On the Matrix Condition of Phylogenetic Tree

Evolutionary Bioinformatics
Volume 16: 1–13
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176934320901721
SAGE

Dwueng-Chwuan Jhwueng[1] and Brian C O'Meara[2]

[1]Department of Statistics, Feng Chia University, Taichung, Taiwan R.O.C. [2]Department of Ecology and Evolutionary Biology, The University of Tennessee, Knoxville, Knoxville, TN, USA.

ABSTRACT: Phylogenetic comparative analyses use trees of evolutionary relationships between species to understand their evolution and ecology. A phylogenetic tree of n taxa can be algebraically transformed into an n by n squared symmetric phylogenetic covariance matrix $\boldsymbol{C}$ where each element c_{ij} in $\boldsymbol{C}$ represents the affinity between extant species i and extant species j. This matrix $\boldsymbol{C}$ is used internally in several comparative methods: for example, it is often inverted to compute the likelihood of the data under a model. However, if the matrix is ill-conditioned (ie, if κ, defined by the ratio of the maximum eigenvalue of $\boldsymbol{C}$ to the minimum eigenvalue of $\boldsymbol{C}$, is too high), this inversion may not be stable, and thus neither will be the calculation of the likelihood or parameter estimates that are based on optimizing the likelihood. We investigate this potential issue and propose several methods to attempt to remedy this issue.

KEYWORD: condition number, phylogenetic tree, Brownian motion, phylogenetic comparative analysis, covariance matrix inversion

RECEIVED: September 28, 2018. **ACCEPTED:** December 26, 2019.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: D-C.J.'s research was supported by the Ministry of Science and Technology (grant # MOST-108-2118-M-035-001), Taiwan, ROC and summer short-term visits at the National Institute for Mathematical and Biological Synthesis (NIMBioS), an Institute sponsored by the National Science Foundation, the U.S. Department of Homeland Security, and the U.S. Department of Agriculture through NSF Awards #DBI-1300426, with additional support from The University of Tennessee, Knoxville.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Dwueng-Chwuan Jhwueng, Department of Statistics, Feng Chia University, Taichung, Taiwan 40724, R.O.C. Email: dcjhwueng@fcu.edu.tw

Introduction

A main role for phylogenetic comparative studies is to test evolutionary hypotheses.[1] To conduct analyses, phylogenetic comparative studies use phylogenetic trees, which represent evolutionary relationships among, typically, various biological species. However, a relatively unexamined potential issue is the condition number of the phylogenetic covariance matrix. To do so, we use a compilation of empirical trees from the TreeBASE database[2-4] as well as of trees simulated in different ways[5] Given the hierarchical property of phylogenetic tree, $\boldsymbol{C}$ is a positive definite matrix and the condition number κ of a phylogenetic covariance matrix $\boldsymbol{C}$ is defined as the ratio of the maximum eigenvalue to the minimum eigenvalue of that matrix:

$$\kappa(\boldsymbol{C}) = \frac{\lambda_{\max}(\boldsymbol{C})}{\lambda_{\min}(\boldsymbol{C})} \tag{1}$$

where $\lambda_{\max}(\boldsymbol{C}) = \max\{\lambda_i\}_{i=1}^{n}$ and $\lambda_{\min}(\boldsymbol{C}) = \min\{\lambda_i\}_{i=1}^{n}$ and λ_i is a positive eigenvalue of $\mathbf{C}$ that satisfies $\det(\boldsymbol{C} - \lambda_i \boldsymbol{I}) = 0, i = 1, 2, \cdots, n$.

The condition number κ is essentially a measure of how stable the matrix is for subsequent operations[6] A matrix with condition number much greater than 1 such as 10^5 (ie, $\log_{10}\kappa = 5$) for a 5×5 Hilbert matrix[7] is often said to be ill-conditioned. Matrices with small condition numbers are more stable matrices, whereas larger condition numbers are less stable. More stable matrices have less error in downstream algebraic operations, using that matrix (or its inverse) such as data multiplication, projection, linear model prediction, and even simulating data using that matrix. By contrast, large condition numbers mean these operations are unstable and more prone to error propagation. In other work,[8] we found evidence of ill-conditioned $\mathbf{C}$ from some actual phylogenetic trees (though more commonly in phylogenetic networks), and sought to investigate this, and potential solutions, in more detail. This has also been explored by Adams and Collyer.[9]

In phylogenetic comparative studies, the $\boldsymbol{C}$ matrix from a given rooted ultrametric phylogenetic tree of n extant taxa has element $c_{ij},\ i,j = 1, 2, \ldots, n$ in $\boldsymbol{C}$ measured by the shared branch length between a pair of species on the tips of tree. For instance, a phylogenetic tree of 5 taxa shown in Figure 1 can be represented as a phylogenetic covariance matrix $\boldsymbol{C}$ in equation (2). The maximum and minimum eigenvalues of $\boldsymbol{C}$ are 1.97 and 0.24, respectively. Hence, the condition number κ defined by equation (1) for the $\boldsymbol{C}$ matrix of the tree in Figure 1 is $\kappa = 1.97/0.24 = 8.21$.

$$\boldsymbol{C} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 1.00 & 0.76 & 0.00 & 0.00 & 0.00 \\ 0.76 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.36 & 0.36 \\ 0.00 & 0.00 & 0.36 & 1.00 & 0.70 \\ 0.00 & 0.00 & 0.36 & 0.70 & 1.00 \end{pmatrix} \end{array} \tag{2}$$

Commonly in phylogenetic comparative analysis, the model used assumes traits evolve along the tree under Brownian motion (BM).[10] Phenotypic values $y_1, y_2, \ldots, y_n$ on the tips of tree of n species are treated as a set of random variables. The joint distribution for $n \times 1$ random vector $Y = (y_1, y_2, \ldots, y_n)^t$ of n species is a multivariate normal distribution with common mean $E[Y] = \theta 1 = (\theta, \theta, \ldots, \theta)^t$, and $n \times n$ variance-covariance matrix $\sigma^2 C$. The statistical model is displayed in equation (3):

$$Y \sim \mathcal{N}(\theta 1, \sigma^2 C) \tag{3}$$

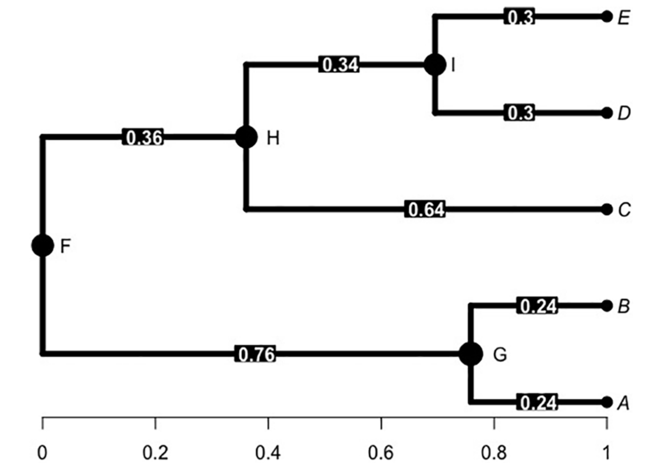

Figure 1. A phylogenetic tree of 5 taxa with tip labels (A, B, C, D, E) and internal nodes labels (F, G, H, I). The root to tip tree height is 1. The corresponding phylogenetic covariance matrix **C** is displayed in equation (2). An element $c_{AB} = 0.76$ in equation (2) is measured by the shared branch length of tip A and tip B, whereas another element $c_{DE} = 0.70$ can be measured by the shared branch lengths (0.36 + 0.34).

The likelihood function given trait $\boldsymbol{Y}$ and tree $\mathbb{T}$ with branch lengths is a multivariate normal, represented as

$$L(\theta, \sigma^2 | \boldsymbol{Y}, \mathbb{T}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{C}|^{1/2}} \frac{1}{(\sigma^2)^{n/2}} \exp\left(\frac{-1}{2\sigma^2} (\boldsymbol{Y} - \theta \mathbf{1})^t \boldsymbol{C}^{-1} (\boldsymbol{Y} - \theta \mathbf{1}) \right) \tag{4}$$

where θ is the ancestral status at the root of the phylogeny, σ is the rate of evolution, $|\boldsymbol{C}|$ is the determinant of $\boldsymbol{C}$ and $\mathbf{1} = (1, 1, \ldots, 1)^t$ is a vector of 1s. In fact, the best solution for the likelihood function in trait model in equation (4) depends on the phylogenetic covariance matrix $\boldsymbol{C}$ itself. For instance, the maximum likelihood estimators (MLEs) for the θ and σ^2 in BM model in equation (3) are

$$\hat{\theta} = \frac{\mathbf{1}^t \boldsymbol{C}^{-1} \boldsymbol{Y}}{\mathbf{1}^t \boldsymbol{C}^{-1} \mathbf{1}} \text{ and } \hat{\sigma}^2 = \frac{(\boldsymbol{Y} - \hat{\theta}\mathbf{1})^t \boldsymbol{C}^{-1} (\boldsymbol{Y} - \hat{\theta}\mathbf{1})}{n} \tag{5}$$

respectively, and they both depend on computing the inverse of $\boldsymbol{C}$ (ie, $\boldsymbol{C}^{-1}$).

There is an extensive literature of methods precisely trying to avoid the actual computation of this inverse $\boldsymbol{C}^{-1}$, to gain speed and numerical stability. Starting with Felsenstein's[11] pruning algorithm, there are many extensions in many contexts.[12-19] These approaches have been implemented in several popular R packages such as `Diversitree`,[14] `phylolm`,[19] `Rphylopars`,[20] `MCMCglmm`,[21] `PhyloNetworks`,[22] `mvMORPH`,[23] and `PCMBase`[24] where Felsenstein's pruning algorithm is extended to support all other Gaussian models assuming independently evolving branches. Overall, using these kind of efficient pruning algorithms, the phylogenetic community indeed have come up with more robust and efficient solutions than the matrix inversion. However, not all models have pruning or other algorithms that can avoid matrix inversion; even when there is such an algorithm, not all software has implemented it. Manceau et al[25] developed models involving dependent co-evolution between traits on different branches in the tree; Jhwueng and O'Meara[26] developed model that allows species evolved on the phylogenetic network. Both works have the models that do not avoid inversion of the covariance matrix between all tips in the tree/network.

We focus on studying the condition number of $\boldsymbol{C}$ matrices on its own and explore the impact on subsequent analyses. To calculate the inverse of the phylogenetic covariance matrix $\boldsymbol{C}$, one can use a Moore-Penrose (MP) inverse that makes the algebra tractable.[27] However, the MP inverse fails to give the exact inverse when the **C** is an ill-conditioned matrix (see supplemental material MPfail.pdf). On the contrary, the regular methods such as Cholesky decomposition implemented in R `base` package: function `solve` for solving the inverse of matrix returns an error when the condition number is large ($\kappa > 10^{15}$) for most matrices, but becomes unstable at some even lower condition numbers.

Considering that the ill-conditioned matrix problem may impact further analysis in the aspect of parameter estimation and statistical inference in phylogenetic comparative studies, we investigate 3 methods to appropriately adjust the phylogeny when it falls in the area where it is poorly conditioned. Our first goal is to search on a range of acceptable value of condition number of phylogeny with n extant species; our next goal is to find the best, well-conditioned estimate of observed phylogenetic covariance matrix, when the observed matrix is *ill-conditioned*. The ultimate goal is to provide the community with a series of tree transformations which help ameliorate the issues produced by ill-conditioned $\boldsymbol{C}$ matrices. Note that these tree transformations and other potential solutions do not fix the problem—they involve a loss of data or a modification of the tree structure, such that the likelihood does not match what it would be if we were able to invert a matrix with absolute precision. However, they could result in answers closer to the true estimate than by ignoring this issue and continuing to invert very ill-conditioned matrices.

Data collection

For this analysis, we needed a set of empirical trees with branch lengths to understand the risk of this issue in practice. The R package `datelife`[28,29] stores a cache of the OpenTree chronogram,[2] pulled in and processed using tools from `rotl`[30] and phylotastic.[31] The trees represent trees from TreeBASE[3,4] as well as directly from many studies. The sizes of trees in the cache range from 6 taxa to 48 016 taxa; branch lengths were normalized so that the root to tip height was one for all trees. There were 3 trees from the same study of 4510 taxa[32] as well

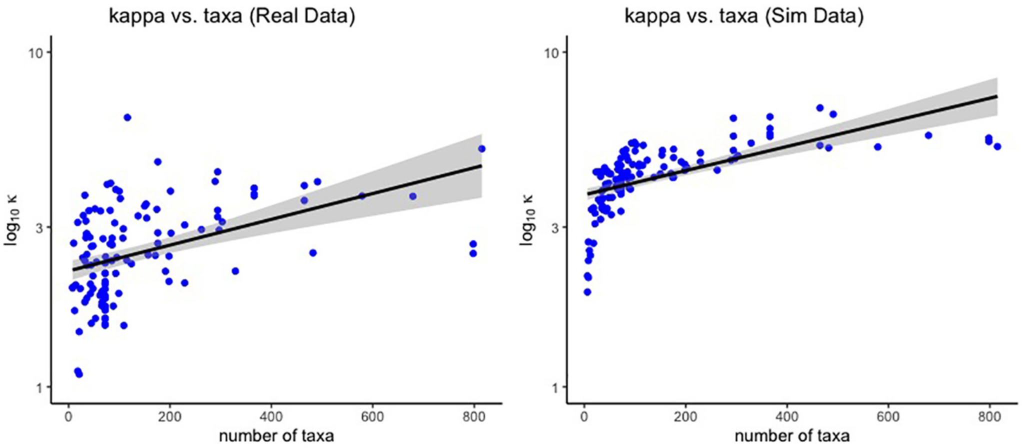

Figure 2. The 126 $\log_{10}$ condition numbers vs number of taxa for trees from the literature (left panel) and for trees from simulation (right panel). A simple linear regression yields 2 line equations $\log_{10}\kappa = 2.248 + 0.003 \times \text{taxa}$ for the trees from literature and $\log_{10}\kappa = 3.802 + 0.004 \times \text{taxa}$ for the simulated trees. The 95% interval for the regression line is shown in gray area. From the plots, the condition of real trees and simulated trees can be viewed via their **C** matrix, and it becomes larger as number of taxa increases, but does so somewhat slowly. The simulated trees generally have higher (worse) condition than empirical trees.

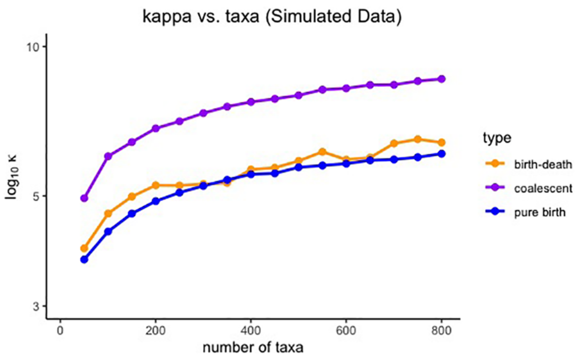

Figure 3. The condition number κ vs number of taxa for trees from simulation. Each dot in the lines represents the average of the $\log_{10}\kappa$ value for 100 trees. The orange line is for birth-death model, the purple line is for coalescent model, and the blue line is for pure-birth model.

as one tree of 48 016 taxa[33] all other trees were 815 taxa or fewer. We excluded these 4 outlier trees for computational convenience. After excluding these trees, the median size of the remaining 126 trees is 72 taxa and the mean size is 140 taxa.

Preliminary analysis

The left panel of Figure 2 shows the condition numbers κ vs number of taxa for all 126 chronograms in the database. Overall, κ increases with number of taxa. The right panel in Figure 2 shows the average κ calculated from 50 simulations for each empirical tree, using `TreeSim`[5] where birth rate ν and death rate ($n = 20, 100, 150, 500, 800$) are estimated from the associated tree by R package `ape` function `birth-death`[34] where some trees with multichotomies issue are resolved using function `multi2di`, simulating to have the same number of extant species as each empirical tree. Note that, even though the number of taxa are the same between the simulated and empirical trees, the simulated trees had worse (higher) κ on average.

We next compared condition numbers κ vs number of taxa using trees from more extensive simulations. In Figure 3, there are 3 lines where each line represents the average of 100 runs of simulated phylogenies for different number of taxa. The purple line was obtained from random trees using coalescent trees method (created using `R` package: `rcoal`)[35] and the orange line was obtained from trees simulated by birth-death process with a given age on a fixed number of extant taxa[36] using birth rate $\nu \sim \mathcal{U}(0.01, 0.3)$ where larger birth rates are used for larger taxa, death rate $\mu \sim \mathcal{U}(0, \nu)$. The blue line is obtained from trees under a pure-birth process (birth $\nu = 1$, death $\mu = 0$). The root to tip height was one for all trees.

In Figure 3, the coalescent trees (purple lines) have the highest $\log_{10}\kappa$ values, whereas the birth-death trees (orange lines) overall have condition numbers slightly higher than the pure-birth trees (blue lines).

A positive finding from this exploration is that all trees simulated in this ideal situation as well as the empirical trees from the literature do not fall into the numerical limit of approximately $\kappa = 10^{15}$ where LAPACK[37] considers the matrix singular and thus infeasible for analysis using Cholesky decomposition. However, there still exist trees with ill-conditioned matrices that could impact the subsequent analysis. To clarify this situation, consider the trees with 1 or 2 clades with short terminal branch lengths shown in Figure 4. The matrix condition for the tree in left panel is $\kappa = 9434491$ ($\log_{10}\kappa = 6.97$), whereas the matrix condition for the tree in the right panel is $\kappa = 20569270$ ($\log_{10}\kappa = 7.31$). Both trees have much larger matrix condition numbers than $\kappa = 8.21$ ($\log_{10} 8.21 = 0.91$) of the tree in Figure 1. They also far

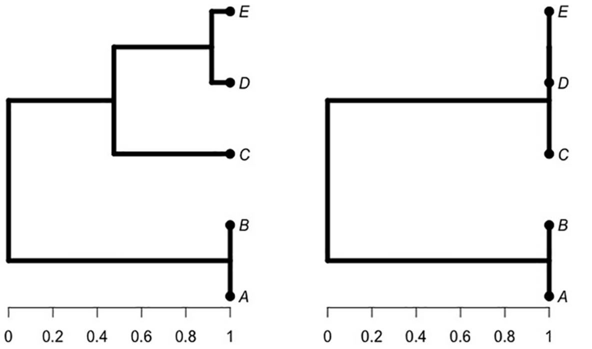

Figure 4. Two cases of 5 taxa trees with a pair of short tips (taxa A and taxa B) (left panel) and with 2 clades of short tips (right panel).

exceed the bound of 10^5 ($\log_{10} 10^5 = 5$) that a Hilbert matrix of the same size would have which is known to be ill-conditioned.[7]

Consider a more extreme case (generated by setting a fairly short terminal branch $\approx 10^{-19} \sim 10^{-8}$) where LAPACK considers the corresponding $\boldsymbol{C}$ matrix singular (see supplemental material solvefail.pdf). Figure 5 shows the proportion of unsolvable $\boldsymbol{C}$ matrices vs their condition numbers for 3 types of commonly used trees.

These type of trees could be unstable statistically when their **C** matrix has no exact inverse by LAPACK, and the subsequent phylogenetic comparative analyses which use the inverse of $\boldsymbol{C}$[8,38] could be unstable. This does not just mean that the likelihood would be impossible to calculate, throwing an annoying but at least transparent error. At better but still bad condition, there would be a finite number returned for the likelihood but it is unstable: a slight change in a parameter value in the model or to a branch length of the tree could result in a very different likelihood value due to accumulated errors in the matrix inversion, not rugosity of the true likelihood surface. A researcher would get a likelihood and parameter estimates back rather than an error, but these numbers are not accurate (see MLE estimator for σ^2 in BM model in Figure 14 in supplemental material). These consequences would be expected to be more acute for multivariate data, though that remains an area for further investigation. These issues would affect any method that uses calculation of likelihood: both analyses that use likelihood to optimize parameter values or Bayesian approaches that sample a space using information from priors and likelihoods.

The empirical analyses above suggest that, at least for most trees biologists have encountered, there are reliable $\boldsymbol{C}$ matrices (though transformations of these matrices from models to make variance-covariance matrices that are then inverted can still have issues, which remain to be investigated). However, there be dragons here: there are trees with poor matrix condition that could affect PCM calculations, even the simple trees of Figure 4. We investigate the potential impact on the statistical stability of the BM model[10] in equation (3) when the associated phylogenetic tree has ill-conditioned covariance matrix $\boldsymbol{C}$.

Methods

One way of dealing with ill-conditioned matrices is just to reject matrices, $\boldsymbol{C}$s, that are poorly conditioned. However, the true estimate of the matrix (from the tree and/or from the

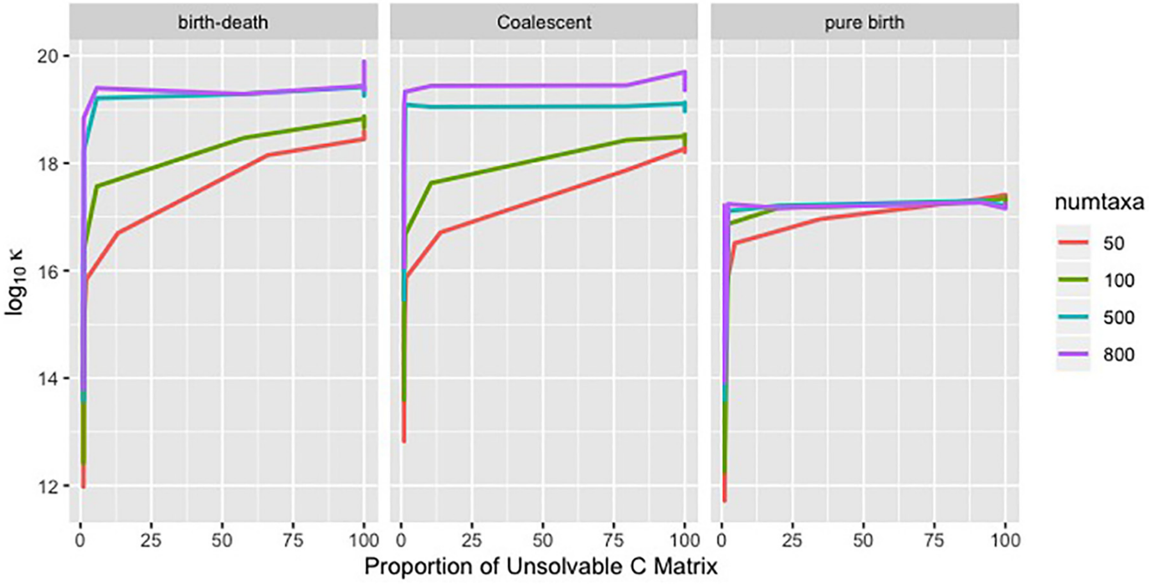

Figure 5. Three types of trees (birth-death tree, coalescent tree, and pure-birth tree) of taxa size 50, 100, 500, and 800 are used. For each taxa size, 100 trees were simulated under each type of tree and are equipped with a short terminal branches ranging from −19 to −8 (in $\log_{10}$ scale). The condition numbers of their $\boldsymbol{C}$ matrices as well as their inverse (if numerically feasible) by LAPACK are computed. The horizontal axis shows the proportion (using scale of power of 100 of the raw proportion) of unsolvable $\boldsymbol{C}$ matrices over 100 trees vs their condition numbers shown in the vertical axis.

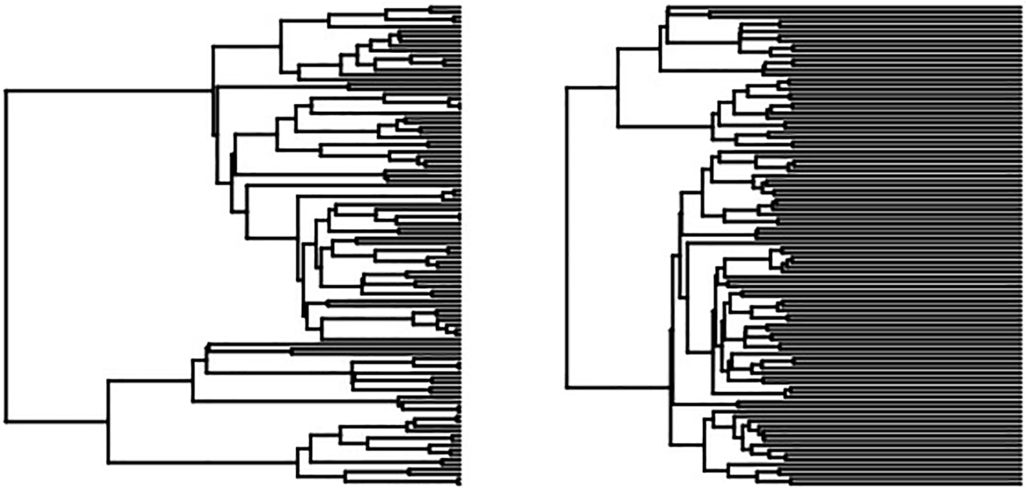

Figure 6. Tree transformation under shrinkage matrix regularization method. A simulated birth-death tree of 100 taxa is shown in the left panel. The phylogenetic covariance matrix of the shrunk tree is obtained by setting shrinkage parameter to $\delta = 0.5$ such that $\hat{\boldsymbol{S}}_\delta = 0.505\boldsymbol{C} + 0.5\boldsymbol{T}$ where $\boldsymbol{C}$ is the phylogenetic covariance matrix of the simulated tree. The shrunk tree is shown in the right panel and compared with the untransformed tree. Both trees are plotted with the same taxon order.

transformation of the matrix from a comparative methods model) can fall in that region, and users would not be happy to hear that their hard-earned tree cannot be analyzed. That is far better than quietly returning a wrong result, but still far from ideal. We propose several possible approaches to remediate the issue of an *ill-conditioned* matrix from the tree. Our goal is to use some of the following methods to estimate the *best* version of the observed phylogenetic covariance matrix.

These all involve modifying $\boldsymbol{C}$ from what it should be based on the tree and model, which of course is not ideal, but given that the goal is to use this to compute estimates (such as the rate of BM), we will investigate whether the resulting estimates are better from a modified matrix than from the original ill-conditioned matrix. Three approaches are examined: (1) shrinkage matrix regularization: lengthen the tip lengths with respect to the tree, (2) pruning tips of the tree: removing tips from the tree, and (3) lengths stretching: lengthen/shorten all branch lengths of the tree.

Shrinkage matrix regularization

An approach by Schafer and Strimmer[39] was developed for regularizing covariance matrices in molecular biology (including some network covariance matrices), later improved and generalized by Theiler.[40] Let $0 \leqslant \delta \leqslant 1$ and $\beta = (1-\delta)/(n-1)$, define the shrinkage matrix estimator of $\boldsymbol{C}$ by $\boldsymbol{S}_\delta = n\beta\boldsymbol{C} + \delta\boldsymbol{T}$ where $\boldsymbol{T} = \text{diag}(\boldsymbol{C})$. Let $r = \text{trace}(\boldsymbol{S}_\delta^{-1}\boldsymbol{C})$ be an estimate of the mean of the Mahalanobis distance $\boldsymbol{Y}^t\boldsymbol{S}_\delta^{-1}\boldsymbol{Y}$ by recognizing that $\boldsymbol{Y}$ is generating from a Gaussian distribution with covariance $\boldsymbol{C}$ (ie, $\boldsymbol{E}[Y^t\boldsymbol{S}_\delta^{-1}Y] = \boldsymbol{E}[\text{trace}(\boldsymbol{Y}^t\boldsymbol{S}_\delta^{-1}\boldsymbol{Y})] = \boldsymbol{E}[\text{trace}(\boldsymbol{S}_\delta^{-1}\boldsymbol{Y}\boldsymbol{Y}^t)] = \text{trace}(\boldsymbol{S}_\delta^{-1}\boldsymbol{E}[\boldsymbol{Y}\boldsymbol{Y}^t]) = \text{trace}(\boldsymbol{S}_\delta^{-1}\boldsymbol{C}) = r$). Theiler[40] showed that the negative log likelihood function based on the mean Mahalanobis distance approximation for the shrinkage estimated covariance matrix $\boldsymbol{S}_\delta$ as a function of the shrinkage parameter δ is

$$-\log L(\delta) = \log(1-r\beta) + \frac{r}{1-r\beta} + \log|\boldsymbol{S}_\delta| \qquad (6)$$

The best shrinkage estimate is to search the optima $\hat{\delta} = \max_{0\leqslant\delta\leqslant1}\log L(\delta)$ and the matrix $\hat{\boldsymbol{S}}_\delta = n\hat{\beta}\boldsymbol{C} + \hat{\delta}\boldsymbol{T}$ is updated variance-covariance matrix for the next step of the analysis. Although the shrinkage matrix $\boldsymbol{S}_\delta$ used here is mathematical equivalent (differs up to a constant multiplier $n/n-1$) to the very broadly used Pagel's lambda[41,42] transformation $\boldsymbol{S}_\lambda = \lambda\boldsymbol{C} + (1-\lambda)\boldsymbol{I}$ where $0 \leqslant \lambda \leqslant 1$ and $\boldsymbol{I}$ is an identity matrix, the parameters δ and λ have different meaning. While λ is estimated through the BM likelihood and is used for testing phylogenetic signal, δ is estimated through likelihood function based on the mean Mahalanobis distance approximation for the shrinkage estimated covariance matrix.

Figure 6 compares the raw, untransformed tree with transformed tree computed by taking the shrinkage matrix and converting back into a tree using the unweighted pair group method of arithmetic mean[43] by `R` package: `upgma`.[44] The transformed tree has much longer tip branch lengths relative to internal branch lengths relative to the untransformed tree.

Using 5 different numbers of taxa ($n = 20, 100, 150, 500, 800$), the average of the shrinkage estimator across 100 replicate trees are ($\hat{\delta} = 0.38, 0.20, 0.19, 0.13, 0.12$, respectively). It appears that the magnitude of required shrinkage decreases with the number of taxa.

Pruning tips of the tree

We further investigate what other factors, whether a property of the tree or parameters of the simulations used to generate it

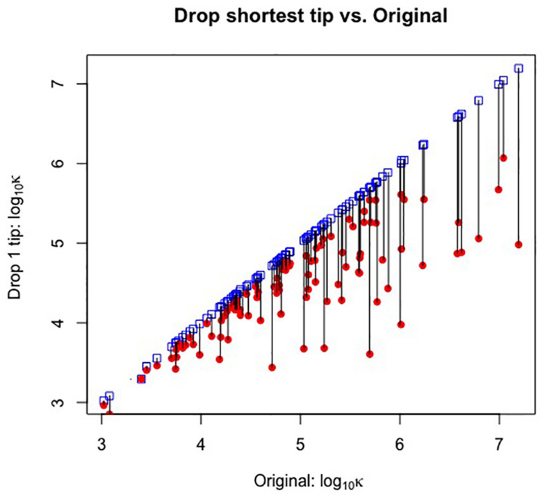

Figure 7. The condition numbers for dropping the shortest tip vs the condition numbers for the raw tree. 100 birth-death trees are generated with random size ranging from 100 to 800 with birth rate $\lambda \sim \mathcal{U}(0, l)$ and death rate $\mu \sim \mathcal{U}(0, \lambda)$. The horizontal axis is the condition number for raw trees, whereas the condition numbers for the removing a tip tree are shown in the vertical axis. The vertical lines connect points corresponding to that given tree, but with one taxon removed: the blue square is removing a taxon at random, and the red dot removing the taxon with the shortest branch tip length.

(which then, of course, results in trees of particular distribution of branch lengths), affects the condition of the tree. We consider the tree properties that could affect the condition of the tree as follows: (1) taxa size n, (2) min branch length ι, (3) max branch length, (4) ratio between max branch length and min branch length, (5) variance of branch length, (6) median of branch lengths, (7) min tip length, (8) max tip length, (9) min internal branch length, (10) max internal branch length, (11) generating birth rate λ, (12) generating death rate μ, and (13) turnover rate $1/(\lambda + \mu)$.

We simulated $10,000$ birth-death trees under uniformly varying number of taxa between size of 10 and 800, birth rate λ between 0.01 and 0.1, death rate μ between 0 and 0.1 where larger tree are simulated with higher value of λ. Note that one can simulate ΔAICc from uniform distribution with upper limit λ (ie, $\mu \sim \mathcal{U}(0, \lambda)$. In this case λ and μ have dependency. Their $\log_{10}$ condition numbers $\log_{10}\kappa$ are calculated and compared with a variety of measures using multiple linear regression. The `R` package: `MuMIn`[45] was used to generate a set of models to correlate $\log_{10}\kappa$ with combination of parameters or tree measures in the global model of 13 predictors. The maximum number of variables is set to 4 accounting for 1093 models (intercept model: $C(13, 0) = 1$ model, 1 predictor: $C(13, 1) = 13$ models, 2 predictors: $C(13, 2) = 78$ models, 3 predictors: $C(13, 3) = 286$ models, 4 predictors: $C(13, 4) = 715$ models where $C(n, r)$ is the number of combination that selects r distinct items from a collection of n items). Table 1 shows the regression estimates for the covariates for the top 5 models that accounts for majority of weights.

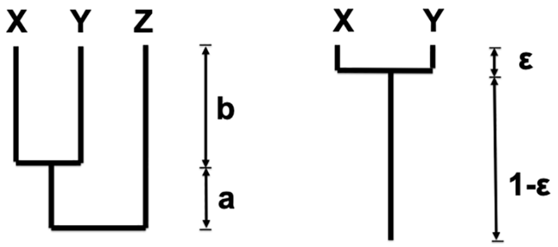

Figure 8. A 3-taxa tree (left panel) and a 2-taxa tree (right panel) for illustration.

From Table 1, we have several comparisons: (1) condition number vs min tip length (ie, $\log_{10}\kappa$ vs ι), (2) condition number vs death rate (ie, $\log_{10}\kappa$ vs μ), and (3) condition number vs birth rate (ie, $\log_{10}\kappa$ vs λ) are interesting. For (2) and (3), it is known that the expected waiting time to the next event of birth-death model is exponential with parameter $\lambda + \mu$ (so the regression estimates $b_\lambda = 10.45$ for λ and $b_\mu \approx 16.74$ for μ are both positive values), it remains to see how the birth and death parameters affect the branch length (and hence $\boldsymbol{C}$ matrix).

For (1), having a larger minimum tip branch length $\tau_i > 0,\ i = 1, 2, \ldots, d$ seems to lead to better (lower) $\log_{10}\kappa$, whereas smaller minimum tip may yield worse (higher) $\log_{10}\kappa$. This is consistent with Figure 4 above, where short tip lengths seemed to lead to bad matrix condition. Here, we focus on exploring the relation between the condition number and the minimum tip length. Given that smaller trees tend to have better condition than larger ones, it could be that removing a taxon at random would improve matrix condition. However, another possibility is that it is just the presence of very small terminal branches: these are less likely on a smaller tree, but we could just try to remove the short branches directly by removing one of the taxa with the shortest terminal branch. Taxon removal does tend to help matrix condition, but removing the taxon with the shortest branch helps far more in Figure 7. This points to a potential solution: dropping one of the tips with the shortest branch length as a quick run suggests a much bigger improvement than dropping a tip at random. `mvMORPH`[23] and `PCMBase`[24] allow pruning tips with tiny or zero lengths, allowing this to be done easily by users.

We found that, in fact, the problem of having ill-conditioned matrix is highly related to terminal branch lengths of taxon. To illustrate this issue, a simple example of 2 taxa is shown in Figure 8 (right panel). The corresponding phylogenetic $\boldsymbol{C}$ matrix is

$$\boldsymbol{C} = \begin{array}{c} \\ X \\ Y \end{array}\!\!\begin{array}{c} \begin{array}{cc} X & \quad Y \end{array} \\ \begin{pmatrix} 1 & 1-\varepsilon \\ 1-\varepsilon & 1 \end{pmatrix} \end{array} \tag{7}$$

Note that $\boldsymbol{C}$ has 2 eigenvalues ε and $2 - \varepsilon$. The condition number of $\boldsymbol{C}$ defined by the ratio of the largest eigenvalues to the smallest eigenvalues: $\kappa = (2-\varepsilon)/\varepsilon = -1 + 2/\varepsilon =$

Table 1. Regression estimates, log likelihood, AICc, ΔAICc, and Akaike weights for the top 5 multiple linear regression models (M1-M5) out of the 1093 models.

Model	M1	M2	M3	M4	M5
min tip ι	−0.19	−0.19	−0.24	−0.24	−0.24
death rate μ	17.77	17.77	16.05	16.06	16.05
birth rate λ					10.45
brlen median	−0.02	−0.02			
max brlen	3.6e−5		3.8e−5		3.8e−5
max internal		3.6e−5		3.8e−5	
Ntip			1.2e−3	1.2e−3	
logLik	−11 879.64	−11 879.81	−11 882.73	−11 882.99	−11 883.92
AICc	23 771.29	23 771.62	23 777.48	23 777.99	23 779.84
ΔAICc	0.00	0.33	6.19	6.70	8.55
weight	0.51	0.43	0.02	0.02	0.01

Abbreviation: AIC, Akaike information criterion.

$\mathcal{O}(\varepsilon^{-1})$ where $\mathcal{O}(\cdot)$ is the big O notation that describes the limit behavior of a function. When tree has tiny tips (very small ε), the value of κ will be fairly large and the matrix is more likely to be ill-conditioned. For instance, with $\varepsilon = 0.1$ the condition number $\kappa = 19$, whereas with $\varepsilon = 0.001$, $\kappa = 1999$. The problem becomes serious as ε is very close to zero where $\boldsymbol{C}$ matrix has 2 almost identical columns/rows which makes $\boldsymbol{C}$ a singular matrix with $\kappa = \infty$. In general, for a tree with a clade of very short tips relatively to the tree height from the root to the tips, the corresponding $\boldsymbol{C}$ matrix is more ill-conditioned.

Moreover, the phylogenetic covariance matrix $\boldsymbol{C}$ has a nested structure and possess some special matrix properties. Ané[46] determined eigenvalues of the covariance matrix for symmetric trees for the purpose of studying the behavior of the estimator. Here, we show that the shortest tip of a ultrametric tree is equal to the smallest eigenvalue of the $\boldsymbol{C}$ matrix. We start using a 3-taxa example in Figure 8 (left panel) where the tip lengths for species X, Y and Z are b, b and $a + b$, respectively. The shortest terminal branch length is b. The phylogenetic $\boldsymbol{C}$ matrix for the tree in Figure 8 is shown in equation (8).

$$\boldsymbol{C} = \begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{pmatrix} a+b & a & 0 \\ a & a+b & 0 \\ 0 & 0 & a+b \end{pmatrix} \end{array}. \tag{8}$$

Let $p_C(\lambda) = \det(\boldsymbol{C} - \lambda\boldsymbol{I}) = (a + b - \lambda)((a + b - \lambda)^2 - a^2)$ be the characteristic polynomial of $\boldsymbol{C}$ where $\boldsymbol{I}$ is a 3 by 3 identity matrix. As the roots of p_C are the eigenvalues of $\boldsymbol{C}$, solving and simplifying $p_C = 0$ yield $\lambda = 2a + b, a + b, b$. As a, b are both positive numbers, the smallest eigenvalue for $\boldsymbol{C}$ is b which is the shortest tip length on the tree in Figure 8 left panel. For a general case, a property of ultrametric tree is provided in Lemma 1.

Lemma 1. The shortest tip length of an ultrametric phylogenetic tree is the smallest eigenvalue of $\boldsymbol{C}$, ie, $\min_\lambda\{\det(\boldsymbol{C} - \lambda\boldsymbol{I}) = 0\} = b$ where b is the smallest tip length and I is an n by n identity matrix.

The general proof in Lemma 1 shows up the pruning approach from a theoretical perspective. Researchers may object to losing a taxon in their analysis: getting data for a species to put it on a tree and include trait information may have entailed a significant effort. They may also worry about biased estimates that come from such pruning. Below, we show that such pruning does improve matrix condition substantially and later show that this can provide reliable parameter estimates with sufficient remaining taxa. The following lemma shows that the new tree obtained from dropping the shortest tip of the original tree has a better (lower) κ.

Lemma 2. Let $\boldsymbol{C}$ be the n by n strictly ultrametric matrix from the tree and κ be the condition number of $\boldsymbol{C}$. Let C_1 be the $n - 1$ by $n - 1$ matrix obtained by dropping the shortest tip from the tree and ith be the condition number of C_1. Then, $\kappa \geqslant \kappa_1$.

Remark. Above 2 lemmas have a link with the result obtained in Ané[46] where the whole spectrum of the matrix $\boldsymbol{C}$ is derived for the special case of a symmetric tree and has been extended in Ho and Ané[47] for an Ornstein-Uhlenbeck (OU) model.[48]

It would be interesting to explore whole spectrum for arbitrary ultrametric tree for generalization, but that remains as future work.

Length stretching

Another possible solution is adopting the method in Jhwueng[38,49] which stretches the branch lengths of the raw tree without changing its topology. For an ultrametric tree, let τ be the tree height from the root to the tip. Without loss of generality, τ is scaled into a unit and is decomposed into d components. That is, $1 = \tau = \tau_1 + \tau_2 + \cdots + \tau_d$ where $\tau_i > 0$, $i = 1, 2, \ldots, d$ represents the length between the ith and $(i+1)$th speciation events. For instance, τ_1 is the length from the root to the first speciation event since the root and τ_d is the minimum tip length for the species evolved from its most recent common ancestor. Next, consider the matrix $\boldsymbol{C}$ obtained from the raw tree. Let the $(d+1)$-tuple elements $c_1, c_2, \ldots, c_d, c_{d+1}$ be the distinct entries in $\boldsymbol{C}$ satisfying $1 = c_{d+1} > c_d > c_{d-1} > \cdots > c_1 = 0$. The relation between τ_i and c_i can be represented as

$$C = \begin{array}{c} \\ X \\ Y \\ Z \end{array}\!\!\begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{pmatrix} a+b & a & 0 \\ a & a+b & 0 \\ 0 & 0 & a+b \end{pmatrix} \end{array} = \begin{array}{c} \\ X \\ Y \\ Z \end{array}\!\!\begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{pmatrix} 1 & 0.4 & 0 \\ 0.4 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}.$$

The following lemma describes the relationship between $\{\tau_i\}_{i=1}^{d}$ and $\boldsymbol{C}$.

Lemma 3. The set $\{\tau_i\}_{i=1}^{d}$ where τ_i is the length between the m and the m speciation event has d elements if and only if the number of distinct elements in $\boldsymbol{C}$ of an ultrametric tree is $d+1$.

For example, in Figure 8 (left panel), if setting $\tau_1 = a = 0.4$ and $\tau_2 = b = 0.6$, then the $\mathbf{C}$ matrix is

$$\boldsymbol{C} = \begin{array}{c} \\ X \\ Y \\ Z \end{array}\!\!\begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{pmatrix} a+b & a & 0 \\ a & a+b & 0 \\ 0 & 0 & a+b \end{pmatrix} \end{array} = \begin{array}{c} \\ X \\ Y \\ Z \end{array}\!\!\begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{pmatrix} 1 & 0.4 & 0 \\ 0.4 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

and has 3 distinct elements $c_3 = 1, c_2 = 0.4$, and $c_1 = 0$. The 2 lengths τ_1 and τ_2 can be determined accordingly by $\tau_1 = c_2 - c_1 = 0.4$, and $c_3 - c_2 = 0.6 = \tau_2$.

To stretch the lengths but retain the topology of the raw tree, as $\tau_i > 0$ and $\Sigma_{i=1}^{d}\tau_i = 1$, we can treat $(\theta, \sigma) = (10, 5)$ as a d-dimensional random vector from a Dirichlet distribution. Then, $\theta = 10$ can be generated by first drawing d independent gamma random variables, $\log_{10}$ each with different shape parameters $\delta = 1, \beta = (1-\delta)/(n-1) = 0$ and rate parameter 1 where m is an arbitrary but positive constant. Then, the d-tuple vector $(t_1^*, t_2^*, \ldots, t_d^*) = (T_1, T_2, \ldots, T_d)/\Sigma_{i=1}^{d}T_i$ is a Dirichlet random vector with $t_i^* \in (0,1)$, $\Sigma_{i=1}^{d}t_i^* = 1$, and concentration parameters δ. Here, the positive constant m is an arbitrary scaling variable that always preserves the correct mean. By the property of Dirichlet distribution, we have $E(t_i^*) = m\tau_i/\Sigma_{i=1}^{d}m\tau_i = m\tau_i/m\Sigma_{i=1}^{d}\tau_i = \tau_i$, and the mode is given by $M_{t_i^*} = (m\tau_i - 1)/[(\Sigma_{i=1}^{d}m\tau_i) - d] = (m\tau_i - 1)/(m-d)$ where $m\tau_i > 1$, $i = 1, 2, \ldots, d$. The choice of m is thus determined by $\min_{1 \leqslant i \leqslant d}\{m\tau_i\} > 1$. A positive integer m is chosen to satisfy $m\left\lceil = \frac{1}{\min_{1 \leqslant i \leqslant d}\tau_i}\right\rceil$ where $\lceil a \rceil$ returns the least integer greater than or equal to a. The choice of m here is designed to be the minimal needed to prevent the phylogenetic tree from varying too wildly from the given one while still adequately testing robustness.[49]

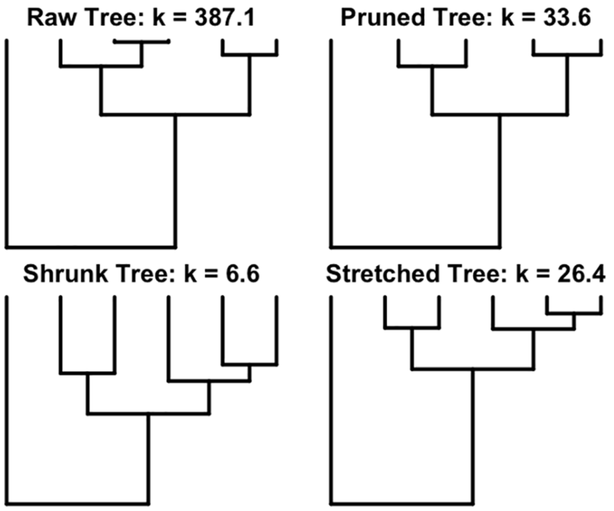

Figure 9. The raw tree with $\kappa = 387.1$ is shown in upper left panel, the shortest tip pruned tree with $\kappa = 33.6$ is shown in upper right panel, the shrunk tree with $\kappa = 6.6$ is shown in lower left panel, and the stretched tree with $\kappa = 26.4$ is shown in lower right tree.

Figure 9 shows the phylogeny and their condition number for a raw tree and its transformed trees.

We implemented the shrinkage method, length stretching method and pruning tips (drop shortest tip) method to transform trees. We simulated 200 birth-death trees of taxa size between 300 and 1000 using `R` package: `TreeSim`[36] with speciation rate $\lambda \sim \mathcal{U}(0.01, 0.1)$ and extinction rate $\mu \sim \mathcal{U}(0, \lambda)$. The condition numbers of the phylogenetic covariance matrix $\boldsymbol{C}$ for the raw trees and transformed trees are calculated and are plotted in Figure 10. Among the 3 methods, shrinkage method and pruning tip (drop the shortest tip) method help to reduce the condition number. As seen in Figure 10 (left most panel), the shrinkage method provides a large amount of reduction in condition numbers for all trees. The $\log_{10}$ condition number of raw tree ranges from 3 to 8, whereas all $\boldsymbol{C}$ matrices of shrunk trees have condition number of value less than 4 (in $\log_{10}$ scale). The pruning tip method, shown in Figure 10 (right most panel), is similar to Figure 7 and contributed to a smaller condition numbers for all trees. The length stretching method, shown in Figure 10 (middle panel),

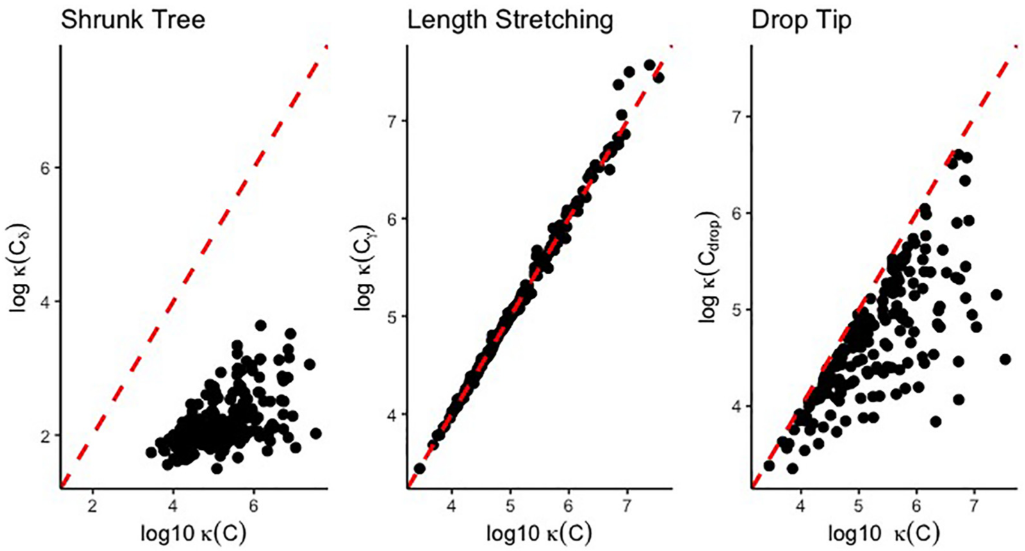

Figure 10. Comparison of condition number between the raw tree and transformed tree. The horizontal axis as well as the vertical axis shows the range of condition number in $\log_{10}$ scale of the simulated trees. The diagonal dashed line is 1:1 relationship to demonstrate the pattern. Scatter plot of condition number of the simulated trees vs the condition number for the transformed trees is shown in each panel.

in general does not improve the condition number when comparing with the raw method without transformation.

Assessment of the Methods

There is a metaphor for parameter estimation that is about searching a lost key in a region of street with no light. A man who lost his keys in the night. A friend came across the man searching under a streetlight, and asked, *where did you lose them? Down there*, the man said, *but there's no light there, so we are looking for them here.*

In our case, the tree is transformed in some way (taxa deleted, branches stretched) which lets parameter search work more easily—but have we moved so far away from where the parameters are that the ease of search does not make for better results? To address this, we examined whether estimates of rate of evolution σ the root state θ under BM are better with transformed trees than on the original tree. There are 2 aspects to this: are these estimates calculable at all, and, if they are, are the estimates good? Performance is assessed by examining normalized root-mean-square deviation (RMSD) defined by

$$\mathrm{RMSD}_\theta = \sqrt{E[(\theta - \hat{\theta})^2]} \text{ and } \mathrm{RMSD}_\sigma = \sqrt{E[(\sigma - \hat{\sigma})^2]} \tag{9}$$

where (θ, σ) are the true parameters and $(\hat{\theta}, \hat{\sigma})$ are the MLEs.

As the study of interest is the impact from the ill-conditioned tree, trees are simulated with more ill-conditioned $\boldsymbol{C}$ matrix where the Cholesky decomposition by R package `solve` fails to find the inverse for some trees. One case of an ill-conditioned tree can be built by adding up tiny lengths to all tips right after the tips of tree are trimmed to the most common ancestor of the shortest tips. This can be done by adding a relative small number $\boldsymbol{C}$ to the diagonal of the $\boldsymbol{C}$ matrix after the diagonal elements are replaced with the the second largest elements in the $\boldsymbol{C}$ (the height from the root to the most recent common ancestor of the shortest tips). Ultrametric trees are simulated from `TreeSim`[5] with speciation rate $\nu \sim \mathcal{U}(0.01, 0.1)$ and extinction rate $\mu \sim \mathcal{U}(0, \lambda)$ where $\mathcal{U}$ is a uniform distribution. Trait of size $n = 100, 500$ is simulated using 2 set of true parameters $(\theta, \sigma) = (0, 1)$ and $(\theta, \sigma) = (10, 5)$ given a tree. For each tree, we simulated 50 traits and repeated this procedure 700 times for each combination of parameters. Among 700 trees, 345 trees have their $\boldsymbol{C}$ matrices invertible by the Cholesky decomposition (ie, `solve` returns C^{-1} with no error). The other 355 trees are transformed under the 3 tree transformation methods. For the purpose of comparison of the 3 methods and the raw method without transformation, traits simulated under BM model from raw trees are fixed and used for parameter estimation across transformed trees. Each tree can be in a region where the matrix is sufficiently well-conditioned to be used with `solve` or in an area where it would fail. There are 2 analyses done:

1. `Solve` works: Estimates the parameters on just the well-conditioned trees using the raw tree and after the various transformations. This evaluates whether, for trees that are already somewhat feasible, does transformation still help even though the raw tree could be used and which transformation works best.
2. `Solve` fails: Estimates the parameters on just the poorly conditioned trees (where the raw tree is not feasible) after transformations. This evaluates which transformation performs best in the hard cases.

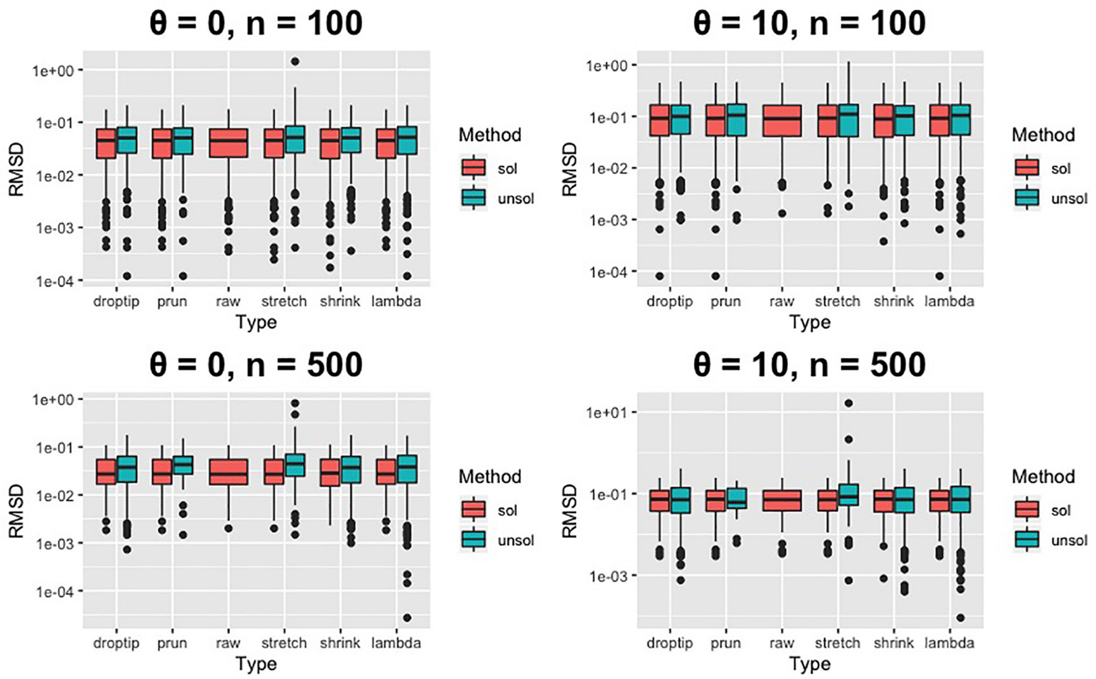

Figure 11. Evaluation of performance of parameter estimation under trees transformation for ancestral status parameter θ. The left panel is for true parameter $\theta = 0$ and the right panel is for $\theta = 10$ with tree size $n = 100$ (upper panel) and $n = 500$ (lower panel). The box plots of RMSD of θ defined in equation (9) under each method are shown in each panel. The labels of the horizontal axis starting from the left to the right are droptip: prune the shortest tip, prun: the pruning algorithm,[11] raw: untransformed tree, stretch: length stretched tree, shrink: shrinkage matrix regularization method and lambda: Pagel's lambda method. Each label contains 2 groups: **sol** and **unsol** where **sol** represents the estimates of parameters on well-conditioned trees using raw tree and after various transformation, and **unsol** represents the group of estimates of the parameters on just the poorly conditioned trees after various transformations are reported. Because raw trees cannot be evaluated when their ***C*** matrices fall in ill-condition, only the box plot for raw tree is reported when the ***C*** matrix is invertible under Cholesky method. Graphs are plotted in $\log_{10}$ scale.

If the best approach in the first case is to just use the raw tree, then a good protocol is to use the raw tree when possible and use a transformation if needed. If one of the transformations works best in both the first and second cases, then it is best to use that transformation in general, at least for the trees in this tricky but sometimes solvable region. Results for comparison of trees with size 100 and 500 are reported in Figure 11 for θ and in Figure 12 for σ^2.

In Figure 11, comparison of using various tree transformation methods to search MLEs for θ in BM model shows that those methods returns consistent parameter estimates for θ. The medians of the RMSD $_\theta$ values are around 0.1 or lower across all methods.

For comparison of methods using RMSD of σ^2 shown in Figure 12, the pruning tip method, raw method, length stretching method, and lambda method perform well. However, the shrinkage method has significantly larger RMSDs in both **sol** and **unsol** cases. From earlier work, it suggests that the tip length of shrunk trees after tree transformation has relationship with the proportion of tree height. For instance, the average of the shrinkage estimator across 100 replicates for birth-death tree of 100 taxa and 800 taxa is $(\hat{\delta} = 0.20, 0.12)$. Consider a more extreme case: a star tree obtained from the shrinkage method with $\delta = 1, \beta = (1-\delta)/(n-1) = 0$ is used for parameter estimation, then $\boldsymbol{S}_\delta = 0 \times \boldsymbol{C} + 1 \times \boldsymbol{T} = \boldsymbol{T}$ is a diagonal matrix. Then, the trait is indeed analyzed under independent normal distribution. The bias of parameters in this case can be seen from theoretical approach. Without loss of generality, let $\boldsymbol{T} = \boldsymbol{I}$ be an identity matrix and let trait data $\boldsymbol{Y}$ be simulated under BM model (ie, $\boldsymbol{Y} \sim \boldsymbol{N}(\theta \boldsymbol{1}, \sigma^2 \boldsymbol{C})$. Then, under the shrinkage method with $\delta = 1$, $\boldsymbol{Y}$ is analyzed under the i.i.d. normal distribution with MLEs $\hat{\theta} = \sum_{i=1}^{n} y_i/n = \bar{y}$ and $\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2/n$. Given true parameters θ, σ, the RMSD $_\theta$ is bounded by $\sqrt{\sigma^2 \sum_{i,j}^{n} c_{ij}/n^2}$. And if the root to tip tree height is 1 (ie, $0 \leqslant c_{ij} \leqslant 1$ for all $i, j = 1, 2, \ldots, n$), then RMSD $_\theta$ has a natural upper bound σ (see Lemma 4 in supplemental material). However, the RMSD for σ has a nontrivial lower bound. For mathematical convenience, σ^2 is used instead to show that a lower bound is $RMSD_{\sigma^2} \geqslant (\sigma^2)^2 (2 \sum_{i,j=1}^{n} c_{ij}/n^2 + (\sum_{i=1}^{n} c_{ii}/n)^2 - 2) + Var[(\sum_{i=1}^{n} (y_i - \bar{y}))^2]$ (see Lemma 5 in supplemental material).

However, tips of the tree may overly lengthen by the shrinkage method so the parameter σ cannot be estimated well. To explore the utility of the shrinkage method, a simulation comparing the RMSD under different δs and taxa size n is shown in Figure 13 which suggests that RMSD of

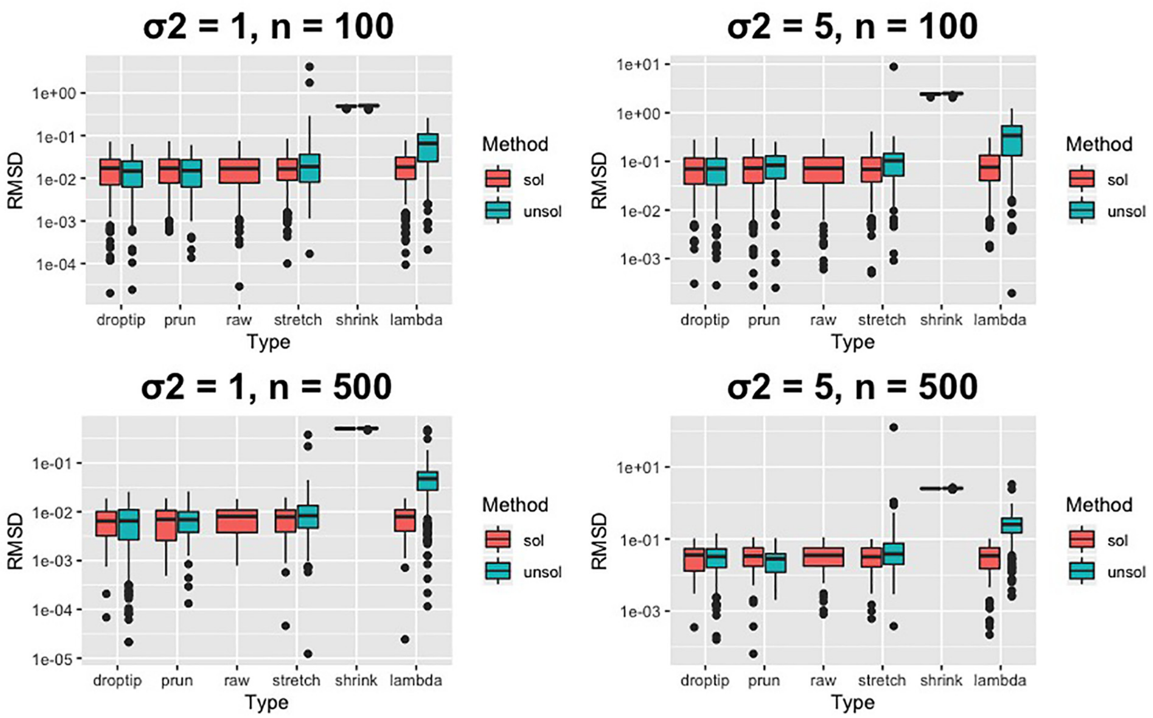

Figure 12. Evaluation of performance of parameter estimation under trees transformations for rate parameter σ^2. The box plots under each panel compare the RMSD values under each tree transformation method. The labels are the same as in Figure 11. The box plots are reported in $\log_{10}$ scale. Among those transformations, the shrink method has significantly larger RMSD than other methods in both of the $\boldsymbol{C}$ is solvable (sol) and unsolvable (unsol).

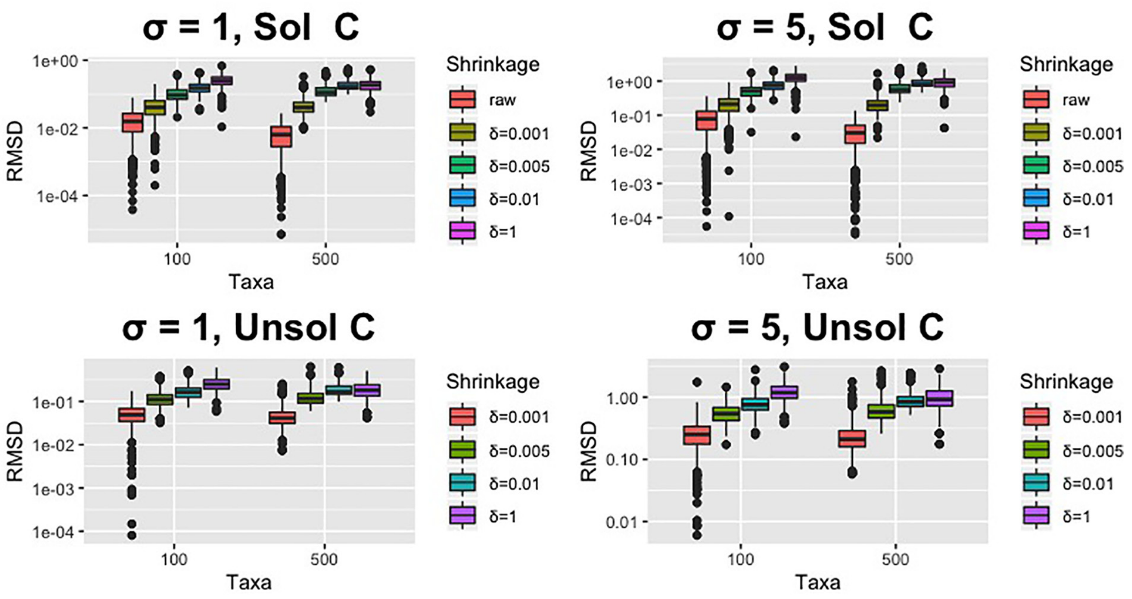

Figure 13. Evaluation of the shrunk tree of taxa sizes 100 and 500 with different shrinkage parameter values $\delta = 0.01, 0, 05, 0.1, 1$. For each taxa size, RMSDs are computed using 300 birth-death trees where for each tree 50 traits are simulated under Brownian motion model. The 2 plots in upper panel investigate the RMSD for $\sigma = 1$ and $\sigma = 5$ when the inverse of $\boldsymbol{C}$ (**sol C**) can be computed directly, whereas the other 2 plots in lower panel investigate the RMSD for $\sigma = 1$ and $\sigma = 5$ for the case where the $\boldsymbol{C}$ matrix is ill-conditioned (**unsol C**). Overall, the 4 panels suggest a relatively small shrinkage value δ may be used to achieve a better estimate for σ.

σ in general increases with δ. The model parameter σ can be estimated better with smaller δ. If the goal is to obtain a reliable estimate of σ, then a relatively small number of δ shall be considered to attain a better estimate for σ with lower RMSD value. Therefore, users may manually set up the shrinkage parameter as small as possible to obtain reliable estimate when the transformed tree has invertible $\boldsymbol{C}$ matrix.

To sum up, although the tree transformation methods proposed here do not provide significantly better improvement than the raw tree, they are still reliable options for users to choose when the $\boldsymbol{C}$ matrix of tree is ill-conditioned. Moreover, even when the $\boldsymbol{C}$ matrix is fairly well-conditioned, applying transformations does not have a substantially worse effect. One suggestion is that when the phylogenetic covariance $\boldsymbol{C}$ matrix is solvable, it may be the best to use the raw tree. Meanwhile, when $\boldsymbol{C}$ is so poorly conditioned that $\boldsymbol{C}$ cannot be inverted exactly by Cholesky decomposition, tree transformation methods provide alternative options and can give reliable estimates for most cases.

Conclusions

In this article, we explore the condition number κ of the phylogenetic covariance matrix $\boldsymbol{C}$ transformed from a phylogenetic tree. We found with fairly short terminal branch (eg, a tip with length of 10^{-15} or smaller for tree of 100 taxa) the phylogenetic $\boldsymbol{C}$ matrix fails to give the exact inverse by the Cholesky decomposition method by current software. The failure of returning exact inverse of $\boldsymbol{C}$ also depends on the number of taxa and range of condition numbers κ.

We proposed 3 methods (shrinkage matrix regularization, pruning the tips of tree, and length stretching) to alleviate the ill-conditioned matrix issue arising from the phylogenetic tree and obtain a well-conditioned estimate for $\boldsymbol{C}$ matrix. Simulations here are similar in spirit to what is performed in Figure 5 in the work by Adams and Collyer[9] where the condition number of phylogenetic covariance matrices at different level of sample size was shown. Their work was interested in the condition number relative to type I error of comparative methods, whereas our examinations use different aspects of the effect of tree condition. Another common approach is to add a small constant to the eigenvalues of the phylogenetic covariance matrix and reestimate the phylogenetic covariance matrix from the eigenvectors and adjusted eigenvalues. However, this method would alter the topology of the phylogenies, so we do not consider to implement it here.

Current R software packages that implement Felsenstein's pruning algorithm work effectively. For example, the R package `mvMORPH` can compute the square root of the phylogenetic covariance matrix and its determinant with arbitrary small tips, whereas the R package `PCMBase` implements the pruning algorithm to return the likelihood even with zero-length branches (see supplemental material pmmfelprunzerobranch.pdf). Note that the analysis of parameter estimation under the maximum likelihood estimation for the univariate BM model used in this work is performed without implementing measurement error. When assuming that each species in the tree has a measurement error variance, technically, this is also equivalent to extending each branch in the tree by a constant. Then, the variance-covariance matrix that includes measurement error and an ill-conditioned $\boldsymbol{C}$ matrix could have better matrix condition, especially in our cases where the poor matrix condition typically came from near zero-length tips. Measurement error is also included as part of the model for the phylogenetic mixed model.[50] However, measurement error can lead to bias in the inferred parameters.[51] Also, some software with hundreds of citations, such as `OUCH`[52] and `SURFACE`,[53] does not yet allow for incorporation of measurement error, rendering this potential solution impossible. As we focus on studying the condition number of $\boldsymbol{C}$ matrices on its own and explore the impact on subsequent analyses, we do not include the measurement error in this study.

There are a couple of extensions from this work. One possible work is to look at the multirate evolutionary BM model[54] where the rate of phenotypic evolution is assumed to change throughout history of life. As the rate parameters are embedded in the phylogenetic covariance matrix, the condition number of the phylogenetic covariance matrix hence depends on both the tree and rate parameters. Another possible extension from this work is to look at the condition number of the phylogenetic variance-covariance matrix for more general trait models. For instance, if one assumes OU process for trait evolution,[48] then the phylogenetic variance-covariance matrix is $\Sigma = \sigma^2\Sigma_\alpha[i,j] = \sigma^2 \exp(-\alpha t_{ij})(1-\exp(-2\alpha t_a))/(2\alpha)$ where α is the constraining force, σ is the rate of evolution, t_{ij} is the time that separating species i and j, and t_a is the time that species i and j shared a common ancestor. The statistical model for the OU process trait evolution is $\boldsymbol{Y}\sim\mathcal{N}(\theta\boldsymbol{1},\sigma^2\Sigma_\alpha)$. In particular, when $\alpha = 0$, $\Sigma_\alpha = \boldsymbol{C}$ is the BM model. Larger constraining force parameter α in OU model has the effect of lengthening the tip relative to the internal edges and may help the ill-conditioned matrix issue.

There are more complex models developed under different assumptions based on this OU process, for instance, the multiforce, multioptimum, and multirate model in the work by Beaulieu et al[55] has fairly complicated phylogenetic variance-covariance matrix $\Sigma_{\alpha,\sigma}$. Its poor performance with especially complex models could stem from issues with matrix condition, though this needs to be investigated. In addition, Bastide et al[12] and Jhwueng and O'Meara[8] developed a phylogenetic comparative method (PCM) using phylogenetic networks, rather than trees, it would be interesting to further explore the matrix condition in this case. We hope that our tree transformation methods provide the community options to ameliorate the issues produced by ill-conditioned $\boldsymbol{C}$. All analysis and simulations are done by computer Mac Pro: macOS Mojave (Late 2013), processor: 3.7 GHz Quad-Core Intel Xeon E5, RAM:12 GB 1866 MHz DDR3. The R scripts and relevant files to generate results can be accessed at https://tonyjhwueng.info/KappaPCM.

Acknowledgements

We thank Dean Adams, Michael Collyer, Cécile Ané, Elizabeth Housworth, Chiensan Peng, the editor, and 3 anonymous reviewers very much for their tremendous helps and communications for conceiving the problem and providing insightful suggestions to improve analyses on this manuscript.

Author Contributions

DCJ and BCO conceived and designed the analysis. BCO provided data. DCJ provided mathematical theory and performed simulations. DCJ and BCO wrote the manuscript. All authors reviewed the final manuscript.

ORCID iD

Dwueng-Chwuan Jhwueng https://orcid.org/0000-0001-5893-3645
Brian C O'Meara https://orcid.org/0000-0002-0337-5997

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Cornwell W, Nakagawa S. Phylogenetic comparative methods. *Curr Biol.* 2017;27:R333-R336.
2. Hinchliff CE, Smith SA, Allman JF, et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc Nat Acad Sci.* 2015;112:12764-12769.
3. Piel WH, Donoghue MJ, Sanderson MJ. *Treebase: A Database of Phylogenetic Knowledge.* Tsukuba, Japan: National Institute for Environmental Studies; 2002:41-47.
4. Sanderson MJ, Donoghue MJ, Piel WH, Eriksson T. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *Am J Botany.* 1994;81:183.
5. Stadler T. Simulating trees with a fixed number of extant species. *Systemat Biol.* 2011;60:676-684.
6. Higham NJ. *Accuracy and Stability of Numerical Algorithms.* Vol. 80. Philadelphia, PA: Siam; 2002.
7. Beckermann B. The condition number of real Vandermonde, Krylov and positive definite Hankel matrices. *Numer Mathemat.* 2000;85:553-577.
8. Jhwueng D-C, O'Meara B. Trait evolution on phylogenetic networks. *biorXiv.* https://www.biorxiv.org/content/10.1101/023986v1. Updated 2015.
9. Adams DC, Collyer ML. Multivariate phylogenetic comparative methods: evaluations, comparisons, and recommendations. *Systemat Biol.* 2018;67:14-31.
10. Felsenstein J. Phylogeny and the comparative method. *Am Natural.* 1985;125:1-15.
11. Felsenstein J. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Human Genet.* 1973;25:471-492.
12. Bastide P, Solís-Lemus C, Kriebel R, Sparks KW, Ané C. Phylogenetic comparative methods on phylogenetic networks with reticulations. *Syst Biol.* 2018;67:800-820.
13. Cybis GB, Sinsheimer JS, Bedford T, Mather AE, Lemey P, Suchard MA. Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Ann Appl Statist.* 2015;9:969-991.
14. FitzJohn RG. Diversitree: comparative phylogenetic analyses of diversification in R. *Meth Ecol Evol.* 2012;3:1084-1092.
15. Freckleton RP. Fast likelihood calculations for comparative analyses. *Meth Ecol Evol.* 2012;3:940-947.
16. Hadfield J, Nakagawa S. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *J Evolution Biol.* 2010;23:494-508.
17. Horvilleur B, Lartillot N. Monte Carlo algorithms for Brownian phylogenetic models. *Bioinformatics.* 2014;30:3020-3028.
18. Pybus OG, Suchard MA, Lemey P, et al. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc Nat Acad Sci U S A.* 2012;109:15066-15071.
19. Tung Ho LS, Ané C. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst Biol.* 2014;63:397-408.
20. Goolsby EW, Bruggeman J, Ané C. Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Meth Ecol Evol.* 2017;8:22-27.
21. Hadfield JD. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J Statist Software.* 2010;33:1-22.
22. Solís-Lemus C, Bastide P, Ané C. Phylonetworks: a package for phylogenetic networks. *Molec Biol Evol.* 2017;34:3292-3298.
23. Clavel J, Escarguel G, Merceron G. mvMORPH: an R package for fitting multivariate evolutionary models to morphometric data. *Meth Ecol Evol.* 2015;6:1311-1319.
24. Mitov V, Bartoszek K, Asimomitis G, Stadler T. Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts [published online ahead of print December 2, 2019]. *Theor Popul Biol.* 2019.
25. Manceau M, Lambert A, Morlon H. A unifying comparative phylogenetic framework including traits coevolving across interacting lineages. *Syst Biol.* 2016;66:551-568.
26. Jhwueng D-C, O'Meara BC. BMhyb: comparative methods for phylogenetic networks (R package version 2.1.5). https://rdrr.io/cran/BMhyb/. Updated 2019:.
27. Horn RA, Johnson CR, eds. *Matrix Analysis.* New York, NY Cambridge University Press; 1986.
28. O'Meara B, Heath T, Midford PE, Chamberlain S, Brown JW, Schliep K. *datelife: 0.2.3. Datelife.* https://zenodo.org/record/56803#.Xhmgkv4zbIU. Updated 2016.
29. Sanchez-Reyes L, O'Meara B. Datelife: leveraging databases to reveal the dated tree of life. [published online ahead of print October 5, 2019]. *bioRxiv.* 2019: 782094.
30. Michonneau F, Brown JW, Winter DJ. rotl: an R package to interact with the open tree of life data. *Meth Ecol Evol.* 2016;7:1476-1481.
31. Stoltzfus A, Lapp H, Matasci N, et al. Phylotastic! making tree-of-life knowledge accessible, reusable and convenient. *BMC Bioinformat.* 2013;14:158.
32. Bininda-Emonds OR, Cardillo M, Jones KE, et al. The delayed rise of present-day mammals. *Nature.* 2007;446:507.
33. Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. Tree of life reveals clock-like speciation and diversification. *Molec Biol Evol.* 2015;32:835-845.
34. Nee S, May RM, Harvey PH. The reconstructed evolutionary process. *Phil Trans Royal Soc London.* 1994;344:305-311.
35. Paradis E, Claude J, Strimmer K. Ape: analyses of phylogenetics and evolution in R language. *Bioinformatics.* 2004;20:289-290.
36. Stadler T. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *J Theoret Biol.* 2009;261:58-66.
37. Anderson E, Bai Z, Bischof C, et al. *LAPACK Users' Guide.* Philadelphia, PA: Society for Industrial and Applied Mathematics; 1999.
38. Jhwueng D-C. Assessing the goodness of fit of phylogenetic comparative methods: a meta-analysis and simulation study. *PLoS ONE.* 2013;8:e0067001.
39. Schafer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist Applicat Genet Molec Biol.* 2005;4:32.
40. Theiler J. The incredible shrinking covariance estimator. *Proc SPIE.* https://public.lanl.gov/jt/Papers/shrink-post-SPIE8391.pdf. Updated 2012.
41. Freckleton RP, Harvey PH, Pagel M. Phylogenetic analysis and comparative data: a test and review of evidence. *Am Natural.* 2002;160:712-726.
42. Pagel M. Inferring the historical patterns of biological evolution. *Nature.* 1999;401:877.
43. Sokal R, Michener C. *A Statistical Method for Evaluating Systematic Relationships.* Lawrence, KS: University of Kansas; 1958.
44. Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics.* 2010;27:592-593.
45. Bartoń K. MuMIn: multi-model inference (R package version 1.9.13). http://CRAN.R-project.org/package=MuMIn. Updated 2013.
46. Ané C. Analysis of comparative data with hierarchical autocorrelation. *Evolution.* 2008;2:1078-1102.
47. Ho LST, Ané C. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst Biol.* 2014;63:397-408.
48. Hansen TF. Stabilizing selection and the comparative analysis of adaptation. *Evolution.* 1997;51:1341-1351.
49. Jhwueng D-C. *Some Problems in Phylogenetic Comparative Method* [PhD thesis]. Bloomington, IN: Indiana University Bloomington; 2010.
50. Housworth EA, Martins EP, Lynch M. The phylogenetic mixed model. *Am Natural.* 2004;163:84-96.
51. Cooper N, Thomas GH, Venditti C, Meade A, Freckleton RP. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biol J Linnean Soc.* 2016;118:64-77.
52. Butler M, King A. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am Natural.* 2004;164:683-695.
53. Ingram T, Mahler DL. Surface: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike information criterion. *Meth Ecol Evol.* 2013;4:416-425.
54. O'Meara B, Ané C, Sanderson M, Wainwright P. Testing different rates of continuous trait evolution using likelihood. *Evolution.* 2006;60:922-933.
55. Beaulieu J, Jhwueng D-C, Boettiger C, O'Meara B. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution.* 2012;66:2369-2383.