compute the moments of the completed dataset given  
 925 the observed traits at the tips, namely:

$$\mathbf{E}_j = \mathbb{E} [\mathbf{X}^j \mid \mathbf{Y}^d], \quad \mathbf{V}_j = \mathbb{V}\text{ar} [\mathbf{X}^j \mid \mathbf{Y}^d], \quad \mathbf{C}_{j,\text{pa}(j)} = \mathbb{C}\text{ov} [\mathbf{X}^j; \mathbf{X}^{\text{pa}(j)} \mid \mathbf{Y}^d] \quad (8)$$

926 where we dropped the dependency in  $\boldsymbol{\theta}^{(h)}$  for the sake of legibility, but all these moments  
 927 are indeed taken with the laws given by the current parameters. We do so thanks to an  
 928 upward-downward recursion on the tree, as described below. This algorithm can apply to a  
 929 broad classes of Gaussian processes, provided that the moments of the traits at a child  
 930 node are of the form:

$$\forall j \in \llbracket 2, m+n \rrbracket, \begin{cases} \mathbb{E} [\mathbf{X}^j \mid \mathbf{X}^{\text{pa}(j)}] = m_j(\mathbf{X}^{\text{pa}(j)}) = \mathbf{Q}_j \mathbf{X}^{\text{pa}(j)} + \mathbf{r}_j \\ \text{Var} [\mathbf{X}^j \mid \mathbf{X}^{\text{pa}(j)}] = \boldsymbol{\Sigma}_j \end{cases}$$

931 For a BM, we get

$$\mathbf{Q}_j = \mathbf{I}_p, \quad \mathbf{r}_j = \boldsymbol{\Delta}^j \quad \text{and} \quad \boldsymbol{\Sigma}_j = \ell_j \mathbf{R}.$$

932 A multivariate OU could also be handled, with:

$$\mathbf{Q}_j = e^{-\mathbf{A}\ell_j}, \quad \mathbf{r}_j = (\mathbf{I}_p - e^{-\mathbf{A}\ell_j})\boldsymbol{\beta}^j \quad \text{and} \quad \boldsymbol{\Sigma}_j = \boldsymbol{\Gamma} - e^{-\mathbf{A}\ell_j}\boldsymbol{\Gamma}e^{-\mathbf{A}^T\ell_j}.$$

933 Although we do not use these last formulas here (thanks to the equivalence between OU  
 934 and BM in our setting), they are implemented in **PhylogeneticEM**, and could be readily  
 935 used in an extension of the method to non-ultrametric trees with fossil taxa. To properly  
 936 handle missing data in a unified framework, we first re-define *ad hoc* inversion and  
 937 determinant operations that allow us to easily write the degenerated Gaussian likelihood  
 938 that appears along the way.

939 *Missing data.*— For a multivariate trait observed at node  $i$ , define the application  
 940  $f_{d_i} : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^{d_i \times d_i}$  that, given a matrix, returns the matrix with only rows and columns  
 941 corresponding to observed traits. Define also the “pseudo-inverse”  $f_{d_i}^{-1} : \mathbb{R}^{d_i \times d_i} \rightarrow \mathbb{R}^{p \times p}$   
 942 that put the observed traits back into their places, and fills the un-defined lines and

943 columns with zeros. This allows us to define a “low-dimensional inverse” as:

$$[\mathbf{S}]_{\text{ld}}^{-1} = f_{d_i}^{-1} ([f_{d_i}(\mathbf{S})]^{-1}), \quad \forall \mathbf{S} \in \mathbb{R}^{p \times p}$$

944 for all  $\mathbf{S}$  such that  $f_{d_i}(\mathbf{S})$  is invertible. We also define a “low dimensional determinant”, as:

$$|[\mathbf{S}]_{\text{ld}}^{-1}| = |[f_{d_i}(\mathbf{S})]^{-1}|, \quad \forall \mathbf{S} \in \mathbb{R}^{p \times p}.$$

945 These conventions amount to taking infinite values for the variance-covariance terms  
946 of non-observed traits. This allows us to write the following:

$$(2\pi)^{(p-d)/2} \Phi_{\mathbf{m}, \mathbf{S}}(\mathbf{x}) = \Phi_{f_d(\mathbf{m}), f_d(\mathbf{S})}(f_d(\mathbf{x})).$$

947 where  $\Phi_{\mathbf{m}, \mathbf{S}}$  denotes the density of a multivariate Gaussian, with expectation vector  $\mathbf{m}$  and  
948 variance matrix  $\mathbf{S}$ . That is, we write the density of a  $d$ -dimensional Gaussian as the density  
949 of a  $p$ -dimensional one, but with the exact same likelihood value, up to a normalizing  
950 constant  $(2\pi)^{(p-d)/2}$ . If  $d = 0$  (no data at one tip), then  $[\mathbf{S}]_{\text{ld}}^{-1}$  is a matrix of 0, and we take  
951 by convention  $|[\mathbf{S}]_{\text{ld}}^{-1}| = 1$ , so that  $\Phi_{f_d(\mathbf{m}), f_d(\mathbf{S})}(f_d(\mathbf{x})) = 1$ .

952 *Upward recursion.*— For a given node  $j$  in the tree, we denote by  ${}^j\mathbf{Y}^d$  the set of all traits  
953 observed at all the tips below node  $j$ . The aim of the upward recursion is to compute the  
954 Gaussian pdf  $f_{j\mathbf{Y}^d|\mathbf{X}^j}({}^j\mathbf{Y}^d; \mathbf{a})$  of  ${}^j\mathbf{Y}^d | \mathbf{X}^j$ , which we write as proportional to a Gaussian  
955 density in  $\mathbf{a}$ :

$$f_{j\mathbf{Y}^d|\mathbf{X}^j}({}^j\mathbf{Y}^d; \mathbf{a}) = A_j({}^j\mathbf{Y}^d) \Phi_{M_j({}^j\mathbf{Y}^d), S_j({}^j\mathbf{Y}^d)}(\mathbf{a}).$$

956 **Initialization:** For each tip  $i$ , the observed values  $(\mathbf{Y}^d)^i$  given the vector of values  $\mathbf{Y}^i$

957

follow a Dirac distribution:

$$\forall i \in \llbracket 1, n \rrbracket, f_{(\mathbf{Y}^d)^i | \mathbf{Y}^i} ((\mathbf{Y}^d)^i; \mathbf{a}) = \delta_{(\mathbf{Y}^d)^i}(\mathbf{a}).$$

958

We can express this in the correct format:

$$\forall i \in \llbracket 1, n \rrbracket, f_{(\mathbf{Y}^d)^i | \mathbf{Y}^i} ((\mathbf{Y}^d)^i; \mathbf{a}) = (2\pi)^{(p-d)/2} \Phi_{\mathbf{Y}^i, \mathbf{0}}(\mathbf{a})$$

959

but taking the “low dimensional” inverses and determinants defined above.

960

**Propagation:** The upward recursion formulas result from the standard properties of the

961

conditional distribution of a multivariate Gaussian distribution plus the fact that  $L$

962

daughters of a given node  $\mathbf{X}^j$  are conditionally independent so

$$f_{j \mathbf{Y}^d | \mathbf{X}^j} ({}^j \mathbf{Y}^d; \mathbf{a}) = \prod_{\ell=1}^L f_{j_\ell \mathbf{Y}^d | \mathbf{X}^j} ({}^{j_\ell} \mathbf{Y}^d; \mathbf{a}).$$

963

We get

$$\left\{ \begin{array}{l} S_j({}^j \mathbf{Y}^d) = \left( \sum_{\ell=1}^L \mathbf{Q}_{j_\ell}^T (S_{j_\ell}({}^{j_\ell} \mathbf{Y}^d) + \Sigma_{j_\ell})^{-1} \mathbf{Q}_{j_\ell} \right)^{-1} \\ M_j({}^j \mathbf{Y}^d) = S_j({}^j \mathbf{Y}^d) \sum_{\ell=1}^L \mathbf{Q}_{j_\ell}^T (S_{j_\ell}({}^{j_\ell} \mathbf{Y}^d) + \Sigma_{j_\ell})^{-1} (M_{j_\ell}({}^{j_\ell} \mathbf{Y}^d) - \mathbf{r}_{j_\ell}) \\ \log A_j({}^j \mathbf{Y}^d) = -\frac{(L-1)p}{2} \log(2\pi) + \frac{1}{2} \log |S_j({}^j \mathbf{Y}^d)| \\ \quad + \sum_{\ell=1}^L \log A_{j_\ell}({}^{j_\ell} \mathbf{Y}^d) - \frac{1}{2} \log |S_{j_\ell}({}^{j_\ell} \mathbf{Y}^d) + \Sigma_{j_\ell}| \\ \quad - \frac{1}{2} \sum_{\ell=1}^L (M_{j_\ell}({}^{j_\ell} \mathbf{Y}^d) - \mathbf{r}_{j_\ell})^T (S_{j_\ell}({}^{j_\ell} \mathbf{Y}^d) + \Sigma_{j_\ell})^{-1} (M_{j_\ell}({}^{j_\ell} \mathbf{Y}^d) - \mathbf{r}_{j_\ell}) \\ \quad + \frac{1}{2} M_j({}^j \mathbf{Y}^d)^T S_j({}^j \mathbf{Y}^d)^{-1} M_j({}^j \mathbf{Y}^d) \end{array} \right.$$

964 where we keep track of the log of the constant  $A_j$ , for numerical accuracy. Remark  
965 that we only need to handle the infinite terms properly as described above, using the  
966 “low dimensional” inverses and determinants when needed. These terms will  
967 disappear as we go up to a node that has at least one tip with some observation for  
968 this particular trait. In the pathological case where a trait is never observed, the  
969 corresponding term remains infinite throughout the recursion, and hence does not  
970 bring any information as to the value of that trait, and does not change the  
971 likelihood. The variance of a root non-observed trait is then just the one put a priori  
972 in  $\Gamma$  (see below).

973 **Root node and likelihood:** Once at the root, we have  $f_{\mathbf{Y}^d|\mathbf{X}^1}(\mathbf{Y}^d; \mathbf{a})$ , which is the  
974 likelihood of the observations given the root state  $\mathbf{X}^1 = \mathbf{a}$ , and we write:

$$f_{\mathbf{X}^1|\mathbf{Y}^d}(\mathbf{a}; \mathbf{Y}^d) \propto f_{\mathbf{Y}^d|\mathbf{X}^1}(\mathbf{Y}^d; \mathbf{a})f_{\mathbf{X}^1}(\mathbf{a})$$

975 which gives

$$\begin{cases} \text{Var} [\mathbf{X}^1 | \mathbf{Y}^d] = (\Gamma^{-1} + S_1(\mathbf{Y}^d)^{-1})^{-1} \\ \mathbb{E} [\mathbf{X}^1 | \mathbf{Y}^d] = \text{Var} [X_1 | \mathbf{Y}^d] (\Gamma^{-1}\boldsymbol{\mu} + S_1(\mathbf{Y}^d)^{-1}M_1(\mathbf{Y})). \end{cases}$$

976 *Downward recursion.*— We now derive a recursion that goes from the root back to the tips  
977 to compute the conditional moments required to evaluate (5). Going down the tree, we  
978 need to compute, for each node  $X_j$ ,  $2 \leq j \leq m$ ,  $\mathbf{E}_j$ ,  $\mathbf{V}_j$  and  $\mathbf{C}_{j,\text{pa}(j)}$  as in (8). (additionally  
979 conditioning on  $\mathbf{X}^1$  if the root is fixed).

980 **Initialization:** The initialization of the downward is given by the last step of the upward.



981 If the root is random, we have

$$\left\{ \begin{array}{l} \mathbf{V}_1 = \mathbb{V}\text{ar} [\mathbf{X}^1 | \mathbf{Y}^d] = (\mathbf{\Gamma}^{-1} + S_1(\mathbf{Y}^d)^{-1})^{-1} \\ \mathbf{E}_1 = \mathbb{E} [\mathbf{X}^1 | \mathbf{Y}^d] = \mathbb{V}\text{ar} [X_1 | \mathbf{Y}^d] (\mathbf{\Gamma}^{-1} \boldsymbol{\mu} + S_1(\mathbf{Y}^d)^{-1} M_1(\mathbf{Y})) \\ \mathbf{C}_{1,\text{pa}(1)} = \text{NA} \end{array} \right.$$

982 whereas, if we work conditionally to the root, we have  $\mathbf{V}_1 = \mathbb{V}\text{ar} [\mathbf{X}^1 | \mathbf{Y}^d, \mathbf{X}^1] = \mathbf{0}$ ,

983  $\mathbf{E}_1 = \mathbb{E} [\mathbf{X}^1 | \mathbf{Y}^d, \mathbf{X}^1] = \boldsymbol{\mu}$  and  $\mathbf{C}_{1,\text{pa}(1)} = \text{NA}$ .

984 **Propagation:** We have

$$f_{\mathbf{X}^{\text{pa}(j)}, \mathbf{X}^j | \mathbf{Y}^d}(\mathbf{a}, \mathbf{b}; \mathbf{Y}^d) = f_{\mathbf{X}^{\text{pa}(j)} | \mathbf{Y}^d}(\mathbf{a}; \mathbf{Y}) f_{\mathbf{X}^j | \mathbf{X}^{\text{pa}(j)}, \mathbf{Y}^d}(\mathbf{b}; \mathbf{a}, \mathbf{Y}^d)$$

985 We know the first term from the recurrence, and we can compute the second term

986 thanks to the upward step:

$$f_{\mathbf{X}^j | \mathbf{X}^{\text{pa}(j)}, \mathbf{Y}^d}(\mathbf{b}; \mathbf{a}, \mathbf{Y}^d) = f_{\mathbf{X}^j | \mathbf{X}^{\text{pa}(j)}, j \mathbf{Y}^d}(\mathbf{b}; \mathbf{a}, j \mathbf{Y}^d) \propto f_{\mathbf{X}^j | \mathbf{X}^{\text{pa}(j)}}(\mathbf{b}; \mathbf{a}) f_{j \mathbf{Y}^d | \mathbf{X}^j}(j \mathbf{Y}^d; \mathbf{b})$$

987 As  $j \mathbf{Y}^d | \mathbf{X}^j \sim \mathcal{N}(M_j(j \mathbf{Y}^d), S_j(j \mathbf{Y}^d))$  and  $\mathbf{X}^j | \mathbf{X}^{\text{pa}(j)} \sim \mathcal{N}(m_j(\mathbf{X}^{\text{pa}(j)}), \Sigma_j)$ , we get

$$\mathbf{X}^j | \mathbf{X}^{\text{pa}(j)}, \mathbf{Y}^d \sim \mathcal{N}(\bar{m}_j(\mathbf{X}^{\text{pa}(j)}), \bar{\Sigma}_j)$$

988

with

$$\left\{ \begin{array}{l} \bar{\Sigma}_j = (S_j(j\mathbf{Y}^d)^{-1} + \Sigma_j^{-1})^{-1} \\ \quad = S_j(j\mathbf{Y}^d) (S_j(j\mathbf{Y}^d) + \Sigma_j)^{-1} \Sigma_j = \Sigma_j (S_j(j\mathbf{Y}^d) + \Sigma_j)^{-1} S_j(j\mathbf{Y}^d) \\ \bar{m}_j(\mathbf{X}^{\text{pa}(j)}) = \bar{\Sigma}_j (S_j(j\mathbf{Y}^d)^{-1} M_j(j\mathbf{Y}^d) + \Sigma_j^{-1} m_j(\mathbf{X}^{\text{pa}(j)})) \\ \quad = \underbrace{S_j(j\mathbf{Y}^d) (S_j(j\mathbf{Y}^d) + \Sigma_j)^{-1} \mathbf{Q}_j}_{\mathbf{Q}_j} \mathbf{X}^{\text{pa}(j)} \\ \quad \quad + \underbrace{S_j(j\mathbf{Y}^d) (S_j(j\mathbf{Y}^d) + \Sigma_j)^{-1} \mathbf{r}_j + \Sigma_j (S_j(j\mathbf{Y}^d) + \Sigma_j)^{-1} M_j(j\mathbf{Y}^d)}_{\bar{\mathbf{r}}_j} \end{array} \right.$$

989

Hence:

$$f_{\mathbf{X}^j | \mathbf{X}^{\text{pa}(j)}, \mathbf{Y}^d}(\mathbf{b}; \mathbf{a}, \mathbf{Y}^d) \propto \exp\left(-\frac{1}{2}(\mathbf{b} - \bar{m}_j(\mathbf{a}))^T \bar{\Sigma}_j^{-1} (\mathbf{b} - \bar{m}_j(\mathbf{a}))\right)$$

990

And, as  $\begin{pmatrix} \mathbf{X}^j \\ \mathbf{X}^{\text{pa}(j)} \end{pmatrix} \Big|_{j\mathbf{Y}^d} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{E}_j \\ \mathbf{E}_{\text{pa}(j)} \end{pmatrix}, \begin{pmatrix} \mathbf{V}_j & \mathbf{C}_{j,\text{pa}(j)} \\ \mathbf{C}_{j,\text{pa}(j)}^T & \mathbf{V}_{\text{pa}(j)} \end{pmatrix}\right)$ , by Gaussian

991

conditioning, we get, for any  $\mathbf{a}$ :

$$\begin{cases} \bar{m}_j(\mathbf{a}) = \mathbf{E}_j + \mathbf{C}_{j,\text{pa}(j)} \mathbf{V}_{\text{pa}(j)}^{-1} (\mathbf{a} - \mathbf{E}_{\text{pa}(j)}) \\ \bar{\Sigma}_j = \mathbf{V}_j - \mathbf{C}_{j,\text{pa}(j)} \mathbf{V}_{\text{pa}(j)}^{-1} \mathbf{C}_{j,\text{pa}(j)}^T \end{cases}$$

992

From this we get:

$$\mathbf{C}_{j,\text{pa}(j)} = \bar{\mathbf{Q}}_j \mathbf{V}_{\text{pa}(j)}, \quad \mathbf{E}_j = \bar{\mathbf{r}}_j + \bar{\mathbf{Q}}_j \mathbf{E}_{\text{pa}(j)}, \quad \mathbf{V}_j = \bar{\Sigma}_j + \bar{\mathbf{Q}}_j \mathbf{V}_{\text{pa}(j)} \bar{\mathbf{Q}}_j^T.$$

993 And, finally:

$$\left\{ \begin{array}{l} \mathbf{C}_{j,\text{pa}(j)} = S_j(j\mathbf{Y}^d) (S_j(j\mathbf{Y}^d) + \Sigma_j)^{-1} \mathbf{Q}_j \mathbf{V}_{\text{pa}(j)} \\ \mathbf{E}_j = S_j(j\mathbf{Y}^d) (S_j(j\mathbf{Y}^d) + \Sigma_j)^{-1} (\mathbf{Q}_j \mathbf{E}_{\text{pa}(j)} + \mathbf{r}_j) + \Sigma_j (S_j(j\mathbf{Y}^d) + \Sigma_j)^{-1} M_j(j\mathbf{Y}^d) \\ \mathbf{V}_j = S_j(j\mathbf{Y}^d) (S_j(j\mathbf{Y}^d) + \Sigma_j)^{-1} \left( \Sigma_j + \mathbf{Q}_j \mathbf{V}_{\text{pa}(j)} \mathbf{Q}_j^T (S_j(j\mathbf{Y}^d) + \Sigma_j)^{-1} S_j(j\mathbf{Y}^d) \right) \end{array} \right.$$

994 **Missing Data:** In presence of missing data, the downward formulas read

$$\left\{ \begin{array}{l} \mathbf{C}_{j,\text{pa}(j)} = \bar{\Sigma}_j \Sigma_j^{-1} \mathbf{Q}_j \mathbf{V}_{\text{pa}(j)} \\ \mathbf{E}_j = \bar{\Sigma}_j \Sigma_j^{-1} (\mathbf{Q}_j \mathbf{E}_{\text{pa}(j)} + \mathbf{r}_j) + \bar{\Sigma}_j S_j(j\mathbf{Y}^d)^{-1} M_j(j\mathbf{Y}^d) \\ \mathbf{V}_j = \bar{\Sigma}_j (\mathbf{I}_p + \Sigma_j^{-1} \mathbf{Q}_j \mathbf{V}_{\text{pa}(j)} \mathbf{Q}_j^T \Sigma_j^{-1} \bar{\Sigma}_j) \end{array} \right.$$

995 where  $\bar{\Sigma}_j^{-1} = S_j(j\mathbf{Y}^d)^{-1} + \Sigma_j^{-1}$  can be is computed using the “low dimensional  
996 inverse” defined earlier for  $S_j(j\mathbf{Y}^d)$ , if needed.

997 Remark that theses formulas involve the inversion of two matrices ( $\Sigma_j$  and  $\bar{\Sigma}_j^{-1}$ ), each of  
998 dimension  $p$  (typically small), which is not computationally intensive.

### 999 *EM Initialization*

1000 Because it is only guaranteed to converge to a local optimum, the EM algorithm is  
1001 highly sensitive to its starting point. As consequence, it needs to be provided with good  
1002 initial guesses for the shifts positions and value, as well as the variance matrix  $\mathbf{R}$ . Initial  
1003 values are determined as follows:

- 1004 1. Do a lasso regression, assuming all traits are independent, choosing a penalty so that  
1005  $K$  shifts are found.
- 1006 2. Find the groups of tips created by those shifts, and center each group by its empirical  
1007 mean.

- 1008 3. Use the centered data to estimate an empirical variance matrix. This is done using  
1009 the Minimum Covariance Determinant (MCD) method, with function `covMcd` from  
1010 package `robustbase` (Rousseeuw et al. 2014).
- 1011 4. Use this estimated matrix to correct for correlations, before running a lasso again.
- 1012 5. For this second lasso, choose a penalty that selects for  $K + K_{\text{lag}}$  shifts, with  $K_{\text{lag}}$  a  
1013 fixed value (default to 5). Then, using a Gauss-lasso procedure, select the best  $K$   
1014 shifts (in term of log-likelihood) among those.

1015 This last step can be combinatorially intensive. To keep it fast, we bound the number of  
1016 trials. It has proven to enhance the results of the algorithm substantially.

### 1017 *Grid on $\alpha$*

1018 The inference presented above works for the rescaled BM, when the parameter  $\alpha$  is  
1019 supposed to be known. In practice, this parameter needs to be estimated. One simple way  
1020 to do that is to use a grid on  $\alpha$ . For each value on the grid, one can find an associated  
1021 estimator, and then find the maximum likelihood estimator of the parameters by taking  
1022 the best likelihood, for each number of shifts  $K$ . For instance, we plot below (Fig. 16) the  
1023 likelihood profile in  $K$  for 30  $\alpha$  values on a grid, for the New World Monkey dataset  
1024 (Aristide et al. 2016).

1025 This grid of  $\alpha$  values can be provided by the user, depending on some *a priori*  
1026 knowledge she might have of the problem at hand. If no grid is provided, one is  
1027 automatically computed, with  $n_\alpha$  values, evenly spaced on a log scale ranging between  $\alpha_{\min}$   
1028 and  $\alpha_{\max}$ . Those extrema values are chosen in the following way.

1029  $\alpha_{\min}$  The minimum value is chosen so that the maximum phylogenetic half-life  
1030 ( $t_{1/2} = \ln(2)/\alpha$ ) is equal to  $A \ln(2)h$ , where  $h$  is the height of the tree, and  $A$  is a

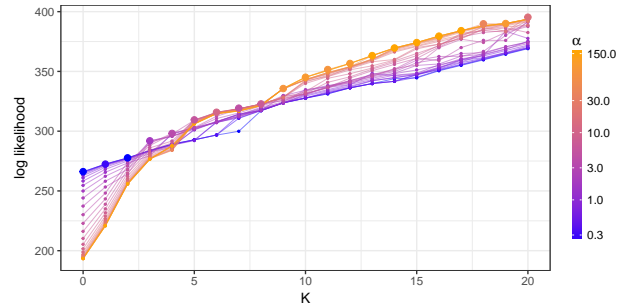


Figure 16: Likelihood profile for all the  $\alpha$  values, on the New World Monkey dataset. Each colored line represents the likelihood of the solution for a given  $\alpha$ . The maximum value of the likelihood for each  $K$  is emphasized. The maximum is not reached by the same value of  $\alpha$  for each  $K$ . Colors in log scale.

1031 constant, by default equal to 3. This ensures that the lowest  $\alpha$  makes for a  
 1032 phylogenetic half-life approximately two times as high as the tree. Lower values of  $\alpha$   
 1033 would make the process looking too much like a BM.

1034  $\alpha_{\max}$  The maximum value of  $\alpha$  is chosen so that the correlations between tips is bounded  
 1035 by  $e^{-B/2}$ , with  $B$  a constant by default equal to 2. This is obtained by noting that  
 1036 the correlation between two tips  $i$  and  $j$  for a given trait  $k$  is given by (for a  
 1037 stationary root):

$$\text{Cov}[Y_{ik}; Y_{jk}] = \frac{\frac{R_{kk}}{2\alpha} e^{-2\alpha d_{ij}}}{\sqrt{\frac{R_{kk}}{2\alpha} \frac{R_{kk}}{2\alpha}}} = e^{-2\alpha d_{ij}} \leq e^{-2\alpha d_{\min}}$$

1038 where  $d_{\min}$  is the minimum phylogenetic distance between two tips. Hence we choose  
 1039  $\alpha_{\max} = B/(2d_{\min})$ .

1040

## SIMULATIONS APPENDICES

1041

### *Kullback-Leibler Divergences*

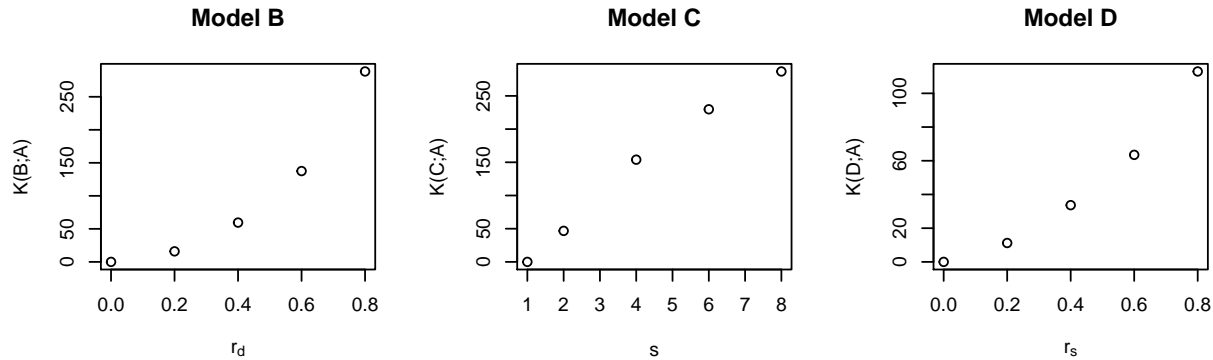


Figure 17: KL divergences from the base model

1042

Denote by  $\mathbf{I}_p$  the identity matrix of size  $p$ ,  $\mathbf{J}_p = \mathbf{1}^T \mathbf{1}$  the matrix filled with ones, and

1043

$\mathbf{S}_p = \text{Diag}(s^{-(p+1)/2+q}; 1 \leq q \leq p)$  (so that  $|\mathbf{S}_p| = 1$ ). We consider the four following models:

1044

**Model A:**  $\mathbf{A} = \alpha \mathbf{I}_p$  and  $\mathbf{R} = \sigma^2 \mathbf{I}_p$

1045

**Model B:**  $\mathbf{A} = \alpha \mathbf{I}_p$  and  $\mathbf{R} = \mathbf{R}_{r_d} = \sigma^2 (\mathbf{I}_p + r_d (\mathbf{J}_p - \mathbf{I}_p))$

1046

**Model C:**  $\mathbf{A} = \alpha \mathbf{S}_p$  and  $\mathbf{R} = \sigma^2 \mathbf{S}_p$

1047

**Model D:**  $\mathbf{A} = \alpha (\mathbf{I}_p + r_s (\mathbf{J}_p - \mathbf{I}_p))$  and  $\mathbf{R} = \frac{\sigma^2}{\lambda} \mathbf{I}_p$

1048

The general formula for a Kullback divergence between two multivariate Gaussian

1049

distributions with means  $\boldsymbol{\mu}_i$  and variances  $\mathbf{V}_i$  ( $i \in \{1, 2\}$ ) is:

$$2\mathcal{K}[\mathcal{N}_1; \mathcal{N}_2] = \text{tr}(\mathbf{V}_2^{-1} \mathbf{V}_1) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \mathbf{V}_2^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - np + \ln \frac{\det \mathbf{V}_2}{\det \mathbf{V}_1}$$

1050 We assume that the root is in the stationary state. From the general formula for a  
 1051 multivariate OU, we derive the form of the variances for these four models (Bartoszek et al.  
 1052 2012; Clavel et al. 2015):

1053 **General Formula:**  $\mathbf{V}^{(i,j)} = \mathbf{P} \left( \left[ \frac{1}{\lambda_q + \lambda_r} e^{-\lambda_q(t_i - t_{ij})} e^{-\lambda_r(t_j - t_{ij})} \right]_{1 \leq q, r \leq p} \odot \mathbf{P}^{-1} \mathbf{R} \mathbf{P}^{-T} \right) \mathbf{P}^T,$

1054 where  $\mathbf{P}$  is the orthogonal matrix of diagonalization of  $\mathbf{A}$ , associated with eigenvalues  
 1055  $(\lambda_1, \dots, \lambda_p)$ .

1056 **Model A:**  $\mathbf{V}_A = \frac{\sigma^2}{2\alpha} \mathbf{M}_\alpha \otimes \mathbf{I}_p$  with  $\mathbf{M}_\alpha = (e^{-\alpha d_{ij}})_{1 \leq i \leq j \leq n}$

1057 **Model B:**  $\mathbf{V}_B = \frac{\sigma^2}{2\alpha} \mathbf{M}_\alpha \otimes \mathbf{R}_{r_d}$

1058 **Model C:**  $\mathbf{V}_C^{(i,j)} = \frac{\sigma^2}{2\alpha} \text{Diag} (e^{-\alpha(\mathbf{S}_p)_{qq} d_{ij}}; 1 \leq q \leq p)$

1059 **Model D:**  $\mathbf{V}_D^{(i,j)} =$

1060  $\frac{\sigma^2}{2\lambda\alpha} \mathbf{P} \text{Diag} \left( \frac{1}{1-r_s} e^{-\alpha(1-r_s)d_{ij}}, \frac{1}{1-r_s} e^{-\alpha(1-r_s)d_{ij}}, \frac{1}{1-r_s} e^{-\alpha(1-r_s)d_{ij}}, \frac{1}{1+3r_s} e^{-\alpha(1+3r_s)d_{ij}} \right) \mathbf{P}^T$

1061 For model C, taking  $\mathbf{R} = \sigma^2 \mathbf{S}_p$  ensures that the variances at the tips for all the  
 1062 (independent) traits are equal to  $\gamma^2 = \frac{\sigma^2}{2\alpha}$ .

1063 For model D, the characteristic polynomial of matrix  $\frac{1}{\alpha} \mathbf{A}$  is

1064  $\chi(X) = (X + r_s - 1)^3 (X - 3r_s - 1)$ , so we wrote

1065  $\mathbf{A} = \alpha \mathbf{P} \text{Diag} (1 - r_s, 1 - r_s, 1 - r_s, 1 + 3r_s) \mathbf{P}^T$ . This leads to a variance at the tips of

1066  $\frac{\sigma^2}{2\alpha\lambda} \mathbf{P} \text{Diag} \left( \frac{1}{1-r_s}, \frac{1}{1-r_s}, \frac{1}{1-r_s}, \frac{1}{1+3r_s} \right) \mathbf{P}^T$ . By computing this matrix product (easy linear

1067 algebra formula), we find that  $\mathbf{P} \text{Diag} \left( \frac{1}{1-r_s}, \frac{1}{1-r_s}, \frac{1}{1-r_s}, \frac{1}{1+3r_s} \right) \mathbf{P}^T = (\lambda - \kappa) \mathbf{I}_p + \kappa \mathbf{J}_p$ , with

1068  $\lambda = \frac{1+(p-2)r_s}{(1-r_s)(1+(p-1)r_s)}$  and  $\kappa = -\frac{r_s}{(1-r_s)(1+(p-1)r_s)}$ . Dividing the variance matrix by a factor  $\lambda$

1069 hence ensures that the diagonal variances at the tips are still equal to  $\gamma^2 = \frac{\sigma^2}{2\alpha}$ .

1070 We can then express the Kullback distance of models B, C and D to model A, using

1071 the general formula:

$$\begin{aligned}
 2\mathcal{K} [i; A] &= \text{tr}(\mathbf{V}_A^{-1}\mathbf{V}_i) - np + \ln \frac{\det \mathbf{V}_A}{\det \mathbf{V}_i} + \left\| (\mathbf{T} \otimes \mathbf{I}_p)[\mathbf{W}(\mathbf{A}_A) - \mathbf{W}(\mathbf{A}_i)] \text{vec}(\Delta^T) \right\|_{\mathbf{V}_A^{-1}} \\
 &= \frac{2\alpha}{\sigma^2} \text{tr}((\mathbf{M}_\alpha^{-1} \otimes \mathbf{I}_p)\mathbf{V}_i) - np + np \ln \frac{\sigma^2}{2\alpha} + p \ln \det \mathbf{M}_\alpha - \ln \det \mathbf{V}_i \\
 &\quad + \left\| (\mathbf{T} \otimes \mathbf{I}_p)[\mathbf{W}(\mathbf{A}_A) - \mathbf{W}(\mathbf{A}_i)] \text{vec}(\Delta^T) \right\|_{\mathbf{V}_A^{-1}}
 \end{aligned}$$

1072 For  $\mathcal{K} [B; A]$ , we can get a closed formula that does not depend on the topology (the  
 1073 expectations term cancels out):

$$2\mathcal{K} [B; A] = n \ln[(1 - r)^3(1 + 3r)]$$

1074 For the two other distances, there are no such nice simplified formula, and the result  
 1075 depends on the topology (even when there are no shifts). To get an idea of the distance  
 1076 when there are no shifts, we computed it on 100 randomly generated trees, and took the  
 1077 mean. With shifts, we computed the distances for the trees and shift positions chosen and  
 1078 shown above.

1079 *Note on the ARI (Hubert and Arabie 1985)*

1080 *Partitions.*— Let  $S$  be a set with  $n$  elements, and  $U, V$  two different partitions of  $S$ , with  
 1081 respectively  $R$  and  $C$  groups. Denote by  $n_{ij}$  the number of elements of  $S$  that are both in  
 1082 groups  $i \in \llbracket 1, R \rrbracket$  of  $U$  and  $j \in \llbracket 1, C \rrbracket$  of  $V$ , and by  $n_i = \sum_{j=1}^C n_{ij}$  (respectively,  
 1083  $n_j = \sum_{i=1}^R n_{ij}$ ) the number of elements of  $S$  that are in group  $i \in \llbracket 1, R \rrbracket$  of  $U$  (resp.  
 1084  $j \in \llbracket 1, C \rrbracket$  of  $V$ ).

1085 *Rand Index.*— We further define:



1086 •  $a$  the number of pairs of  $S$  that are in the same groups in both partitions  $U$  and  $V$ ,

$$a = \sum_{i=1}^R \sum_{j=1}^C \binom{n_{ij}}{2}$$

1087 •  $b$  the number of pairs of  $S$  that are in different groups in both partitions  $U$  and  $V$ ,

$$b = \binom{n}{2} - \left[ a + \left( \sum_{i=1}^R \binom{n_{i\cdot}}{2} - a \right) + \left( \sum_{j=1}^C \binom{n_{\cdot j}}{2} - a \right) \right] = \binom{n}{2} + a - \sum_{i=1}^R \binom{n_{i\cdot}}{2} - \sum_{j=1}^C \binom{n_{\cdot j}}{2}$$

1088 Then the Rand index is defined as the number of agreeing pairs on the total number of  
1089 pairs:

$$\text{Rand} = \frac{a + b}{\binom{n}{2}}$$

1090 *Adjusted Rand Index.*— Assume that the null model is a generalized hypergeometric  
1091 models, where the partitions and the number of elements in each group are fixed (i.e. the  
1092  $n_{i\cdot}$  and  $n_{\cdot j}$  are fixed), and the element randomly distributed among them. Then:

$$\mathbb{E} \left[ \binom{n_{ij}}{2} \right] = \binom{n_{i\cdot}}{2} \binom{n_{\cdot j}}{2} / \binom{n}{2}$$

1093 The ARI is then defined as (1 is the maximum value of the Rand index):

$$\text{ARI} = \frac{\text{Rand} - \mathbb{E}[\text{Rand}]}{1 - \mathbb{E}[\text{Rand}]}$$

1094 which can be re-written as:

$$\text{ARI} = \frac{\sum_{i=1}^R \sum_{j=1}^C \binom{n_{ij}}{2} - \sum_{i=1}^R \binom{n_{i\cdot}}{2} \sum_{j=1}^C \binom{n_{\cdot j}}{2} / \binom{n}{2}}{\frac{1}{2} \left( \sum_{i=1}^R \binom{n_{i\cdot}}{2} + \sum_{j=1}^C \binom{n_{\cdot j}}{2} \right) - \sum_{i=1}^R \binom{n_{i\cdot}}{2} \sum_{j=1}^C \binom{n_{\cdot j}}{2} / \binom{n}{2}}$$

1095 *One class partition.*— Assume that  $R = 1$ , i.e. that one of the partition has only one class.

1096 Then:

$$\sum_{i=1}^R \sum_{j=1}^C \binom{n_{ij}}{2} = \sum_{j=1}^C \binom{n_{1j}}{2} = \sum_{j=1}^C \binom{n_{\cdot j}}{2}$$

1097 and

$$\sum_{i=1}^R \binom{n_{i\cdot}}{2} \sum_{j=1}^C \binom{n_{\cdot j}}{2} = \binom{n_{1\cdot}}{2} \sum_{j=1}^C \binom{n_{\cdot j}}{2} = \binom{n}{2} \sum_{j=1}^C \binom{n_{\cdot j}}{2}$$

1098 so that  $\text{ARI} = 0$ . Hence, if one of the true solution or the estimated solution has no shift,

1099 then the ARI is automatically equal to 0.

1100

### *Supplementary Figures*

1101 *Sensitivity / Precision.*— Because only the clustering of the tips induced by the shifts, and

1102 not their exact position on the branches of the tree, are identifiable, we used the ARI,

1103 rather than sensitivity and precision, to assess methods of shift detection. With this *caveat* in

1104 mind, we plot these quantities here for the interested reader. To do that, we removed the

1105 6.53% of solutions that were not identifiable in the results of the methods.

1106 These graphs confirm our conclusions drawn in the main text, with PhylogeneticEM,

1107 more conservative, having a better precision, along with a similar sensitivity than  $\ell_1$ ou. It

1108 is interesting to note that, even when the model is violated for PhylogeneticEM, the

1109 methods keeps a better or similar precision (see e.g. Model C in Fig. 19).

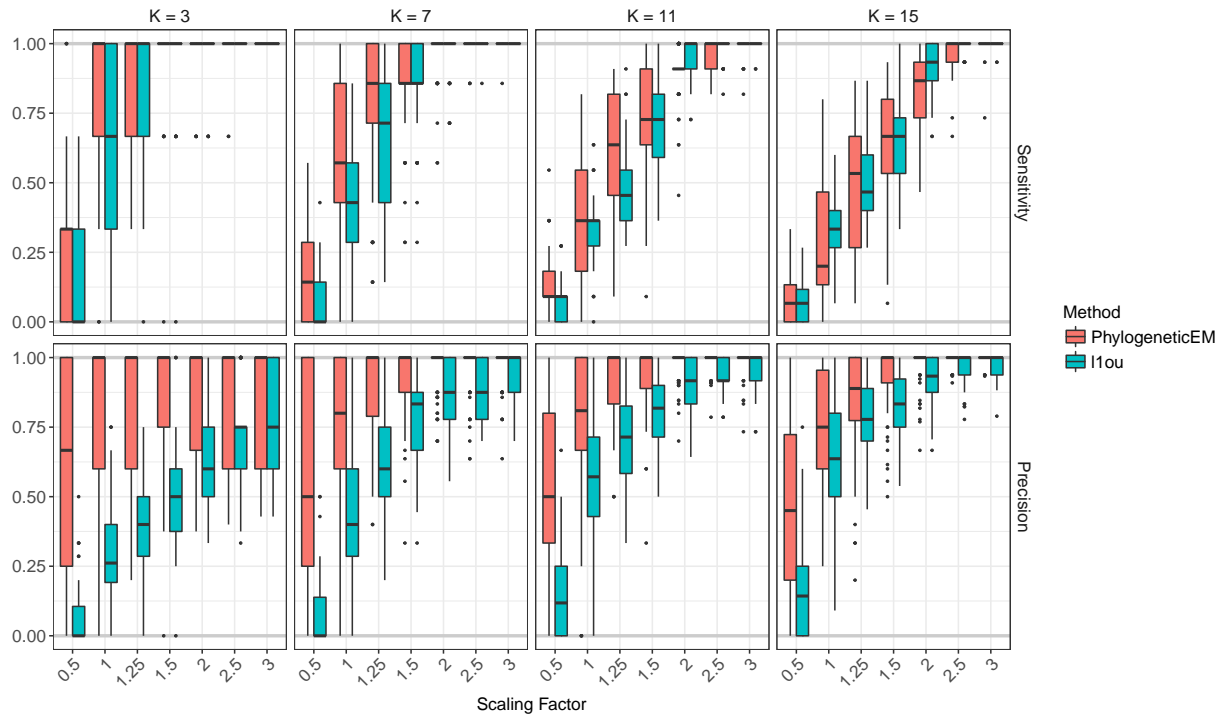


Figure 18: Sensitivity (top) and precision (bottom) for the solutions found by PhylogeneticEM (red) and  $l1ou$  (blue). Each box corresponds to one of the configuration shown in Figure 2, with a scaling factor varying between 0.5 and 3, and a true number of shift between 3 and 15 (solid lines, bottom).

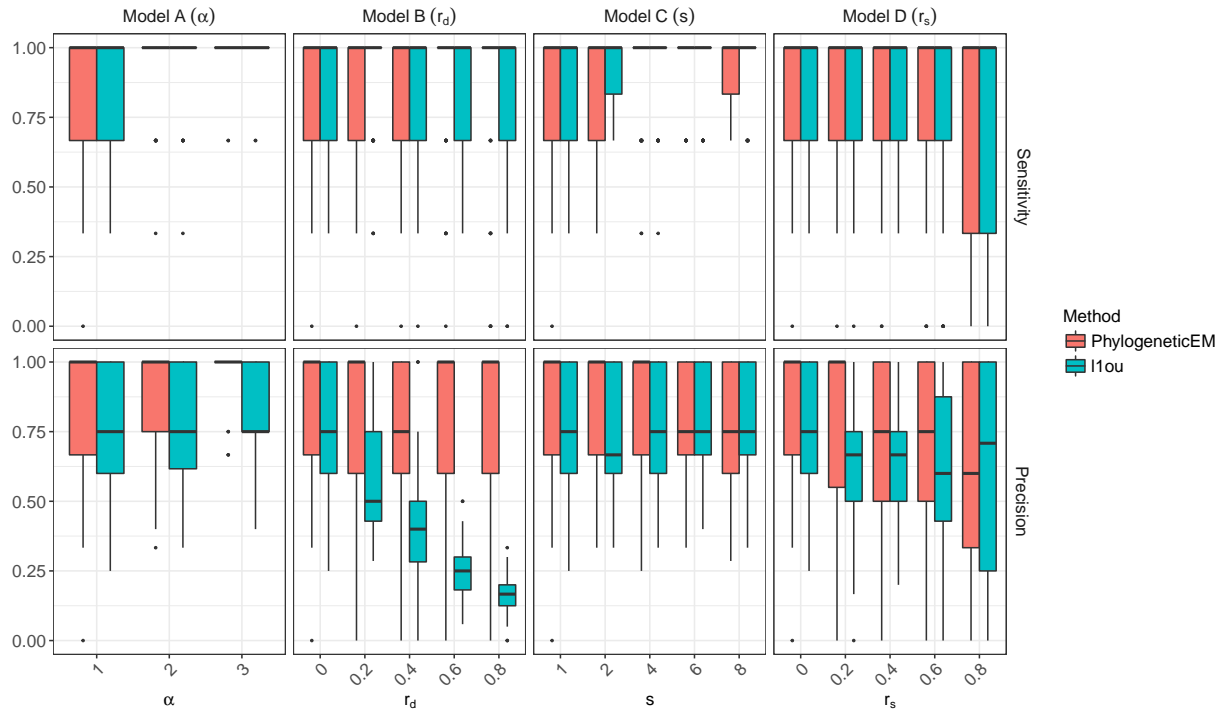


Figure 19: Sensitivity (top) and precision (bottom) for the solutions found by PhylogeneticEM (red) and  $\ell_{1ou}$  (blue). Each panel corresponds to a different type of mis-specification (except Model A) and the parameters  $r_d$ ,  $s$  and  $r_s$  control the level of mis-specification, with leftmost values corresponding to no mis-specification. For the ARI, the solid lines represent the maximum (1) and expected (0, for a random solution) ARI values.