Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7**, 811–824 doi: 10.1111/2041-210X.12534

Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models

Mohammad Khabbazian[1], Ricardo Kriebel[2], Karl Rohe[3] and Cécile Ané[2,3]*

[1]*Department of Electrical and Computer Engineering, University of Wisconsin, 1415 Engineering Drive, Madison, WI, USA;* [2]*Department of Botany, University of Wisconsin, 430 Lincoln Drive, Madison, WI, USA; and* [3]*Department of Statistics, University of Wisconsin, 1300 University Avenue, Madison, WI 53706, USA*

Summary

1. The detection of evolutionary shifts in trait evolution from extant taxa is motivated by the study of convergent evolution, or to correlate shifts in traits with habitat changes or with changes in other phenotypes.

2. We propose here a phylogenetic lasso method to study trait evolution from comparative data and detect past changes in the expected mean trait values. We use the Ornstein–Uhlenbeck process, which can model a changing adaptive landscape over time and over lineages.

3. Our method is very fast, running in minutes for hundreds of species, and can handle multiple traits. We also propose a phylogenetic Bayesian information criterion that accounts for the phylogenetic correlation between species, as well as for the complexity of estimating an unknown number of shifts at unknown locations in the phylogeny. This criterion does not suffer model overfitting and has high precision, so it offers a conservative alternative to other information criteria.

4. Our re-analysis of Anolis lizard data suggests a more conservative scenario of morphological adaptation and convergence than previously proposed. Software is available on GitHub.

Key-words: adaptation, convergent evolution, lasso, ℓ1ou, phylogenetic Bayesian information criterion, phylogenetic comparative method, regularization

Introduction

Recent advances in DNA sequencing technology and phylogenetic methods enabled accurate reconstructions of the evolutionary relationships among very large groups of species, and opened new avenues to study phenotypic trait evolution. The inference of evolutionary trees with thousands of taxa or thousands of genes demands complex mathematical models and computational tools (see for instance Bininda-Emonds *et al.* 2007; Wickett *et al.* 2014). Likewise, the inference of phenotypic trait evolution on very large trees demands complex models that are capable of handling heterogeneity across a wide range of species. Hansen (1997) used an Ornstein–Uhlenbeck (OU) process to model the macroevolution of a phenotype subject to selection pressure towards an 'optimal' value. This OU model was validated on a large number of fossil lineages (Hunt, Bell & Travis 2008; Hopkins & Lidgard 2012), as well as in cross-species comparative analyses (Harmon *et al.* 2010).

Hansen (1997) proposed to use heterogeneous OU models with different optimal phenotype values on different branches of the tree. These models can then be used to test various hypotheses about phenotypic adaptation (Butler & King 2004). For instance, Scales, King & Butler (2009) evaluated a small set of predefined hypotheses to place the various optima on the tree, to investigate whether fibre-type composition of a leg muscle in lizards is adaptive to the species predator escape strategy, or to its foraging strategy, or both. Mahler *et al.* (2013) also used OU models with varying optima, but without a preselected set of hypotheses for the number and placement of these optima (see also Ingram & Mahler 2013; Ingram & Kai 2014). To do so, they used a stepwise search among OU models to study how natural selection shaped the morphology of Caribbean Anolis lizards (Losos 2009), and then correlated the phylogenetic placements of shifts in OU optima to habitat changes. Repeated evolution of similar phenotypes in similar environments was taken as evidence for a deterministic aspect of macroevolution.

Several methods were proposed for OU models with multiple optima on phylogenetic trees, to infer the number and the position of shifts in trait optimum without predefined hypotheses. This task is difficult both computationally and theoretically, due to the very large number of models to be evaluated and compared statistically. Uyeda & Harmon (2014) developed a Bayesian method, with a Monte Carlo Markov chain implementation in the R package bayou. This method quantifies the uncertainty about the number of shifts and their phylogenetic placement. The results can vary quite heavily, however, depending on

*Correspondence author. E-mail: cecile.ane@wisc.edu

the prior distribution that the user needs to specify for the number of shifts. Ingram & Mahler (2013) developed a maximum likelihood method and a stepwise search algorithm, 'surface', with possibly convergent shifts to the same optimum (see also Mahler & Ingram 2014). Surface uses the Akaike information criterion (AIC) to select the number of shifts. In this setting, however, Ho & Ané (2014a) showed that AIC is biased towards model overfit and suggested using instead a modified Bayesian information criterion (mBIC; Zhang & Siegmund 2007) to reduce the detection of false shifts. In addition to the theoretical difficulties of inferring the correct number of shifts, both bayou and surface can become computationally heavy with large trees, handling a maximum of a few hundred taxa.

We propose here a new method to detect shifts in phenotypic optima under the OU model on trees. The method, ℓ1ou, is based on the lasso (Tibshirani 1996) and can handle extremely large phylogenetic trees with thousands of taxa. For example, analysis of sporangium shape from 886 moss (Bryophyta) species (Rose, Kriebel & Sytsma 2016) takes only 220 min with our method, whereas surface did not complete after 6 weeks. As far as we know, it is the first time that a lasso-type method is proposed for phylogenetically structured data. In the next section, we present our lasso-based methods, along with choices to deal with identifiability issues and with a new phylogenetic-aware information criterion (pBIC) to do model selection. This section can be skipped at first, and its technical details are presented in Appendix S1–S4 (Supporting information). In the following section, we show using simulations that our ℓ1ou method is also more accurate and can take advantage of multiple traits to infer a more robust model. We then illustrate the method and its scalability on data from 100 *Anolis* lizard species and four traits. We implemented the method in R, available at https://github.com/khabbazian/l1ou.

Although we focus on OU models with shifts in the optimal phenotype value, we recognize that many other types of heterogeneity might affect real data, especially at deep evolutionary scales. Changes in the rate of evolution were considered by others, mostly for BM models that exclude adaptation, to test prespecified hypotheses about where rate changes have taken place (O'Meara *et al.* 2006; Stack *et al.* 2011), or to detect the phylogenetic position and number of these rate changes (Eastman *et al.* 2011; Rabosky 2014). Changes in the strength of selection towards the optimum value have also been proposed by Beaulieu *et al.* (2012), although simultaneously detecting shifts in several of these parameters was shown to be difficult. We also caution against a literal interpretation of OU model parameters, especially at deep phylogenetic scales. In particular, even if the 'optimal value' is estimated to be constant within a given clade, this value may only reflect a broad adaptive zone, around which the true optimal value constantly fluctuates (Uyeda & Harmon 2014). In this case, it is prudent to interpret α as a parameter for phylogenetic correlation, rather than a direct estimate of the selection strength.

Lasso-based method for shift detection

THE OU MODEL ON A PHYLOGENETIC TREE

We model the evolution of a continuous phenotypic trait $y(t)$ over time t with an Ornstein–Uhlenbeck (OU) process, defined by the following stochastic equation:

$$\mathrm{d}y(t) = \alpha(\theta(t) - y(t))\mathrm{d}t + \sigma \mathrm{d}B(t),$$

where $B(t)$ is the Brownian motion (BM). This process considers trait adaptation to the environment through the parameter $\theta(t)$, called the optimum value of the trait, and which may vary over time. The parameter $\alpha \geq 0$ is the rate of adaptation. Equivalently, the phylogenetic half-life, $\log(2)/\alpha$, is the amount of time it takes for the trait expected value to reach halfway to the optimum value. If $\alpha \approx 0$, or $\log(2)/\alpha$ is much larger than the time interval of interest (e.g. the tree height), then the expected value of $y(t)$ converges slowly to the optimum relative to the observed time period. In this case, $y(t)$ mostly varies around the ancestral state and the OU process reduces to a BM.

Throughout the paper, we assume a known phylogenetic tree for the species of interest. We also assume that this tree is rooted, binary and ultrametric. The OU process is assumed for the evolution of trait y along each branch of tree, independently for the two daughter branches of each node conditional on the trait value at that node. For simplicity and identifiability of the model parameters, we assume that, although unknown, α and σ^2 are fixed across the tree but that the optimum value θ may vary across time and across branches in the tree.

We make further assumptions on changes in $\theta(t)$ because its estimation suffers from identifiability issues. Ho & Ané (2013, 2014a) showed that a relatively small variation in $\theta(t)$ cannot be distinguished with certainty from variation caused by the BM part of the process, even with an infinite number of present-day species if the tree height is bounded (for trees of growing height such as from the Yule process, Adamczak & Miłoś 2015; Ané, Ho & Roch 2015; Bartoszek & Sagitov 2015). Ho & Ané (2014a) also showed that the exact location and number of changes in the optimum value, also called shifts, cannot be identified when these shifts are on the same branch (see Fig. 1, left). Given these restrictions, we assume that $\theta(t) = \theta_b$ is constant along branch b, so that θ is a piecewise constant function from the root to any species (leaf). In other words, we assume at most one shift on each branch, located at the beginning of the branch if present. This parsimonious model can still describe the effect of many shifts on each branch.

Even with this parsimonious assumption, the shift positions on a tree can still be unidentifiable. For example, Fig. 1 (right) shows different shift placements that all correspond to the same grouping of taxa and would all receive equal likelihoods. We explain below (and prove in Appendix S2, theorem 1) that our method deals with this unidentifiability, and automatically returns a parsimonious model in terms on number of shifts and shift magnitudes (in absolute values).

METHOD FOR ONE TRAIT (UNIVARIATE CASE)

Shift detection as a linear model selection problem

Under our assumption that there exists at most one shift at the beginning of any given branch, the trait values at leaves follow this linear model (see Appendix S1 for the full derivation):

$$Y = \beta_0 \mathbb{1} + X^{(\alpha)}\beta + \varepsilon \qquad \text{eqn 1}$$

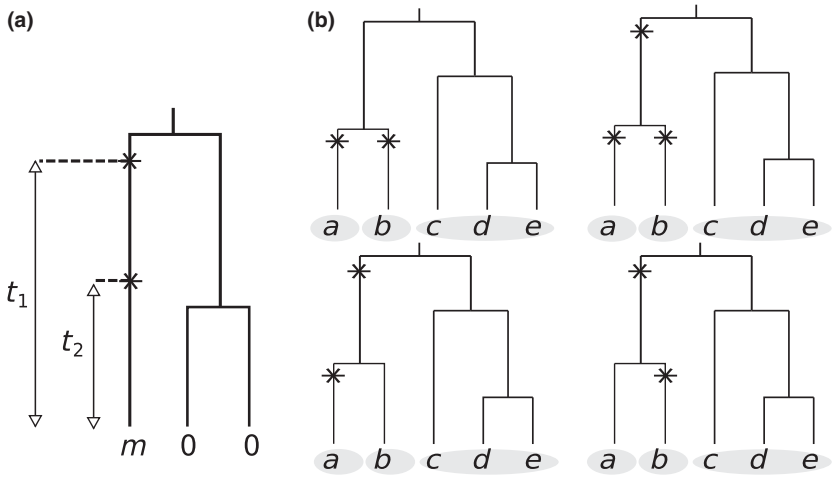

Fig. 1. The number and position of shifts on a given branch cannot be identified. (a) On a single branch, one shift at age t_1 or one shift at age t_2 or two shifts at ages t_1 and t_2 lead to the exact same model with means m, 0, 0 at the leaves, provided that the shift magnitudes $\Delta\theta_i$ (at t_i) satisfy $(1-\exp(-\alpha t_1))\Delta\theta_1 + (1-\exp(-\alpha t_2))\Delta\theta_2 = m$. (b) These four shift configurations generate the same model, with shifts denoted as stars (*). Each has three clusters of tips sharing the same mean: $\{a\}$, $\{b\}$, and $\{c, d, e\}$. The top right configuration is not parsimonious and cannot be returned by ℓ1ou. The other three configurations are all parsimonious and may be returned by ℓ1ou depending on the data.

where β_0 is an overall mean ($\mathbb{1}$ is a vector of ones). The β coefficients contain the magnitude of the shifts in selection optimum, that is changes in θ values, one for each branch b in the tree: $\beta_b = \theta_b - \theta_{p(b)}$ where $p(b)$ is the parent of b. The nonzero elements in β correspond to the set of branches where θ changes, that is, the shift positions. Following Rabosky *et al.* (2014), we call this set of branches with shifts a 'shift configuration'. The design matrix $X^{(\alpha)}$ has n rows (number of taxa) and p columns (number of branches) and depends on α. Define a_b to be the age of b's parent node, that is, the distance from the parent node to its descendant species. For taxon i and branch b,

$$X^{(\alpha)}_{ib} = \begin{cases} 1 - e^{-\alpha a_b} & \text{if } b \text{ is on the path from the root to taxon } i \\ 0 & \text{if taxon } i \text{ is not a descendant of } b \end{cases}$$

(see Appendix S1, for details). Correlations due to shared evolutionary history are captured in the error ε that follows a centred normal distribution with covariance $\Sigma^{(\alpha)}$ derived from the OU model:

$$\Sigma^{(\alpha)}_{ij} = \begin{cases} \sigma^2 e^{-\alpha d_{ij}}(1 - e^{-2\alpha t_{ij}})/(2\alpha) & \text{if the root value is fixed} \\ \sigma^2 e^{-\alpha d_{ij}}/(2\alpha) & \text{if the root value has the stationary distribution} \end{cases} \quad \text{eqn 2}$$

where t_{ij} is the evolutionary time shared between species i and j, and d_{ij} is their tree distance.

The linear regression (1) cannot be solved with ordinary least squares for several reasons. First, $X^{(\alpha)}$ has more columns (branches with potential shifts) than rows (species with observations). Secondly, the columns in $X^{(\alpha)}$ are highly correlated, in particular because one shift on a given branch is equivalent to two shifts of equal magnitudes located on each of the two daughter branches. Finally, the predictors in $X^{(\alpha)}$ depend on the unknown adaptation rate, α. However, if we restrict the set of hypothetical shifts and if we reduce $X^{(\alpha)}$ to these branches accordingly, then (1) may have a least-squares solution. We show that it is indeed the case if the shift configuration is 'identifiable', that is, if every hypothesized shift is 'visible' from at least one taxon (more formally, see Appendix S2). The main problem is then to select the shift configuration that best fits the data, among all the identifiable shift configurations.

Regularization with lasso

To tackle the challenges outlined above, which come from the high-dimension nature of the problem, the typical assumption is that only a relatively small subset of predictors (here, shifts) describes the response. In other words, we assume that β is sparse or that most shift magnitudes are 0. A common way to achieve this is to consider the lasso problem (Tibshirani 1996) whose solution $\hat{\beta}$ minimizes the following ℓ_1-penalized least square criterion:

$$\frac{1}{2}||Y - \hat{\beta}_0\mathbb{1} - X\hat{\beta}||_2^2 + \lambda||\hat{\beta}||_1 \quad \text{eqn 3}$$

where λ is a tuning parameter and the ℓ_1 norm of the shift magnitudes is simply the sum of their absolute values: $\ell_1(\hat{\beta}) = \sum_b |\hat{\beta}_b|$. This penalty term causes many estimated shifts in $\hat{\beta}$ to be zero, which leads to selecting the most relevant features. By varying the tuning parameter λ from zero to ∞, we increase the weight of the penalty and obtain $\hat{\beta}$'s with support of size n shifts (no penalty) to zero shifts (extreme penalty). Compared to an ℓ_2 penalty in ridge regression, for instance, the ℓ_1 penalty has the advantage of sparsity: where the estimated shifts are $\hat{\beta}_b = 0$ exactly on many branches.

The theory of the lasso is well explored (for instance Bühlmann & Van De Geer 2011; Eldar and Kutyniok, 2012). To guarantee statistical selection consistency, small prediction error and uniqueness of the estimate, various sufficient conditions were introduced on the sparsity of the coefficient vector and coherency of the design matrix (Van De Geer, Bühlmann & et al. 2009). For instance, Zhao & Yu (2006) showed that if (i) X satisfies the 'irrepresentable condition', (ii) ε contains independent random variables with finite variance, and (iii) λ is chosen to have the appropriate scale, then with high probability, the nonzero elements of $\hat{\beta}$ are identical to the nonzero elements of the true β. These results allow for p to grow asymptotically faster than n, so long as the number of nonzeros in β grows slower than n. Furthermore, different methods based on convex optimization, combinatorial and greedy algorithms were proposed to compute the exact or approximate solution. Efron *et al.* (2004) showed an intuitive connection between the lasso and stepwise selection solutions. They proposed the fast LARS algorithm to find the lasso estimates $\hat{\beta}$ that minimize (3) at every value of λ.

We now rewrite model (1) to derive an appropriate ℓ_1 penalty so as to estimate a parsimonious shift configuration and to account for phylogenetic correlation. If this correlation was ignored, a straight ℓ_1 penalty would bias shift detection in favour of large clades in the tree, for which similarity might otherwise be explained by common ancestry. We first consider the case when α is known, which implies that

$X := X^{(\alpha)}$ and the phylogenetic covariance $\Sigma := \Sigma^{(\alpha)}$ are known. To remove phylogenetic correlation, we consider $\Sigma^{-1/2}Y$, whose components are uncorrelated, but whose mean is $\Sigma^{-1/2}(\beta_0\mathbb{1} + X\beta)$. Therefore, our lasso estimate is the solution β that minimizes the following ℓ_1-penalized criterion:

$$\frac{1}{2}||\Sigma^{-1/2}(Y - \beta_0\mathbb{1} - X\beta)||_2^2 + \lambda||\beta||_1 \qquad \text{eqn 4}$$

Throughout the document, this will be referred to as the *phylogenetic lasso*. We use the R package lars to solve this optimization problem for all values of the tuning parameter λ (see Fig. 2 for an example) (Efron *et al*. 2004). An extra model selection phase is then required to find the appropriate λ and the corresponding estimated number of shifts.

Under some mild conditions and for every λ, we prove in Appendix S2 (theorem 1) that there is a unique solution $\hat{\beta}$ minimizing (4), and that the support of $\hat{\beta}$ is an identifiable shift configuration. Furthermore, in Appendix S4 we explain a linear algorithm to calculate $\Sigma^{-1/2}$ efficiently in linear time. This algorithm is based on the method proposed by Stone (2011).

Model selection for the number of shifts

In traditional models with uncorrelated errors, tuning the penalty weight λ is typically done with tools such as cross-validation, minimum expected information loss (AIC) or maximum model posterior probability (e.g. BIC, Schwarz, 1978). In our problem, cross-validation is not appropriate since leaving out some taxa may erode small clades with a shift, taking away part of the signal of interest. In surface, the following criterion is used:

$$\mathrm{AIC_c}(\mathcal{M}_k) = -2\log\,\mathrm{lik}(\mathcal{M}_k) + 2p + \frac{2p(p+1)}{nm - p - 1}$$

where $\mathcal{M}_k$ is the hypothesis that there are k shifts, $\mathrm{lik}(\mathcal{M}_k)$ is the maximum likelihood of the best k-shift configuration, and m is number of traits, all assumed to share the same shift configuration. Here $p = k + m(k + 3)$ is the number of parameters, counting the position of each shift as one parameter, and $k + 3$ parameters specific to each trait (shift magnitudes, β_0, α and σ). Ho & Ané (2014a) showed that minimizing AIC leads to strong model overfitting, however. Therefore, we adapt BIC to better estimate the model posterior probability in the situation when errors are phylogenetically correlated.

The traditional BIC score of $\mathcal{M}_k$ can be defined as

$$\mathrm{BIC}(\mathcal{M}_k) = -2\log\mathrm{lik}(\mathcal{M}_k) + (k + m(k+3))\log(n)$$

where again each shift location is counted as a parameter and $k + 3$ parameters are specific to each trait.

In Appendix S3 we show that a phylogenetic correction must be applied to better approximate the marginal probability that the true model has k shifts, leading to the following phylogenetic BIC for $m = 1$ trait:

$$\mathrm{pBIC}(\mathcal{M}_k) = -2\log\mathrm{lik}(\mathcal{M}_k) + 2k\log(2n-3) + 2\log(n) + \log\det\left(X_{\mathcal{M}_k}^{(\hat{\alpha})\prime} v\Sigma^{(\hat{\alpha})^{-1}} X_{\mathcal{M}_k}^{(\hat{\alpha})}\right) \qquad \text{eqn 5}$$

where $X_{\mathcal{M}_k}^{(\alpha)}$ is the matrix $X^{(\alpha)}$ reduced to the columns corresponding to the k estimated branches with a shift but expanded with a column of ones to include the intercept, and v is the observed trait variance. Informally, $2k\log(2n-3)$ is the penalty term for the shift positions and comes from approximating twice the log of the number of configurations with k shifts, when the tree grows ($n\to\infty$). The penalty for the shift magnitudes and the intercept is captured by the last term, which appears when these parameters are integrated out with a non-informative flat prior. Interestingly, this penalty is not a simple function of the number of parameters. The determinant term depends on α and more importantly, on the location of the shifts through the structure of $X_{\mathcal{M}_k}^{(\alpha)}$. For instance, if α is infinite and if the configuration has two shifts that separate the taxa into three distinct groups of sizes n_1, n_2 and n_3, then

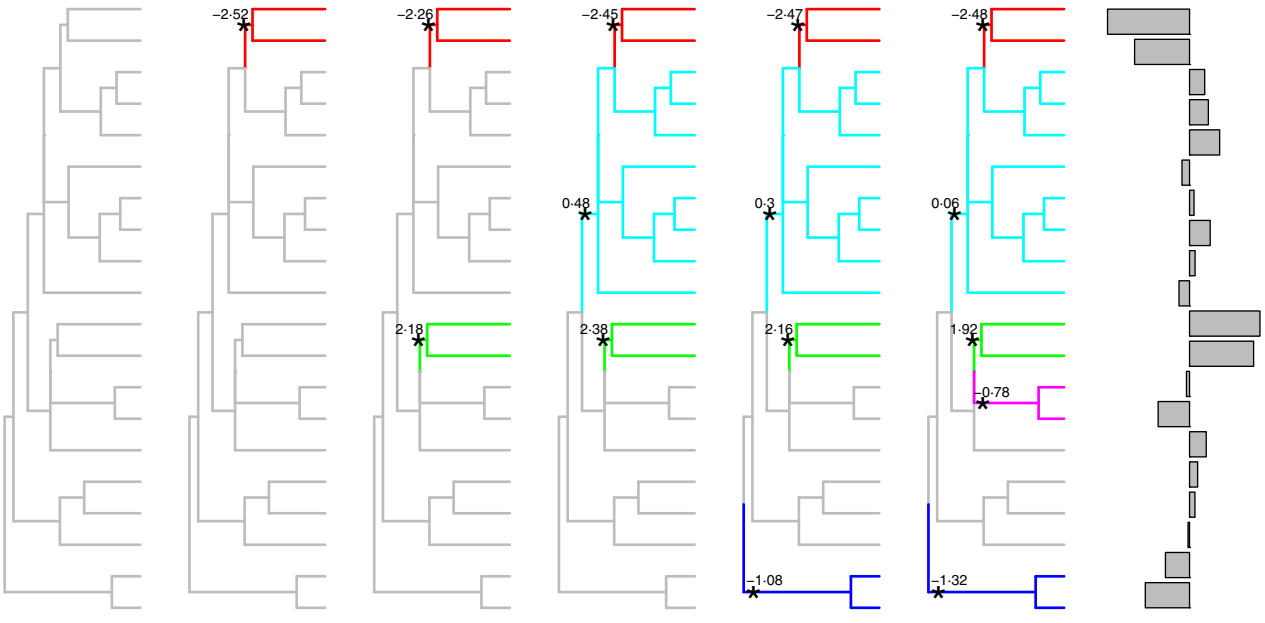

Fig. 2. Example of our lasso solution path. The number of estimated shifts depends on the penalty parameter, with 0 to 5 estimated shifts as λ decreases progressively from infinity to $\lambda = 3.07$ (one estimated shift $\beta_b \neq 0$), $\lambda = 3.31, 2.64, 1.92$ and 1.73 (five estimated shifts). The shift configurations are shown from left to right. Each estimated shift is indicated by a star and by its magnitude $\hat{\beta}_b$. Decreasing λ further would further increase the number of estimated shifts (at $\lambda = 1.09, 0.93$ etc.) The sample data are shown with the bar graph.

the last penalty term is proportional to $\log(n_1) + \log(n_2) + \log(n_3)$, just like in the modified BIC proposed by Ho & Ané (2014a). These numbers of taxa n_i are the effective sample sizes for the intercept and shift values, that is the number of observations that effectively provide information on these parameters (when $\alpha = \infty$). This last penalty term generalizes the effective sample size proposed in Ané (2008), to an OU phylogenetic model with any number of shifts.

While pBIC is written here specifically for an OU process, it can easily be applied to any process with k shifts in the mean and any phylogenetic correlation structure, such as a BM process with jumps. To do so, $X^{(\hat{\alpha})}_{\mathcal{M}_k}$ in (5) needs to be the design matrix controlling how shift coefficients affect the species means, $\Sigma^{(\hat{\alpha})}$ the estimated phylogenetic covariance, and 2 log (n) needs to be replaced by p log (n) where p is the number of parameters for the phylogenetic covariance structure, including σ^2.

For multiple traits, $2k$ log $(2n-3)$ appears only once to penalize the shift configuration shared by all traits, but each trait contributes its own 2 log (n) and determinant terms to penalize the trait-specific shift magnitudes, β_0, α and σ.

In order to choose λ, we compute the information criterion (BIC or pBIC) for each shift configuration found by the lasso solution path, and then, we pick the few top solutions (and their associated λ). While our phylogenetic lasso assumes a fixed α in (4), α is then optimized during the likelihood and pBIC (or BIC) evaluation of each shift configuration found by lasso. The columns of the design matrix in (4) can be correlated, causing the lasso to pick groups with redundant shifts. To drop these shifts, we add an extra 'backward selection' step: any shift whose removal improves the information criterion is dropped. This backward procedure is only performed for the best few models in the solution path to obtain the final estimated model.

Dealing with unknown phylogenetic covariance

Our prior assumption that the adaptation rate α is known is not realistic. So we repeat the procedure twice, once with a conservative starting value for α, and then again with an estimate of α informed by the shift configuration found in the first round (see the outline below with all steps)

We assume in the first round that $\alpha \approx 0$, which leads to the greatest level of phylogenetic correlation, that of a BM. This is conservative because similarity among all species of a clade might be explained by shared ancestry, rather than a shift at the base of the clade. However, $\mathbf{X}^{(\alpha)}$ in (1) is degenerate when $\alpha = 0$ (absence of adaptation to the shifts), so we consider its linear approximation when α is small. Its nonzero terms are $1 - e^{-\alpha a_b} \sim \alpha a_b$, and this approximation is most accurate for young branches (young age a_b). Therefore, for our first round with $\alpha \approx 0$ we rewrite (1) as follows:

$$Y = \widetilde{X}\widetilde{\beta} + \varepsilon \qquad \text{eqn 6}$$

where $\widetilde{X}_{ib} = a_b$ if taxon i is a descendant of b, $\widetilde{X}_{ib} = 0$ otherwise, and $\widetilde{\beta} = \alpha\beta$. The phylogenetic covariance for ε is assumed to be $\Sigma^{(0)}$ from the BM. The phylogenetic lasso (4) is solved in this first round using $\widetilde{X}$ and $\Sigma^{(0)}$. As already noted by Hansen (1997), this multipeaked OU process with $\alpha \approx 0$ corresponds to a BM model with regime-specific trends, with the trend coefficients estimated by $\widetilde{\beta}$ here.

Recall that α is estimated through maximum likelihood during the pBIC (or BIC) evaluation, separately on each candidate configuration, when tuning the lasso penalty λ to do model selection. This is performed with a linear-time algorithm in the R package phylolm v2.2 (Ho & Ané 2014a, b). We then use $\hat{\alpha}$ estimated from the best shift configuration selected in the first round, as input to the phylogenetic lasso (4) for a second round. Simulations show that this second round improves the final estimates of the shift positions. We summarize below these various steps of our method, which we call ℓ1ou + IC, where IC is any information criterion (e.g. pBIC).

1. Find the solution path of the phylogenetic lasso (4) for $\alpha = 0$ (BM covariance), using the linear approximation for $\mathbf{X}^{(\alpha)}$.
2. Calculate $\hat{\alpha}$, $\hat{\beta}$ that maximize the likelihood then calculate IC for each candidate configuration on the path from step 1 (and some simpler configurations, see previous section). Retain the configuration with the best IC.
3. Solve the phylogenetic lasso (4) using $\alpha = \hat{\alpha}$ from the configuration found in step 2.
4. Repeat step 2 but on the path of candidate configurations found in step 3.
5. Retain the shift locations, $\hat{\alpha}$ and $\hat{\beta}$ from the configuration with the best IC among those found in steps 2 and 4.

Detecting convergent regimes

An adaptation of the phylogenetic lasso can determine if some shifts converge to the same optimal value in multiple parts of the tree, as might be expected if different clades share a similar environment. After shift locations have been estimated by ℓ1ou, convergent evolution can be detected by minimizing the following criterion

$$\frac{1}{2}||\Sigma^{-1/2}(Y - \beta_0\mathbb{1} - X\beta)||_2^2 + \lambda||\mathbf{M}\beta||_1. \qquad \text{eqn 7}$$

This differs from the phylogenetic lasso (4) because it penalizes linear combinations of shift magnitudes, $\mathbf{M}\beta$. $\mathbf{M}$ is built so that each row captures the difference in optimal value between two regimes in the tree. To detect convergence among the first two shifts for example, if the configuration estimated by ℓ1ou was as in the left tree of Fig. 3, $\mathbf{M}$ would include a row with entries $(1 - e^{-\alpha a_{b_1}})$ and $-(1 - e^{-\alpha a_{b_2}})$ in the columns corresponding shifts 1 and 2, respectively (a_b is the age of a shift on branch b), and 0 entries otherwise. In general, $\mathbf{M}$ has at most $k(k-1)/2$ rows if k shifts were detected by ℓ1ou, but could have fewer rows because we do not need to test for a convergence that would remove a single shift. Tibshirani & Taylor (2011) provide a fast solution path algorithm to solve the generalized lasso for an arbitrary $\mathbf{M}$, implemented in the R package genlasso. An information criterion can then be used to select the best model (or λ) along the solution path. For pBIC, the design matrix $X^{(\alpha)}_{\mathcal{M}_k}$ is reduced to the convergent model with one column per distinct optimal value. This pBIC formulation is heuristic here (like $\mathrm{AIC_c}$ or BIC) as our derivation of (5) assumed independent shifts.

METHOD FOR MULTIPLE TRAITS (MULTIVARIATE CASE)

Using multiple traits should increase the power and increase the method's robustness to detect shifts. An easy way to analyse multiple traits is to reduce the data to just a few dimensions, such as with principle component analysis (PCA), and separately analyse the first few dimensions that explain most of the variance. Revell (2009) demonstrated that PCA is misleading for phylogenetic data and proposed phylogenetic PCA (pPCA) instead, which assumes a BM covariance among taxa. Recently, Uyeda et al. (2015) showed that both standard PCA and pPCA are biased, in that the top principal components (PC) are most influenced by the traits varying early in the tree. This bias suggests that false shifts might be detected near the root of the tree if ℓ1ou (or other shift detection methods) are

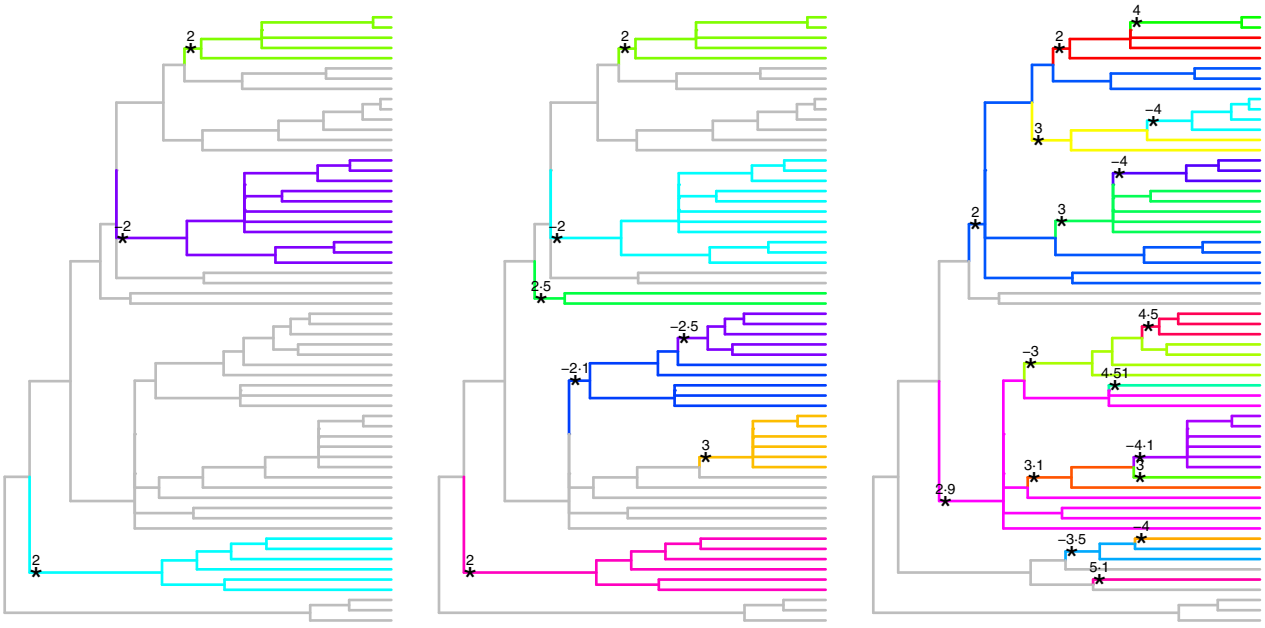

Fig. 3. Tree with 60 taxa used in simulations to compare the accuracy of various methods. Data were simulated under the OU model with no shifts or with multiple shifts (left: 3, centre: 7, right: 17 shifts). The shift positions are annotated with stars and their simulated magnitudes.

used on the first few PC axes. Indeed, this was confirmed in our simulations (see Fig. 7).

To extend our ℓ1ou method to multiple traits, we assume that traits shifted at the same time in the past, on the same branches in the tree. In other words, we group the shift magnitudes for all traits on a given branch together, and we seek to estimate a model where either all shifts in a group are 0 (none of the traits shifted on that branch) or most of the shifts in a group are not 0 (many of the traits shifted on that branch). More formally, we assume (like in surface) that the m traits arose from independent OU processes, each with its own α and σ^2 parameters, but with shifts on a shared set of branches. We write the m observed traits in a long vector $\mathbf{Y}$ of size nm by stacking each trait on top of one another, and we collect the trait-specific adaptation and variance rates in vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}^2$. We also write the shift magnitudes as a long vector $\boldsymbol{\beta}$ by stacking the coefficient of each trait (β_{jb} for trait j on branch b) on top of one another, and we similarly stack the intercepts for all traits into a vector $\boldsymbol{\beta}_0$ of size m. The multivariate response model becomes

$$\mathbf{Y} = \mathbb{1}\boldsymbol{\beta}_0 + \mathbf{X}^{(\alpha)}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $X^{(\alpha)}$ is a block diagonal matrix of size $mn \times mp$ with $X^{(\alpha_j)}$ for trait j on the diagonal, and $\mathbb{1}$ is similarly block diagonal with $\mathbb{1}$ as diagonal terms. The errors $\boldsymbol{\epsilon}$ are assumed to be phylogenetically correlated with variance $\Sigma^{(\alpha_j)}$ for trait j, but independent across traits. It means that, conditional on knowing the true shifts, residual variation ($\boldsymbol{\epsilon}$) is uncorrelated between traits. If shifts are unknown however, traits are correlated because they shift on the same branches. So in fact, we assume that all the between-trait correlation (as could be estimated with straight Pearson correlation coefficients) is due to correlation between shifts.

Yuan & Lin (2006) proposed the group lasso to generalize the lasso when there are predefined groups of coefficients. Here, each branch b in the tree corresponds to a group of coefficients: $(\beta_{jb})_{j \le m}$ across traits. To capture the trend that all coefficients in a group are 0 (or not) together, the group lasso uses the ℓ_1 penalty over groups, rather than over individual coefficients:

$$\sum_{\text{branch } b} \|\beta_{\cdot,b}\|_2 = \sum_{\text{branch } b} \left(\sum_{\text{trait } j} \beta_{jb}^2 \right)^{1/2}.$$

The $||\beta_{\cdot,b}||_2$ acts as an ℓ_1 penalty on the group of shifts on branch b. This group contains all of the shifts on branch b, for every one of the traits. Because it acts as an ℓ_1 penalty on the group, this penalty selects groups (here branches) to be either entirely zero or entirely nonzero. In the special case when there is only one trait, this penalty reduces to the earlier ℓ_1 penalty: $\sum_b |\beta_b|$. Using this group penalty, we consider the following multivariate phylogenetic lasso

$$\min_{\boldsymbol{\beta}} \frac{1}{2} ||\boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \mathbb{1}\boldsymbol{\beta}_0 - \mathbf{X}^{(\alpha)}\boldsymbol{\beta})||_2^2 + \lambda \sum_b ||\beta_{\cdot,b}||_2, \qquad \text{eqn 8}$$

where $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}^{(\alpha)}$ is block diagonal with $\Sigma^{(\alpha_j)}$ on its diagonal. We used the R package grplasso for solving this group lasso step. Unlike LARS, the search for the λ values where the shift configuration changes is done using a grid search, which can be slower. We then select λ and the associated shift configuration using the same ℓ1ou procedure as before, simply replacing (4) by (8) in the lasso steps 1 and 3.

BOOTSTRAP SUPPORT FOR SHIFTS

To quantify uncertainty in the detected shifts, we use an adapted bootstrap procedure, borrowing ideas from Freckleton & Harvey (2006) (see also Pennell *et al.* 2015).

1. Use ℓ1ou to estimate $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. For each trait j, compute $\Sigma_j^{-1/2}$ and $\Sigma_j^{1/2}$ in linear time, where $\Sigma_j = \Sigma_j^{(\hat{\alpha}_j)}$ is the phylogenetic correlation for trait j. Then compute the vector of residuals for trait j: $R_j = \Sigma_j^{-1/2}(Y_j - X^{(\hat{\alpha}_j)}\hat{\beta}_{j\cdot})$.

2. Repeat a very large number of times the following. For each trait j, sample from R_j with replacement to create a bootstrap sample of n residuals $\widetilde{R}_j$. Use ℓ1ou to estimate the shift configuration and shift magnitudes from the bootstrap data with $\widetilde{Y}_j = X^{(\hat{\alpha}_j)}\hat{\beta}_{j\cdot} + \Sigma_j^{1/2}\widetilde{R}_j$,

3. For each branch, calculate the bootstrap support for a shift on that branch as the proportion of bootstrap iterations when a shift was detected on that branch for one of more traits.

This procedure is expected to be conservative, because shifts that are undetected in step 1 cannot receive high bootstrap support. An undetected shift would just contribute one large residual, which would be resampled and 'scattered' throughout the tree in the bootstrap resampling step 2. Note that the bootstrap results from step 2 could be summarized more thoroughly in step 3. For instance, on each branch with a estimated shift, a bootstrap confidence could be obtained for the magnitude of this shift.

Simulations

SHIFTS IN ONE TRAIT

We used simulations to compare the accuracy of different methods: ℓ1ou combined with either pBIC, BIC or the same AIC_c as used in surface (using its forward phase only to focus on the shift configurations rather than the shift magnitudes), bayou, and the stepwise selection method proposed by Ho & Ané (2014a) . This stepwise method is capable of accepting various criteria, but we used here the 'mBIC' also proposed by Ho & Ané (2014a) . Bayou requires the user to choose a prior distribution for each model parameter, and the results are sensitive to this choice. We made choices based on the true parameters used to simulate the data: the number of shifts was given a conditional Poisson prior distribution with mean the true number of simulated shifts. A uniform prior was chosen for α and σ^2 on $[\alpha - 0.5, \alpha + 0.5]$ and $[\sigma^2 - 0.5, \sigma^2 + 0.5]$. An empirical Bayes approach was taken for the shift magnitudes as in Uyeda & Harmon (2014), with a centred normal prior distribution with standard deviation equal to twice that observed in the tip data. Since bayou is a Bayesian method, it returns a posterior distribution on shift configurations. To summarize this distribution, we took a liberal approach and said that a branch was detected to have a shift if the posterior probability of a shift on that branch was 0.10 or greater. For ℓ1ou methods, we used the random root covariance in (2). For all methods, we set the maximum number of shifts to half the number of taxa in the tree.

We simulated data sets under OU models along two different phylogenies of flowering plants in the family Melastomataceae, one with 60 taxa and one with 215 taxa using the function rTraitCont in the R package ape (Paradis, Claude & Strimmer 2004). The first tree (Fig. 3) is the consensus phylogeny from Kriebel, Michelangeli & Kelly (2015) pruned to a single accession per species. It was small enough for all methods to run reasonably fast and was used to compare the methods' accuracies. The second tree was simply used to sample subtrees and compare the methods' running times as a function of tree size.

On the 'small' tree, we simulated traits under four different configurations: either no shift, or 3, 7 or 17 shifts as shown in Fig. 3. We used $\alpha = 1$, corresponding to a moderate half-life (0.69) compared to the tree height, which was set to 1 by rescaling all branch lengths. We set $\sigma^2 = 2$ to fix the stationary variance $\sigma^2/(2\alpha)$ at 1. In the absence of shifts, we varied α while keeping $\sigma^2/(2\alpha) = 1$. In the presence of shifts, we instead varied the shift magnitudes. They were first set to the values shown in Fig. 3. They correspond to moderate magnitudes, just large enough to be detected individually (if their phylogenetic positions were known) with non-negligible power (Ho & Ané 2014a), because means at the tips differ by about 1 stationary standard deviation ($\beta_b(1 - e^{-\alpha a_b}) \approx \sigma/\sqrt{2\alpha} = 1$). These shift magnitudes were then all scaled by the same factor, varying from 1 to 4, to create easier scenarios. For each condition, we generated 200 replicate data sets with 1 trait each and estimated the shift configuration using each method.

To compare the methods' accuracies, we first considered the scenario with no shifts and calculated the number of false positives, that is the average number of detected shifts, necessarily all false. Figure 4 (left) shows that ℓ1ou + AIC_c and surface (both using AIC_c) have many more false positives than the other methods.

Next, we considered scenarios with 3, 7 or 17 true shifts and calculated the recall rate of each method (average proportion of branches with a true shift that were detected as having a shift) and precision (average proportion of detected shifts that were true, that is located on a branch with a true shift). Figure 5 shows that surface and ℓ1ou with AIC_c are liberal methods: both enjoy high recall rates (they find many of the true shifts) but tend to have low precision (they also find many false positives). Given our liberal threshold to call a shift in bayou (PP $\geq$ 0·1), it is not surprising to find that bayou also has a ten-

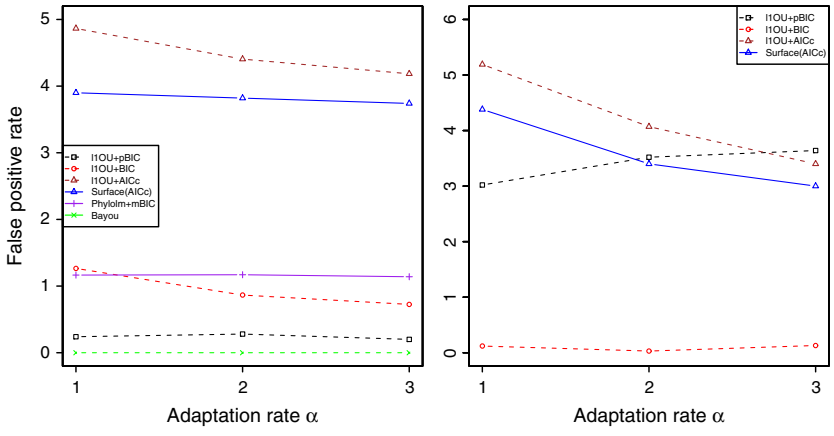

Fig. 4. Number of false positives for different methods to detect shifts in the Ornstein–Uhlenbeck (OU) process. One trait (left) or four independent traits (right) were simulated under a homogeneous OU model with no shift.

dency to be liberal, with high recall rates and low precision. However, both performance measures tended to be lower for bayou than for ℓ1ou with AIC_c. On the other extreme, ℓ1ou was conservative when coupled with pBIC, enjoying the highest precision (the detected shifts were mostly true) but a low recall rate (many true shifts were missed). When coupled with BIC and on a single trait, ℓ1ou provided an intermediate approach, which might provide a good balance to reach a reasonable precision with a reasonable recall rate. The phylolm stepwise method based on mBIC performed consistently more poorly than other methods. Its recall rate was among the lowest, but its precision was always comparable or lower than that or ℓ1ou with pBIC, for instance. Figure 5 (right) also shows that identifying 17 shifts on a 60-taxon tree is much more difficult than detecting three or seven shifts. The performance of all methods went down significantly with 17 shifts. This is not surprising, because each shift was visible by an average of 3·5 extant species, compared to 8·6 when there were only seven shifts. Detecting the exact position of each shift is likely to be much more difficult as the density of shifts increases within the tree.

To compare the methods' running time, we used a 215-taxon plant phylogeny expanded from Kriebel, Michelangeli & Kelly (2015) and randomly subsampled between 32 and 215 taxa to obtain a smaller tree. For each tree size, we generated 2 replicate data sets with a single trait, using $\alpha = 1$, $\sigma^2 = 2$ and a true number of shifts that increased with the tree size (from four shifts on 32 taxa to 26 shifts on 215 taxa). The maximum number of estimated shifts was set for all methods to twice the true number of shifts. We kept a constant number of 400000 generations in bayou, because it is unclear how this number should be set to obtain a comparable mixing convergence across tree sizes. However, good mixing is likely to require more generations on large trees with large numbers of edges to evaluate. Hence, the running time for bayou is likely to be underestimated for large trees. All running times were obtained with a 2·7 GHz processor. Figure 6 displays the average elapsed time of each method, showing that previously pro-

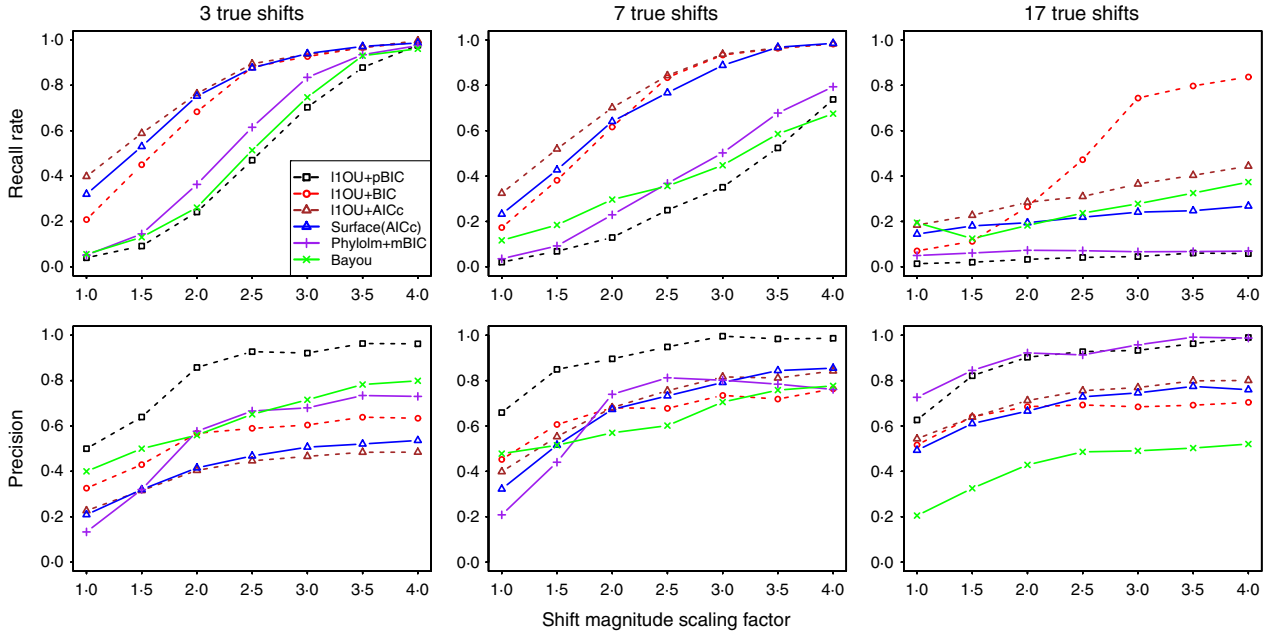

Fig. 5. Recall rate (first row) and precision (second row) of various methods to detect the position of Ornstein–Uhlenbeck (OU) shifts on the tree with 3, 7 and 17 true shifts (see Fig. 3). The magnitudes of all shifts were increased by the same scaling factor.

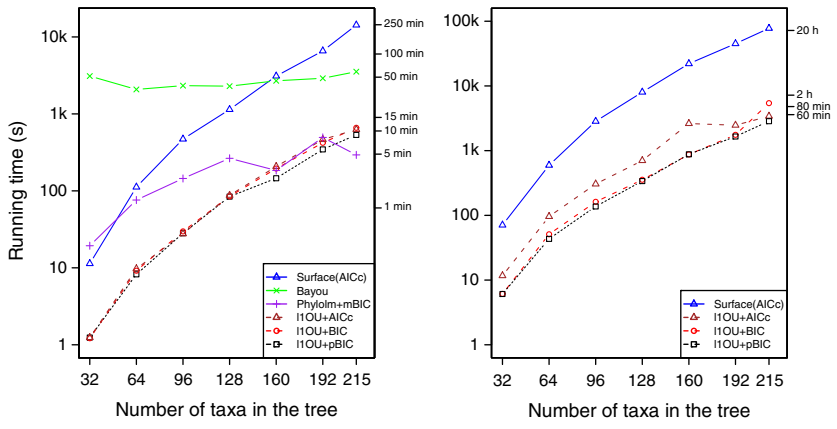

Fig. 6. Average running time of different methods for Ornstein–Uhlenbeck (OU) shift detection vs. the number of species in the tree, for data sets with a single trait (left) or four traits (right). Time is displayed on a log scale.

posed methods do not scale well to trees with a few hundred taxa. On the other hand, ℓ1ou is between one to two orders of magnitude faster than the other methods, with no loss of accuracy.

SHIFT DETECTION FROM MULTIPLE TRAITS

We conducted two simulation experiments with multiple trait data. First, we explored the effect of conducting standard PCA to reduce the problem dimension, before detecting shifts on the first PC axis only. Secondly, we explored the performance of ℓ1ou and surface when applied to multiple traits

We simulated data (100 replicates for each situation) under the same 60-taxon tree as before (Fig. 3) except that each data set contained $m = 20$ independent continuous traits simulated under OU model. For our first experiment, the true model had *no* shifts. We set $\alpha = 2$, corresponding to a moderate half-life 0.34, and $\sigma^2 = 4$ to fix the stationary variance at 1. Figure 7 shows the systematic error caused by using only the first PC, that is the axis with the largest variation in the data. As expected, some branches near the root are consistently detected as having a shift. Even though we used the most conservative method (ℓ1ou + pBIC) to analyse the first PC, at least one shift was detected near the root in 65% of the replicates, on one of the branches marked by a star. When using ℓ1ou + BIC or AIC_c, the occurrence of these false positives increased to 82% and 89%.

Secondly, we considered the same tree as before with 0, 3, 7 or 17 true shifts but with multiple traits (Fig. 3). We used $\alpha = 1$ and $\sigma^2 = 2$ and generated $m = 4$ independent traits under the OU model. We chose four traits because this is representative of a number of applications and because surface was too slow to handle 20 traits for many replicates (about 1 h per replicate).

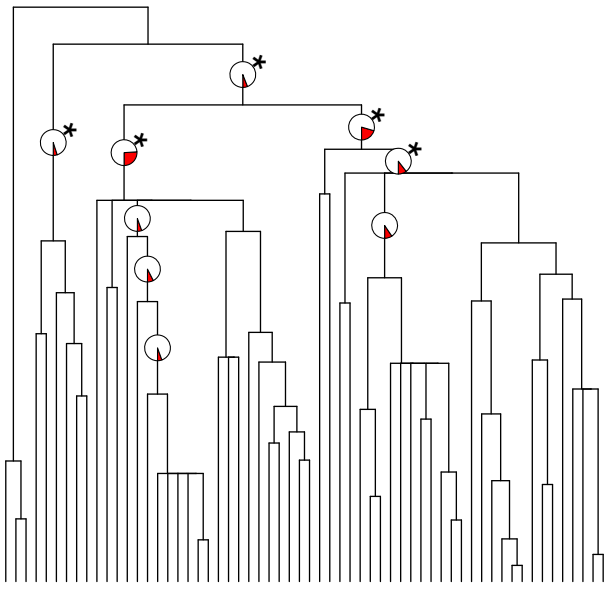

Fig. 7. Shifts detected by analysing the first standard principal component using ℓ1ou + pBIC. Data were generated under the BM model ($\alpha = 0$) with no shifts, $\sigma^2 = 4$ and 20 variables. Pie charts show the proportion of replicates for which a shift was detected on a given branch (shaded area), on branches for which this proportion was 5% or greater. Stars mark branches near the root and subtending moderate or large bipartitions, where most false shifts tended to be detected.

When no shifts were simulated, we further varied α keeping the stationary variance $\sigma^2/(2\alpha) = 1$. Figure 4 (right) shows that all methods except ℓ1ou + BIC had a few false positives. In contrast to analyses with a single trait, ℓ1ou + BIC appeared as most conservative. We then repeated the same simulations except that the four traits had residual correlation, either from correlated drift or from correlated selection. This caused an increase in the number of falsely detected shifts, for all methods (Fig. S10).

In simulations with shifts, the magnitudes shown in Fig. 3 were multiplied by +1 or −1 randomly and independently for each trait. They were then all scaled by a common factor as before, varying from 1 to 4. Bayou and the phylolm stepwise method were not applied since they cannot handle multiple traits. As expected, using multiple independent traits improved both the recall rate and the precision of all methods, compared to using a single trait (Fig. 8). Like before, surface and ℓ1ou + AIC_c were very similar and were the most liberal methods and pBIC tended to be the most conservative With 4 traits ℓ1ou + BIC was also very conservative. However, there were situations when the most liberal methods kept detecting false shifts (precision capped around 50% with three true shifts and three false shifts detected) even when the signal-to-noise ratio increased (large shift magnitudes), while the most conservative method reached both a recall rate of 100% and a precision of 100%.

We also evaluated the accuracy of shift detection when phylogenetic PCA is first applied to reduce the dimension of the data, to detect shift positions by on each pPC axis separately. For each data set generated above, we applied pPCA (assuming a BM model as proposed by Revell 2009) and applied various shift detection methods on the first axis. The multivariate version of ℓ1ou or surface on the original multivariate data had a better or comparable recall rate and precision than the same method applied to the single first pPC (Fig. S2).

Illustrations with data on *Anolis* lizards

Anolis lizards on the Caribbean islands have independently evolved a similar set of 'ecomorphs', such that species of the same ecomorph category from different islands are similar morphologically (Losos *et al.* 1998). Mahler *et al.* (2013) studied similarities among islands by considering 11 traits including body size, limb and tail lengths, and adhesive toepad lamella number across 100 species. They applied pPCA and retained the first four axes, which together explained 93% of variation. Their data and tree are available in the supplementary material of Mahler *et al.* (2013). We applied surface and ℓ1ou + pBIC, BIC or AIC_c to their four pPC traits, using the random root covariance in (2) and allowing for a maximum of 50 shifts. ℓ1ou + pBIC detected 12 shifts (in 13·8 min). Figure 9 shows that each of these shifts is supported by several of the four traits. Surface found 28 shifts (in 2 h and 12 min), which include 11 of the 12 shifts detected here (Fig. S3). The one shift

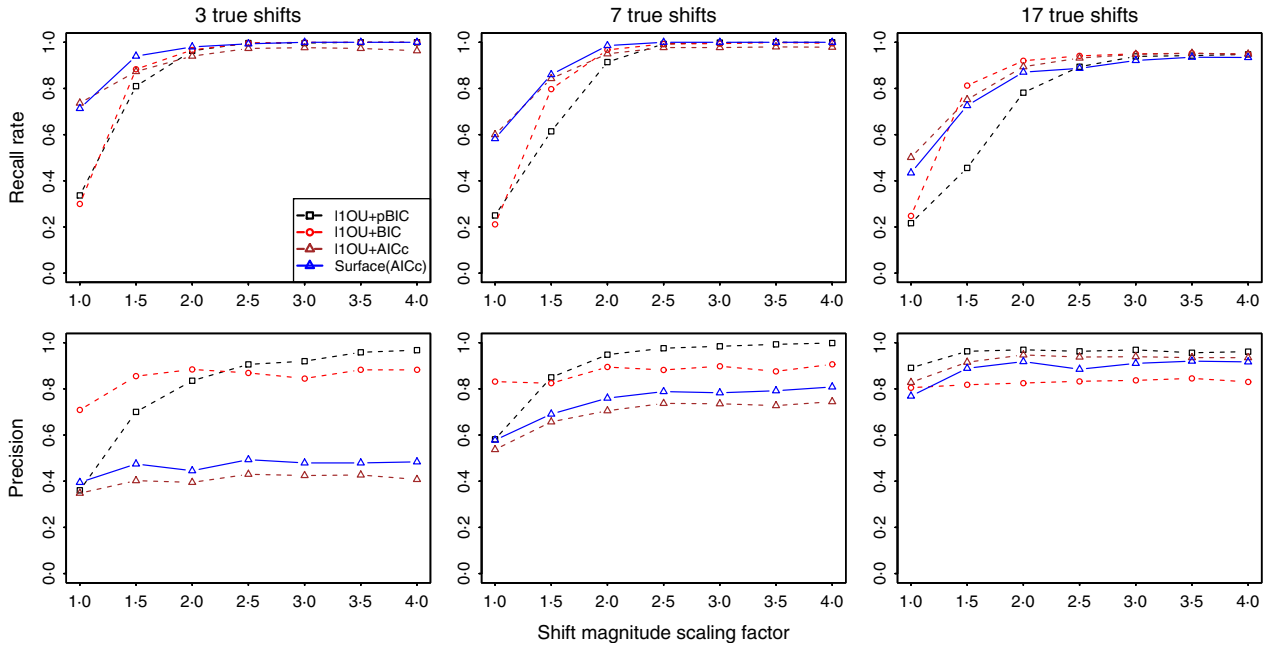

Fig. 8. Recall rate (first row) and precision (second row) of multivariate methods to detect the position of Ornstein–Uhlenbeck (OU) shifts on the 60-taxon tree with three, seven and 17 true shifts (see Fig. 3), from four traits. The magnitudes of all shifts were increased by the same scaling factor.

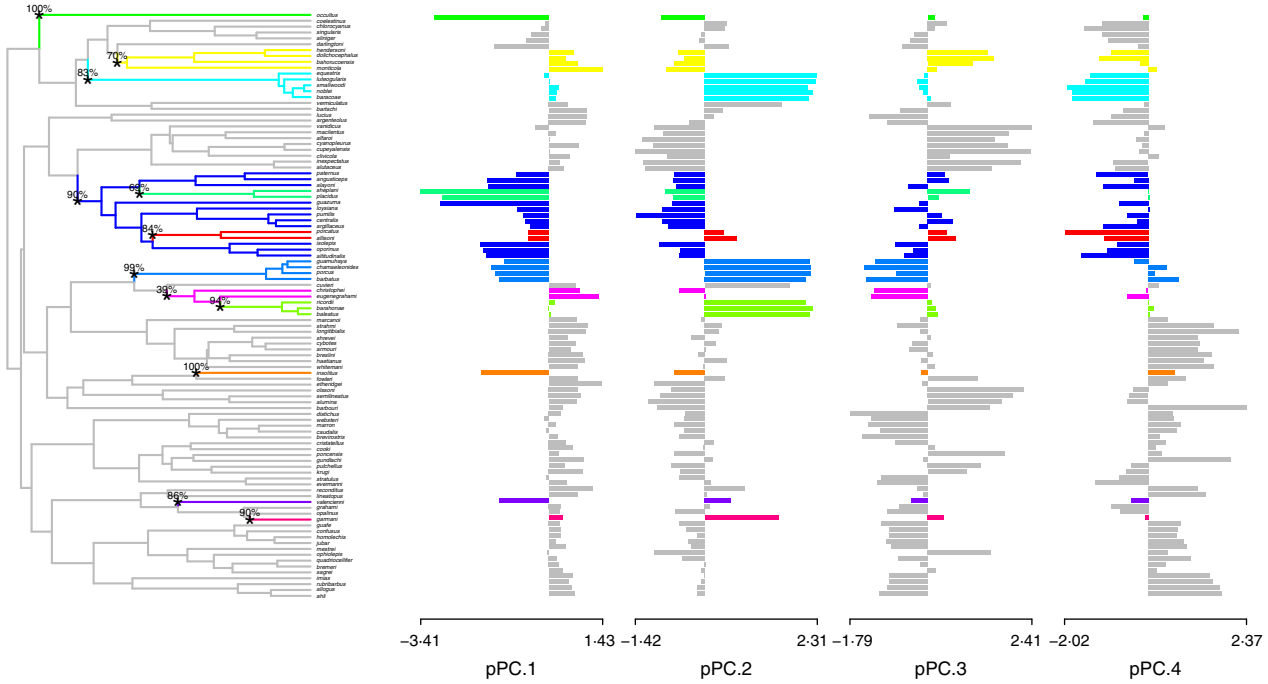

Fig. 9. Shifts in Anolis lizard morphology. Re-analysis of the four traits from Mahler *et al.* (2013) with ℓ1ou and pBIC provided support for 12 evolutionary shifts in optimum morphology under an Ornstein–Uhlenbeck (OU) process. Left: shift configuration. Edges with a shift are annotated with a star and bootstrap support. Right: bar graphs showing the four traits combined for analysis.

not detected by surface had the lowest bootstrap support (39%). All other shifts had support between 69% and 100%. The 28 shifts found by ℓ1ou + AIC_c (in 13·2 min) included all 12 shifts found by ℓ1ou + pBIC and were very similar to those found by surface up to equivalent parsimonious configurations (Fig. S4). With this many shifts, the one configuration returned by ℓ1ou + AIC_c (or by surface) is equivalent to many other configurations that define the same clustering of taxa. Therefore, this one configuration is masking a lot of uncertainty about the shift locations. Because pBIC is quite conservative, we can be more confident in its 12 shifts compared to the extra 16 shifts found by AIC_c or by surface. On these data, ℓ1ou + BIC was most conservative and did not detect any shift. Figure S5 shows the score profile plot of each method. For BIC, this profile shows a local optimum in BIC at nine shifts, seven of which were found by pBIC (Fig. S6).

Of the 12 shifts detected by ℓ1ou + pBIC, four occurred within Cuba (or as a dispersal to Cuba), five within Hispaniola, two within Jamaica and only one with Puerto Rico (or as a dispersal to Puerto Rico), based on a parsimonious geography reconstruction (Fig. S7). Overall, our results suggest that ecomorphological convergence is not as convincing statistically as previously argued. First, over half of the shifts previously detected are suspected to be unreliable or misplaced. Secondly, two of the four islands only have one or two confirmed shifts, weakening the statistical evidence for repeated convergence on separate islands

When analysing the first trait only (pPC1, which alone explained 40% of variation), fewer shifts were detected by all methods, showing the gain in detection power from combining multiple traits. Four shifts were detected with ℓ1ou + pBIC, all of which were also detected by ℓ1ou + AIC_c, which detected 16 shifts total. Using the generalized lasso (7) + AIC_c on these 16 shifts, we detected a high level of convergent evolution with a total of eight regimes only. In comparison, surface detected 12 shifts and five distinct optima, with some similarities but also marked differences (Fig. S8). These two convergent evolution models had very similar AIC_c scores however (−86·37 and −86·40), highlighting great uncertainty about the exact phylogenetic placement of shifts and convergent evolution.

Discussion

In this work, we adapted the lasso, now widely used for standard statistical model selection, to phylogenetic comparative data and the detection of shifts in the mean. The lasso penalizes parameters by their absolute values, which leads to sparse models with most parameters estimated at 0. The OU process that we used can model the response to a changing adaptation landscape, to which the lasso provides a parsimonious solution.

We also proposed a new phylogenetic criterion pBIC that explicitly accounts both for phylogenetic correlation, and for the large number of configurations with a given number of shifts k. This number of models grows extremely fast with k, leading to overfitting issues and high rates of falsely detected shifts with AIC. On the contrary, pBIC was shown to be conservative. Interestingly, the pBIC penalty for a k-shift model is not a simple function of the number of parameters, and/or of the number of configurations with k shifts (Massart 2007). The penalty depends on the best shift configuration and generalizes the notion of a shift's effective sample size (Ané 2008). In particular, shifts leading to small clades are penalized less than shifts leading to large clades, especially if phylogenetic correlation is low, because their effective sample size is smaller. Our ℓ1ou method could be combined with any further improvements to pBIC. Also, pBIC can be generalized to other models with shifted means, making it applicable to models with jumps derived from the BM process for instance (see below).

Bastide, Mariadassou & Robin (2015) recently considered the same problem and highlighted the same identifiability issues on shift configurations. They derived the exact number of non-equivalent (distinguishable) parsimonious configurations of k shifts, which could depend on the tree topology. This number, necessarily smaller than the number of ways to choose k edges, could be used to improve our pBIC derivation (affecting the term $2k \log(2n-3)$). To select k and for a single trait, Bastide, Mariadassou & Robin (2015) used a criterion penalty based on the number of distinguishable configurations, with guaranteed properties if α is known. For one trait, the maximum likelihood configuration with k shifts is found with Expectation-Maximization (Dempster, Laird & Rubin 1977), which is probably more thorough but slower than our approach.

A major strength of our phylogenetic lasso method is its speed, being one or more orders of magnitudes faster than currently existing methods. Parallelization of our implementation could further reduce its running time. This is because the set of candidate models returned by the lasso can be evaluated for pBIC in parallel; this second step is the computational bottleneck, consuming much more time than the first lasso step. To achieve fast running times, we also implemented a linear-time algorithm (Stone 2011), to obtain the square-root and inverse square-root of the covariance matrix, $\Sigma^{(\alpha)}$. This fast algorithm facilitates both the noise-whitening transformation for the phylogenetic lasso and the bootstrap procedure here, but it could have broader benefits for other applications. The matrices $\Sigma^{-1/2}$ and $\Sigma^{1/2}$ are not unique (many matrices satisfy $A'A = \Sigma$), and the matrices returned by the fast algorithm are not symmetric, but they have an advantage of interpretability: each row corresponds to an edge in the tree, including a root edge. Therefore, they provide phylogenetically corrected residuals that map onto the phylogenetic tree. Their applications include model diagnostics and visualizations (Pennell *et al.* 2015) with possible interpretation as to the cause of potential model violations. Here, phylogenetically corrected residuals might be used to detect possible model violations that might correlate with shift configurations.

Our bootstrap procedure, which uses both $\Sigma^{-1/2}$ and $\Sigma^{1/2}$, is comparable to the fully parametric bootstrap method used by Ingram & Mahler (2013) for surface. Our method is partially nonparametric, however, in that we resample the phylogenetically corrected residuals instead of sampling from the OU process, to gain some robustness to potential violation of the OU model assumptions. The results from such bootstrap procedures should be interpreted with caution, however, because they can depend heavily on the shifts simulated under the bootstrap model. If this model only uses the shifts detected conservatively with pBIC, then any true shift that went undetected will necessarily receive low bootstrap support. On the lizard data for instance, adding an extra two shifts to the pBIC configuration increased the pBIC score by 4·99 only, but resulted in greatly increased bootstrap support for the newly added shifts (from close to 0% to 63% and 62%, see Fig. S9). It might be advantageous to use the shifts detected with a more liberal criterion (AIC_c) in the bootstrap simulation model, but analyse the bootstrap data sets with a conservative criterion (pBIC). Hence, these bootstrap values should be interpreted with caution, and more work is needed

to improve parametric bootstrap methods here, when model selection is involved.

For shifts located on neighbouring edges, extra caution should be taken because of identifiability issues. For instance, the data contain no information on whether a shift is at the base on the ingroup clade vs. the outgroup clade (i.e. on either edge connecting to the root). Even if the bootstrap support for a shift is 100% at the base of the ingroup clade, the user should keep in mind that there is still complete uncertainty about the exact placement of this shift on either side of the root, or its timing along either edge. Similarly, shifts detected on two sister clades should be interpreted with caution, even if each one receives 100% bootstrap support. The exact same data could be obtained with a shift on the edge ancestral to these two sister clades, and only one subsequent shift to one of the clades. Here again, the 100% bootstrap values ignore uncertainty due to a lack of identifiability. Bayesian methods can deal with this issue much more elegantly (Uyeda & Harmon 2014), because one might place equal prior probabilities on all the non-distinguishable shift configurations. Posterior probabilities would reflect uncertainty between all these configurations, even uncertainty on the location of a shift along a given edge. Non-identifiable shift configurations might also have different posterior probabilities because their shared maximum likelihood might be achieved at different shift magnitudes, which are not necessarily equally likely a priori. Therefore, a Bayesian framework can distinguish between non-idenfiable shift configurations using biologically reasonable priors on shift magnitudes. Also, even though the posterior mean number of shifts depends on the prior number of shifts, Bayesian posterior distributions might quantify uncertainty over the various configurations with a fixed number of shifts better than bootstrap samples (see also visualization tools in Rabosky *et al.* 2014). This is because bootstrap samples are generated under a unique bootstrap simulation model, from the best estimated shift configuration only. More work could still be done to improve frequentist bootstrap procedures or other ways to quantify uncertainty, for the detection of phylogenetic shifts.

The lack of identifiability between different shift configurations is because the data truly bear on the clustering of taxa into groups. If there is evidence that two sister clades and their outgroup taxa make three different clusters each with its own adaptive optimum, then we might be able to estimate these three clusters with very high confidence. However, there will still remain complete uncertainty (without fossil data) to know how many adaptive shifts occurred, at the base of which clade they occurred, and at what time. Therefore, the proposed method should be treated as an estimation of phylogenetically consistent clusters, rather than exact shift positions.

In many applications, the underlying data are truly on a continuous scale but are discretized to facilitate analysis or to provide a taxonomic description. For instance, moss sporangium shape (Rose, Kriebel & Sytsma 2016) might be described as either 'round' or 'linear', with some subjectivity involved when scoring intermediate species, or training needed to achieve consistent scoring between different observers. For the purpose of defining thresholds to categorize continuous measurements into discrete values, our method would provide an objective and phylogenetically aware method. A liberal model selection criterion like AIC_c would be recommended, to detect sufficiently many categories and to prioritize the influence of the trait data over the species phylogenetic placement.

For the purpose of categorizing a continuous variable or for the study of adaptation, an interesting next step is to detect convergence, when different shifts lead to the same selective optimum value. For one variable, we used the generalized lasso to penalize differences between pairs of optima (Tibshirani & Taylor 2011). However, more work is needed to adapt pBIC, to correctly integrate out the constrained shifts and to account for the number of convergent configurations with k shifts. For multiple traits, the ideas of the generalized and group lasso could be combined in an ℓ_1 penalty that favours convergent regimes shared by all traits. But further work is needed because there is no fast algorithm for this form of penalty yet.

Extending our method to account for residual correlation between traits would be desirable. Simulations showed that none of the available methods are robust to the presence of correlation among traits due to drift (Fig. S10), with a marked increased of falsely detected shifts. Models for correlated traits could also combine primary response variables with potential predictors into one multivariate variable, to model variation in the response explained by shifts as well as predictors (Hansen, Pienaar & Orzack 2008; Bartoszek *et al.* 2012). However, fitting phylogenetic multivariate OU models with arbitrary selection and drift covariance matrices is difficult computationally (e.g. Clavel, Escarguel & Merceron 2015), and new theory would be needed for these models, to select the appropriate number of shifts.

Another extension of our method would be to move away from the OU model with discontinuous jumps in the adaptive optimum but continuous trait evolution. For example, the OU model leads to the same trait distribution on present-day taxa as a BM punctuated by jumps causing discontinuity in the process (at an evolutionary time-scale), provided that branch lengths in the tree are rescaled depending on α (Ho & Ané 2014). If the OU model leads to unreasonably large shifts in optimal values, a BM model might provide jumps that are more reasonable biologically, even though the two models are statistically equivalent. This is likely to occur when phylogenetic correlation is high (low α, or slow adaptation), in which case the OU model needs an unreasonably large shift in the adaptive optimum to explain a moderate jump in the observed mean. While the OU model is statistically equivalent to a process with jumps, our lasso and pBIC in (5) both penalize the magnitude of shifts in the adaptive optima, rather than the magnitude of jumps in the observed means. Hence, our model and implementation would need to be adapted to BM evolution with jumps to penalize changes in observed means rather than in adaptive shifts, through an adaptation of $X^{(\alpha)}$ and of the phylogenetic covariance. Further work could also extend this BM model with jumps to allow for an unknown level of phylo-

genetic correlation, using an extra parameter like Pagel's λ (Lynch 1991; Pagel 1999) and a similar approach to vary λ (instead of α) across different runs of the lasso.

Finally, extending our method to account for measurement error should be easiest when multiple measurements are available per species, using the observed standard errors of species means as in Ives, Midford & Garland (2007). Doing so could be most beneficial if two very closely sister species have quite different trait values, in the range of measurement error. A spurious shift to one of the two sister species might be needed to explain the trait difference if measurement error is ignored, with a possibly overestimated α (underestimated phylogenetic correlation).

Acknowledgments

This work was supported in part by the National Science Foundation (DMS 1106483). We thank Emma Krauska for helpful feedback on an early version of the manuscript and two anonymous reviewers whose constructive feedback helped us to greatly improve this work.

Data accessibility

The Anolis lizard data are available from the Supporting information of Mahler *et al.* (2013) doi 10.1126/science.1232392 (Table S2), and dryad entry doi: 10.5061/dryad.9g182. The R package ℓ1ou is available open source at https://github.com/khabbazian/l1ou.

References

Adamczak, R. & Miłoś, P. (2015) CLT for Ornstein–Uhlenbeck branching particle system. *Electronic Journal of Probability*, **20**, 1–35.

Ané, C. (2008) Analysis of comparative data with hierarchical autocorrelation. *Annals of Applied Statistics*, **2**, 1078–1102.

Ané, C., Ho, L.S.T. & Roch, S. (2015) Phase transition on the convergence rate of parameter estimation under an Ornstein–Uhlenbeck diffusion on a tree. Available at: http://arxiv.org/abs/1406.1568 (accessed 24 December 2015).

Bartoszek, K. & Sagitov S. (2015) Phylogenetic confidence intervals for the optimal trait value. *Journal of Applied Probability*, **52**, 1115–1132.

Bartoszek, K., Pienaar, J., Mostad, P., Andersson, S. & Hansen, T.F. (2012) A phylogenetic comparative method for studying multivariate adaptation. *Journal of Theoretical Biology*, **314**, 204–215.

Bastide, P., Mariadassou, M. & Robin, S. (2015) Detection of adaptive shifts on phylogenies using shifted stochastic processes on a tree. Available at: http://arxiv.org/abs/1508.00225 (accessed 24 December 2015).

Beaulieu, J.M., Jhwueng, D.-C., Boettiger, C. & O'Meara, B.C. (2012) Modeling stabilizing selection: expanding the Ornstein–Uhlenbeck model of adaptive evolution. *Evolution*, **66**, 2369–2383.

Bininda-Emonds, O., Cardillo, M., Jones, K. E., MacPhee, R. D. E., Beck, R. M. D., Grenyer, R., *et al.* (2007) The delayed rise of present-day mammals. *Nature*, **446**, 507–512.

Bühlmann, P. & Van De Geer, S. (2011) *Statistics for High-Dimensional Data: Methods Theory and Applications*. Springer Science & Business Media, Berlin, Heidelberg.

Butler, M.A. & King, A.A. (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist*, **164**, 683–695.

Clavel, J., Escarguel, G. & Merceron, G. (2015) mvMORPH: an R package for fitting multivariate evolutionary models to morphometric data. *Methods in Ecology and Evolution*, **6**, 1311–1319.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **39**, 1–38.

Eastman, J.M., Alfaro, M.E., Joyce, P., Hipp, A.L. & Harmon, L.J. (2011) A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution*, **65**, 3578–3589.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004) Least angle regression. *The Annals of Statistics*, **32**, 407–499.

Eldar, Y.C. & Kutyniok, G. (2012) *Compressed Sensing: Theory and Applications*. Cambridge University Press, Cambridge, UK.

Freckleton, R.P.& Harvey, P.H. (2006) Detecting non-Brownian trait evolution in adaptive radiations. *PLoS Biology*, **4**, e373.

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

Hansen, T.F., Pienaar, J. & Orzack, S.H. (2008) A comparative method for studying adaptation to a randomly evolving environment. *Evolution*, **62**, 1965–1977.

Harmon, L.J., Losos, J.B., Jonathan Davies, T., Gillespie, R.G., Gittleman, J.L., Bryan Jennings, W., *et al.* (2010) Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, **64**, 2385–2396.

Ho, L.S.T. & Ané, C. (2013) Asymptotic theory with hierarchical autocorrelation: Ornstein–Uhlenbeck tree models. *Annals of Statistics*, **41**, 957–981.

Ho, L.S.T. & Ané, C. (2014a) Intrinsic inference difficulties for trait evolution with Ornstein–Uhlenbeck models. *Methods in Ecology and Evolution*, **5**, 1133–1146.

Ho, L.S.T. & Ané, C. (2014b) A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology*, **63**, 397–408.

Hopkins, M.J. & Lidgard, S. (2012) Evolutionary mode routinely varies among morphological traits within fossil species lineages. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 20520–20525.

Hunt, G., Bell, M.A. & Travis, M.P. (2008) Evolution toward a new adaptive optimum: phenotypic evolution in a fossil stickleback lineage. *Evolution*, **62**, 700–710.

Ingram, T. & Kai, Y. (2014) The geography of morphological convergence in the radiations of Pacific Sebastes rockfishes. *The American Naturalist*, **184**, E115–E131.

Ingram, T. & Mahler, D.L. (2013) SURFACE: detecting convergent evolution from comparative data by fitting Ornstein–Uhlenbeck models with stepwise Akaike information criterion. *Methods in Ecology and Evolution*, **4**, 416–425.

Ives, A.R., Midford, P.E. & Garland, T. Jr. (2007) Within-species variation and measurement error in phylogenetic comparative methods. *Systematic Biology*, **56**, 252–270.

Kriebel, R., Michelangeli, F.A. & Kelly, L.M. (2015) Discovery of unusual anatomical and continuous characters in the evolutionary history of Conostegia (Miconieae: Melastomataceae). *Molecular Phylogenetics and Evolution*, **82** (Part A), 289–313.

Losos, J.B. (2009) *Lizards in an Evolutionary Tree: Ecology and Adaptive Radiation of Anoles, Volume 10*. University of California Press, Berkeley.

Losos, J.B., Jackman, T.R., Larson, A., de Queiroz, K. & Rodriguez-Schettino, L. (1998) Contingency and determinism in replicated adaptive radiations of island lizards. *Science*, **279**, 2115–2118.

Lynch, M. (1991) Methods for the analysis of comparative data in evolutionary biology. *Evolution*, **45**, 1065–1080.

Mahler, D.L. & Ingram, T. (2014) Phylogenetic comparative methods for studying clade-wide convergence. *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice* (ed. L.Z. Garamszegi), pp. 425–450. Springer-Verlag, Berlin, Heidelberg.

Mahler, D.L., Ingram, T., Revell, L.J. & Losos, J.B. (2013) Exceptional convergence on the macroevolutionary landscape in island lizard radiations. *Science*, **341**, 292–295.

Massart, P. (2007) *Concentration Inequalities and Model Selection, Volume 1896 of Ecole d'Eté de Probabilités de Saint-Flour*. Springer-Verlag, Berlin, Heidelberg.

O'Meara, B.C., Ané, C., Sanderson, M.J. & Wainwright, P. C. (2006) Testing for different rates of continuous trait evolution using likelihood. *Evolution*, **60**, 922–933.

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

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

Pennell, M.W., Fitzjohn, R.G., Cornwell, W.K. & Harmon, L.J. (2015) Model adequacy and the macroevolution of angiosperm functional traits. *The American Naturalist*, **186**, E33–E50.

Rabosky, D.L. (2014) Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PloS One*, **9**, e89543.

Rabosky, D.L., Grundler, M., Anderson, C., Title, P., Shi, J.J., Brown, J.W., Huang, H. & Larson, J.G. (2014) BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods in Ecology and Evolution*, **5**, 701–707.

Revell, L. J. (2009) Size-correction and principal components for interspecific comparative studies. *Evolution*, **63**, 3258–3268.

Rose, J., Kriebel, R. & Sytsma, K. (2016) Shape analysis of moss (Bryophyta) sporophytes: insights into land plant evolution. *American Journal of Botany*, **103**, 652–662.

Scales, J.A., King, A.A. & Butler, M.A. (2009) Running for your life or running for your dinner: What drives fibertype evolution in lizard locomotor muscles? *The American Naturalist*, **173**, 543–553.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Stack, J.C., Harmon, L.J. & O'Meara, B. (2011) RBrownie: an R package for testing hypotheses about rates of evolutionary change. *Methods in Ecology and Evolution*, **2**, 660–662.

Stone, E.A. (2011) Why the phylogenetic regression appears robust to tree misspecification. *Systematic Biology*, **60**, 245–260.

Tibshirani, R. (1996) Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58**, 267–288.

Tibshirani, R.J. & Taylor, J. (2011) The solution path of the generalized lasso. *The Annals of Statistics*, **39**, 1335–1371.

Uyeda, J.C. & Harmon, L.J. (2014) A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Systematic Biology*, **63**, 902–918.

Uyeda, J.C., Caetano, D.S. & Pennell, M.W. (2015) Comparative analysis of principal components can be misleading. *Systematic Biology*, **64**, 677–689.

Van De Geer, S.A., Bühlmann, P. *et al.* (2009) On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, **3**, 1360–1392

Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N. *et al.* (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, E4859–E4868.

Yuan, M. & Lin, Y. (2006) Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 49–67.

Zhang, N.R. & Siegmund, D.O. (2007) A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, **63**, 22–32.

Zhao, P. & Yu, B. (2006) On model selection consistency of lasso. *The Journal of Machine Learning Research*, **7**, 2541–2563.

Received 7 October 2015; accepted 21 December 2015
Handling Editor: Thomas Hansen

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. The regression model.

Appendix S2. Identifiability and uniqueness of lasso estimation.

Appendix S3. Derivation of a phylogenetic BIC.

Appendix S4. Fast algorithm for the square root covariance and its inverse.

Appendix S5. Supplementary Figures.