On the matrix condition of phylogenetic tree

Dwueng-Chwuan Jhwueng[1]

[1]**Department of Statistics, Feng-Chia University, No. 100 Wenhua Rd., Seatwen Taichung Taiwan**

Corresponding author:
Dwueng-Chwuan Jhwueng[1]

Email address: dcjhwueng@fcu.edu.tw

ABSTRACT

Phylogenetic comparative analyses incorporate phylogenetic tree to study evolutionary relationship among a group of related species. A phylogenetic tree of n taxa can be algebraically transformed into an n by n squared phylogenetic similarity matrix $\boldsymbol{C}$ where each element g_{ij} in $\boldsymbol{C}$ represents the affinity between extant species i and extant species j. Because $\boldsymbol{C}$ plays an important role in phylogenetic comparative analysis, one should take rigorous investigation on the condition of $\boldsymbol{C}$. The condition number of matrix $\boldsymbol{C}$ denoted by κ is defined by the ratio of the maximum eigenvalue of $\boldsymbol{C}$ to the minimum eigenvalue of $\boldsymbol{C}$. While $\boldsymbol{C}$ is an ill-conditioned matrix with fairly large value of κ, results obtained from subsequent analyses such as computing the likelihood that requires inversion of $\boldsymbol{C}$ may not be stable. To remediate this problem, we propose several methods to appropriately adjust the phylogenetic tree and improve the matrix condition of $\boldsymbol{C}$ for the sake of attaining reliable results.

INTRODUCTION

Starting by stating very briefly the central role of comparative studies and phylogenetic trees in evolutionary biology.

The central role for phylogenetic comparative studies played in evolutionary biology is to test evolutionary hypotheses and to provide evidences of organismal evolution and diversification (**?**). The central role for phylogenetic trees in evolutionary biology serves as the scientific evidence of evolutionary relationships among various biological species.

Consider a numerical problem arises from a given phylogenetic tree when a biologist analyzes trait data using phylogenetic comparative analyses

This problem can be explored from the simulation and from studying the real trees to obtain a measure for summarizing some aspects of phylogenies. For instance, one can explore via investigating the statistical properties of real phylogenies (i.e. trees estimated from molecular data under substitution models) or simulated trees generated in various ways. To our knowledge, prior comparisons have focused on the shape such as tree balance (Mir et al., 2013) of simulated trees and whether they seemed reasonable when compared to the shapes of trees estimated from data (Colijn and Plazzotta, 2018).

There are many comparison of statistical properties of the trees themselves in the literature. **?** developed metrics on unlabeled tree shapes, and used them to compare simulated and data-derived trees. The shapes of evolutionary trees are influenced by the nature of the evolutionary process but comparison of trees from different processes are hindered by the challenge of completely describing tree shape. We present a full characterization of the shapes of rooted branching trees in a form that lends itself to natural tree comparisons. We use this characterization to define a metric, in the sense of a true distance function, on tree shapes. The metric distinguishes trees from random models known to produce different tree shapes. It separates trees derived from tropical versus USA influenza A sequences, which reflect the differing epidemiology of tropical and seasonal flue. We describe several metrics based on the same core characterization, and illustrate how to extend the metric to incorporate trees branch lengths or other

features such as overall imbalance. Our approach allows us to construct addition and multiplication on trees, and to create a convex metric on tree shapes which formally allows computation of average tree shapes.

? characterizing and comparing phylogenies from their Laplacian spectrum: phylogenetic trees are central to many areas of biology, ranging from population genetics and epidemiology to microbiology, ecology, and macroevolution. The ability to summarize properties of trees, compare different trees, and identify distinct modes of division within trees is essential to all these research areas. But despite wide-ranging applications, there currently exists no common, comprehensive framework for such analyses. Here we present a graph-theoretical approach that provides such framework. We show how to construct the spectral density profile of a phylogenetic tree from its Laplacian gragh. Using ultrametric simulated trees as well as non-ultrametric empirical trees, we demonstrate that the spectral density successfully identifies various properties of the trees and clusters them into meaningful groups. Finally, we illustrate how the eigengap can identify modes of division within a given tree. As phylogenetic data continue to accumulate and to be integrated into various areas of the life sciences, we expect that this spectral graph-theoretical framework to phylogenetics will have powerful and long-lasting applications.

? described a generalization of the K statistics (**?**)blomberg2003) that is useful for quantifying and evaluating phylogenetic signal in highly dimensional multivariate data.

As a consequence, it would be meaningful to focus more on algebraic and statistical sides where one can look at the condition number of the phylogenetic similarity matrix $\boldsymbol{C}$ (Jhwueng, 2013) for these trees collected in literature, as well as for trees simulated in different ways (Stadler, 2011). Since by the hierarchical property of phylogenetic tree, $\boldsymbol{C}$ is a positive definite matrix, the condition number κ of a phylogenetic similarity matrix $\boldsymbol{C}$ can be defined by 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 an eigenvalue of $\boldsymbol{C}$ that satisfies $\det(\boldsymbol{C} - \lambda_i \boldsymbol{I}) = 0, i = 1, 2, \cdots, n$.

Basically the condition number κ is a measure of how stable the matrix is for subsequence operations (Higham, 2002). It could be good to add some background on the use of this number, and on the range of acceptable valuess for it. Small condition numbers are more stable matrices, while larger condition numbers are less stable. this number describe the numerical properties of the matrix, and its robustness to standard computer operation More stable matrices are generally good, as with them there is 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 in statistics area. By contrast, large condition numbers mean these operations are unstable and more prone to error propagation.

Given a rooted phylogenetic tree of n extant taxa, each element $g_{ij}, i, j = 1, 2, \cdots, n$ in $\boldsymbol{C}$ is measured by the shared branch length between a pair of species on the tips of tree. For instance, a phylogenetic tree of 6 taxa shown in Fig. 1 can be represented as a phylogenetic similarity matrix $\boldsymbol{C}$ in Eq. (2). The maximum and minimum eigenvalues of $\boldsymbol{C}$ are 192.32 and 7.68, respectively. The condition number κ for the tree in Fig. 1 is calculated as $\kappa = 192.32/7.68 = 25.04$.

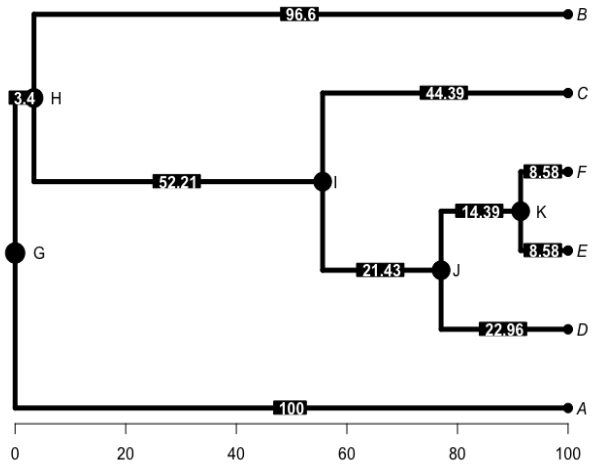

Figure 1. A phylogenetic tree of 6 taxa and its corresponding phylogenetic similarity matrix $\boldsymbol{C}$

$$\boldsymbol{C} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 100.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 100.00 & 3.40 & 3.40 & 3.40 & 3.40 \\ 0.00 & 3.40 & 100.00 & 55.61 & 55.61 & 55.61 \\ 0.00 & 3.40 & 55.61 & 100.00 & 77.04 & 77.04 \\ 0.00 & 3.40 & 55.61 & 77.04 & 100.00 & 91.42 \\ 0.00 & 3.40 & 55.61 & 77.04 & 91.42 & 100.00 \end{pmatrix} \end{array} \tag{2}$$

For n species, let $y_1, y_2, \cdots, y_n$ be the phenotypic values observed on the tips of tree. By the evolutionary dynamics under the Gaussian process such as the Brownian motion (Felsenstein, 1985), the joint distribution for $n \times 1$ random vector $\boldsymbol{Y} = (y_1, y_2, \cdots, y_n)^t$ of n species is a multivariate normal distribution with common mean $\mathbb{E}(\boldsymbol{Y}) = \theta \mathbf{1} = (\theta, \theta, \cdots, \theta)^t$, and $n \times n$ phylogenetic covariance matrix $\sigma^2 \boldsymbol{C}$. The statistical model is displayed in Eq. (3)

$$\boldsymbol{Y} \sim \mathcal{N}\left(\theta \mathbf{1}, \sigma^2 \boldsymbol{C}\right). \tag{3}$$

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

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

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 $\mathbf{1} = (1, 1, \cdots, 1)^t$ is a vector of 1s. In fact, the best solution for the likelihood function in trait model in Eq. (4) depends on the phylogenetic similarity matrix $\boldsymbol{C}$ itself. To calculate the inverse of the phylogenetic similarity matrix $\boldsymbol{C}$, one can use a Moore-Penrose (MP) pseudoinverse that makes the algebra tractable(Horn and Johnson, 1986). However, there are other methods that may turn out to be better for this situation. Our goal is in two folds: first we 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 similarity matrix, when the observed matrix is *ill-conditioned*.

Data Collection

We collect the real tree data in the Tree of Life(ToL) database (Maddison and Schulz, 2007), those trees are converted into chronograms using the method in (Boettiger and Temple Lang, 2012) for interfacing with TreeBASE (Sanderson et al., 1994; Piel et al., 2002). However, trees in TreeBASE do not have branch

lengths so that tree measures could not be evaluated while OpenTree database (Hinchliff et al., 2015) still does not have as many as trees as TreeBASE, but at least they are labeled as chronograms.We access those trees using the `R` package: `rotl` (Michonneau et al., 2016) to get its cache of 130 chronograms that includes sometimes multiple from one study from OpenTree by installing the DateLife (O'Meara et al., 2016) and phylotastic (Stoltzfus et al., 2013) to pull them in(file: opentree_chronograms.rda). All the trees are labeled with the citation for the study. The tree size ranges from 6 taxa to 48016 taxa.

For tree of 48016 taxa (Hedges et al., 2015), our computer reaches its limit to compute the variance covariance matrix. For the three trees of 4510 taxa (Bininda-Emonds et al., 2007), we observed that the size falls out the taxa size set. We ignore these four trees to get 126 trees with 815 or fewer taxa in each for our work. The median size is of 72 taxa and the mean size is of 140 taxa.

Preliminary analysis

As phylogenetic comparative methods must obtain the inverse of $\boldsymbol{C}$ for use in the calculations.

There is an extensive literature of methods precisely trying to avoid the actual computation of this inverse, in order to gain speed and numerical stability. Starting with Felsenstein's pruning algorithm (Felsenstein, 1973) (Felsenstein, 1973), there are many extensions in many contexts, see e.g. (Hadfield and Nakagawa, 2010) Hadfiled & Nakagawa (2010), (FitzJohn, 2012) Fitzjohn (2012), (Freckleton, 2012) Freckleton (2012), (Tung Ho and Ané, 2014) Ho and Ane (2013), (Pybus et al., 2012) Pybus et al. (2012), (Horvilleur and Lartillot, 2014) Lartillot (2014), (Cybis et al., 2015) Cybus et al. (2015), (Bastide et al., 2018) Bastitide et al. (2018). These approaches have been implemented in several popular R packages such as Diversitree (FitzJohn, 2012), phylolm (Tung Ho and Ané, 2014), Rphylopars(Goolsby et al., 2017), MCMCglmm (Hadfield, 2010). This is not to say that studying the condition number of these matrix is not an interesting and valuable question on its own, but just to point out that, using these kind of efficient algorithms, the community have come up with solutions that are more robust and efficient than just a matrix inversion.

We show the condition numbers κ vs. number of taxa for all 126 chronograms in the database in Figure 2 where κ falls nicely in a tight increasing pattern with number of taxa. This could be what real trees look like when viewed by the *stability* of their phylogenetic similarity matrix. Stability decreases as number of taxa increases, but does so somewhat slowly.

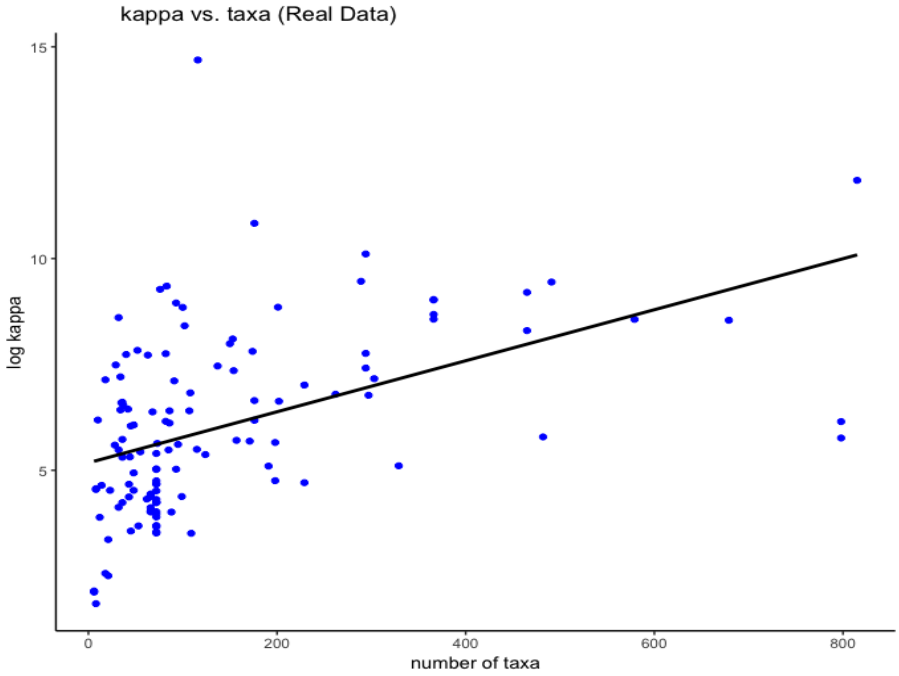

Figure 2. The log conditioned number vs. number of taxa for trees from literatures. A simple linear regression yields the line equation $\log \kappa = 5.18 + 0.006 \times \text{taxa}$. For one unit increase in taxa, we expect to see 0.6% ($e^{0.006025} \approx 1.006$) increase in kappa value.

We next show conditioned number κ vs. number of taxa using trees from simulation. In Figure

3, there are three lines where each line represents the mean of 100 runs of simulated phylogenies at different number of taxa. The green line was obtained from random trees using coalescent trees method (implemented using `R` package: `rcoal` (Paradis et al., 2004) along with computed BM (Felsenstein, 1985) to make a chronogram), and the red line was obtained from trees simulated by birth-death process with a given age on a fixed number of extant taxa (Stadler, 2009) using birth rate $\lambda = 0.4$, death rate $\mu = 0.1$, and tree of age $\log(n)/(\lambda - \mu) = \log(100)/(0.4 - 0.1) = 15.35$. The blue line is obtained from trees under a pure-birth process (birth $= 1$, death $= 0$).

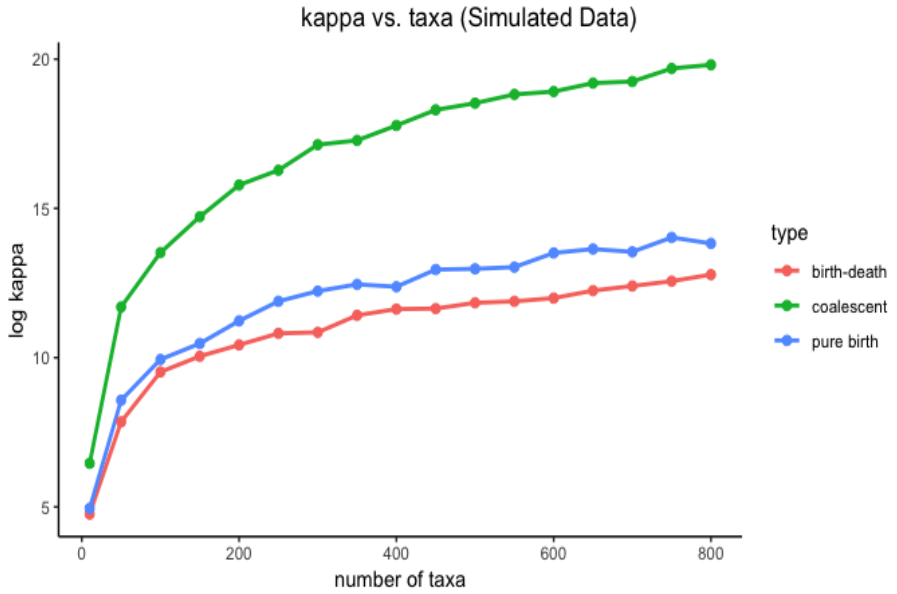

Figure 3. The condition number κ vs. number of taxa for trees from simulation. redline: birth-death model, greenline: coalescent model, blueline: pure birth model.

THE GOAL OF THIS PAPER COULD BE PHYLOGENY OF n TAXA WITH CONDITION NUMBER MORE THAN 10^x ARE SAID TO BE ILL-CONDITIONED.

WE NEED TO RETWRITE THE FOLLOWING PARAGRAPH AS REVIEWERS FOUND AN ERROR IN OUR CODE.

Notice that random splits trees(blue line) do something similar to the line shown in Fig. 3 for pure birth trees(red line). As shown in Figure 3, we see both random split trees and pure birth trees behave unstably. In other words, compare to Figure 2, we observed that real chronograms are more statistically *stable* the reviewer is not sure that the term statistically stable is relevant when using this particular metric. But that might be just a matter of definitio, that could be clarified with a better introduction of the properties of the condition number. than trees generated from random split trees or pure-birth simulators (blue line and red line in Figure 3). This gives a surprising result, and has quite a few consequences; particular for researchers using pure birth or random split tree as inputs for simulations that then evaluate comparative patterns or even comparative methods themselves. In short, if the trees are unstable statistically, then either the data generated from these simulations could be unstable, the subsequent phylogenetic comparative analyses which use the inverse of $\boldsymbol{C}$ could be unstable. And these consequences would be expected to be more acute for multivariate data, though it remains under investigation.

Above preliminary analysis is a motivational interest of ours in the repository of chronograms. We would like to see how the chronograms behave relative to the simulated trees, in terms of statistical stability. From above, it is very likely that matrix condition affects a broader swatch of PCMs than anyone has considered. In fact, this issue has not been well recognized within the community, and of the fact that the statistical properties of the covariance matrix from a phylogeny or network could lead to downstream issues with phylogenetic comparative methods. Observing this, in this paper we will work on the phylogenetic similarity matrix $\boldsymbol{C}$ for the BM model (Felsenstein, 1985) in Eq. (3).

METHODS

To deal with the ill-conditioned matrix issue rising from the given phylogenies, one direct approach is just to reject matrices, $\boldsymbol{C}$s, that are poorly conditioned. But the truth estimate can fall in that region, and users would be less favorable to hear that their trees can not be analyzed though that is far better than quietly returning a wrong result. We propose several possible approaches to remediate the issue of *ill-condition* 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.We describe three approaches

(i) shrinkage matrix regularization: lengthen or shorten the tip lengths.

(ii) pruning tips of the tree: removing tips from the tree.

(iii) length stretching:

Shrinkage matrix regularization

An approach by Schafer and Strimmer (2005) was developed for regularizing covariance matrices in molecular biology (including some network covariance matrices) and Theiler (2012) improved upon them accordingly for general purposes. Given the tree with the matrix $\boldsymbol{C}$ READ Theiler (2012) WORK, NEED REWRITE MODEL, AND MAYBE INCLUDE TRAIT IN THE CODE. Let $0 \leq \delta \leq 1$ 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 $r = \text{trace}(\boldsymbol{C}^{-1}\boldsymbol{S}_\delta)$, the negative log likelihood function for the shrinkage estimated covariance matrix $\boldsymbol{S}_\delta$ as a function of the shrinkage parameter δ is

$$-\log L(\delta) = \log(1-r\beta) + \frac{r}{1-r\beta} + \log \mid \boldsymbol{S}_\delta \mid . \tag{5}$$

The best shrinkage estimate is to search the optima $\hat{\delta} = \max_{0\leq\delta\leq 1} \log L(\delta)$ 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 analysis. The shrinkage method we used here is similar to the very broadly ysed Pagel's lambda transformation (Pagel, 1999) $\boldsymbol{S}_\delta = \lambda C + (1-\lambda)T$, where the biological meaningful heritbility parameter λ differs in β up to an $n/(n-1)$ factor.

We implemented the shrinkage method to compare the condition numbers for the raw trees and shrunk trees. We simulated 100 birth-death trees using `R` package: `TreeSim` (Stadler, 2009) where each tree is of $n = 100$ of extant sampled tips, speciation rate $\lambda = 0.4$, extinction rate $\mu = 0.1$, each tip is included into the final tree with probability 0.5, and the time since origin of the process is $(\log n)/(\lambda - \mu)$. The condition numbers of the phylogenetic similarity matrix $\boldsymbol{C}$ for the 100 raw trees and 100 shrunk trees are calculated. The results is plotted in Figure 4 which shows that the shrunk trees have lower condition numbers than the raw trees. This indicates an improvement for the ill-condition issue.

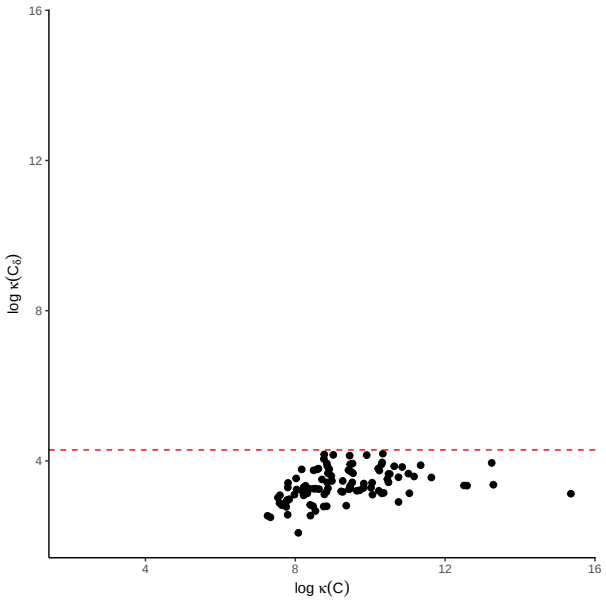

Figure 4. Scatter plot of the condition number of the raw trees $\log \kappa(\boldsymbol{C})$ vs. the condition number for the shrunk trees $\log \kappa(\boldsymbol{C}_\delta)$.

We next visualize the phylogenies that result from the estimated covariance matrices to see how they differ from the observed trees. Then we compare aspects of phylogenies that have well-behaved covariance matrices to those that are ill-behaved. We compare the topology of the raw tree and the shrunk tree reconstructed using the shrinkage matrix under the unweighted pair group method of arithmetic mean `R` package: `(upgma)` (Sokal et al., 1958) and show the result in Figure 5. Those shrunk tree seems to stretch a lot on the tip lengths.

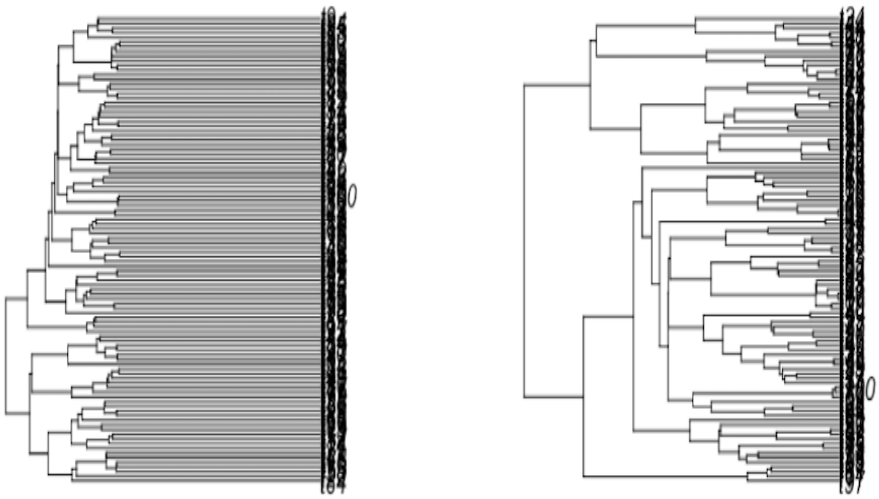

Figure 5. Tree comparison for the shrunk tree (left panel) and the raw tree (right panel).

We also found a trend that the sample sizes have impact on estimate the the value of δ which has direct impact on stretching the tips lengths. We use three different number of taxa $N = 20, 100, 150$ and find that δ will increase as sample size increases. For one hundred replicates, the average of the shrinkage estimator $\hat{\delta}$ are 0.02, 0.41, and 0.63 under the $N = 20, 100$ and 150, respectively.

Tree pruning

We further investigate whether any possible tree parameter has impact on the condition of the tree. We consider the tree parameters that could affect the condition of the tree as following: (1) Ntip (2) age (3) median of branch lengths (4) ratio between max branch length and min branch length (5) variance of branch length (6) sampling fraction (7) birth rate λ (8) max branch length (9) max internal branch length (10) max tip length (11) min branch length (12) min internal branch length (13) min tip length (14) death rate μ.

We conduct an analysis by simulating $100,000$ trees under uniformly varying number of taxa between size of 10 and 100, birth rate λ between 0.01 and 0.1, death rate μ between 0 and λ, sampling fraction between 0.1 and 1, and tree of height of value $(\log n/(\lambda-\mu))$. Their condition numbers κ are calculated and compared to a variety of measures using multiple linear regression. We apply `R` package: `MuMIn` (Bartoń, 2013) to dredge plenty models to correlate $\log\kappa$ with various parameters or tree measures. The number of variables is set to 4 and essentially all the weight was on one of 1471 models (intercept model: 1 model, 1 predictor: 14 models, 2 predictors:91 models, 3 predcitors:364 models, 4 predictors:1001 models). The model, $\log\kappa \sim n+\lambda+\mu+\iota$ where n is the number of taxa and ι is the min tip, accounts for most weight of significance (Akaikei weight $w \approx 1$).

Table 1 shows the coefficient and standard error for the predictors.

Table 1. The best model out of the 1471 model by model dredging.

	(Intercept)	n	λ	μ	ι
Estimate	6.31	0.03	-24.02	34.48	-0.18
Std. Error	0.02	0.00	0.23	0.26	0.00

Note that there are models with positive coefficient of *lambda* but HOW ABOUT THE min tip ? LOOK AT THE dredge.results100000.csv

From previous analysis, it makes sense that higher n would contribute to worse (higher) κ, as does bigger smallest tip branch ι leading to better (lower) κ. We run all pairwise dot-plots between condition numbers and the other tree parameters. From Table 1 , we found two comparisons: (1) condition number vs. min tip (i.e. κ vs. ι) and (2) condition number vs. birth rate (i.e. κ vs. λ) are intersting.

We provide an empirical evidence that taxon removal tends to help, but removing the taxon with the shortest branch helps more in Figure 6. 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.

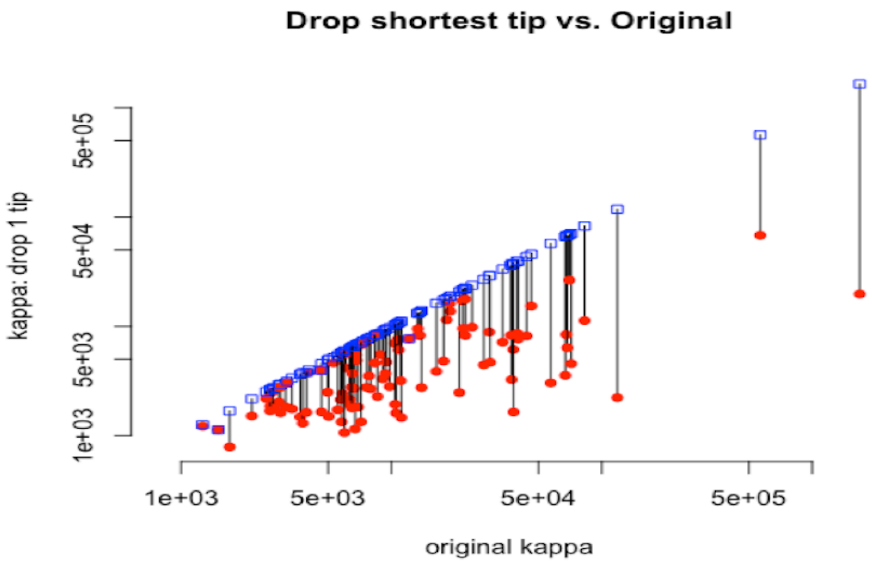

Figure 6. Condition numbers for dropping the shortest tip vs. the condition numbers for the raw tree. The horizontal-axis is the original κ, and the lines connect points corresponding to that given tree, but with one taxon removed: the blue square is removing a taxon at random, and the red dot removing the taxon with the shortest branch tip length.

We found that, in fact, the problem of having ill-conditioned matrix is highly related to branch lengths of taxon. We illustrate this issue using a simple example of two taxa shown in the Figure 7 (right panel). The phylogenetic matrix is

$$\boldsymbol{C} = \begin{matrix} \\ X \\ Y \end{matrix} \begin{matrix} \begin{matrix} X & Y \end{matrix} \\ \begin{pmatrix} 1+\varepsilon & 1 \\ 1 & 1+\varepsilon \end{pmatrix} \end{matrix} \tag{6}$$

Note that $\boldsymbol{C}$ has two eigenvalues ε and $2+\varepsilon$. The condition number of $\boldsymbol{C}$ defined by the ratio of the largest eigenvalues to the smallest eigenvalues is $\kappa = (2+\varepsilon)/\varepsilon = 1+2/\varepsilon = \mathcal{O}(\varepsilon^{-1})$ where $\mathcal{O}(\cdot)$ is the big O notation that describes the limit behavior of a function. When we have tiny tip branch (very small ε), the value of κ will be fairly large and the matrix is more of ill-conditioned. For instance, with $\varepsilon = 0.1, \kappa = 21$ while with $\varepsilon = 0.001$ gives $\kappa = 2001$. The problem becomes serious as ε is very close to zero as we have a matrix of two almost identical columns/rows which makes $\boldsymbol{C}$ a singular matrix with $\kappa = \infty$. In general, for a tree of arbitrary taxa that includes a clade of sub tree described in this case, we will face the ill-condition problem. To explore this issue more thoroughly, we define a measure using the fraction $\omega = \varepsilon/(t+\varepsilon)$ to quantify the ill-condition of the tree where $t+\varepsilon$ is the tree height, ε is the shortest tip length and t is the branch lengths from the root to the most recent common ancestor of the tips with smallest tip lengths. When ε approaches to zero, the fraction ω would approach to zero as well. In this case, as two columns/rows in the matrix are almost identical, the matrix will suffer a serious ill-condition issue of order $\mathcal{O}(\varepsilon^{-1})$. We use simulated birth-death trees of 100 taxa. For a good tree ($\kappa < 500$), $\omega = 1.45 \times 10^{-2}$, while for a bad tree ($\kappa > 1.5 \times 10^{6}$) , $\omega = 3.96 \times 10^{-6}$. Hence the tree condition is of the order $\mathcal{O}(\omega^{-1})$. The reviewer is not sure she follows the point of the above empirical analysis, especially in the light of what comes next, with the precise description of Lemma 1. Could this omega score be used as a faster to compute proxy for the kappa value?

Moreover, phylogenetic similarity matrix $\boldsymbol{C}$ have a nested structures that general covariance matrices might not.We identify that the shortest tip is actually the smallest eigenvalue of the variance covariance matrix. We start by illustrating a 3 taxa example.

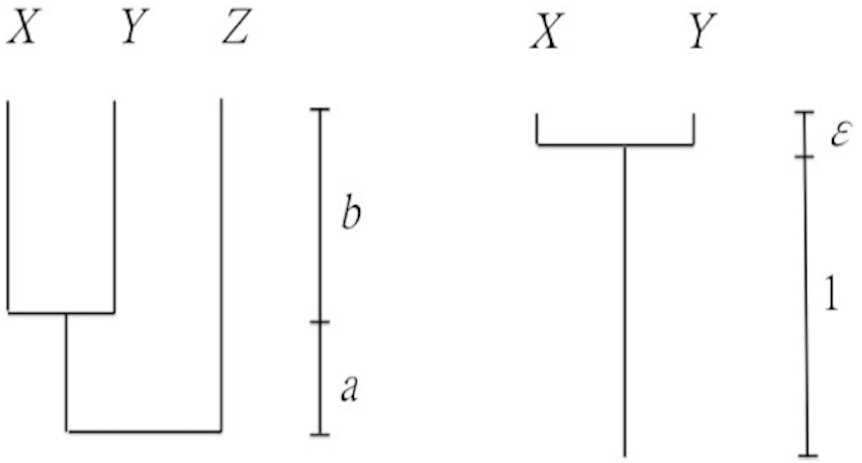

Figure 7. Example of tree of three taxa and two taxa.

In Figure 7 (left panel), the tip lengths for species X, Y and Z are b, b and $a+b$, respectively. Apparently, the branch length of the shortest tip is b. The phylogenetic matrix for the tree in Fig. 7 can be represented as following

$$\boldsymbol{C} = \begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{pmatrix} a+b & a & 0 \\ a & a+b & 0 \\ 0 & 0 & a+b \end{pmatrix} \end{array}. \tag{7}$$

Let λ be an eigenvalue for $\boldsymbol{C}$. We have $\det(\boldsymbol{C}-\lambda\boldsymbol{I})=0$ where $\boldsymbol{I}$ is a 3 by 3 identity matrix. Solving and simplifying this equation leads to $(a+b-\lambda)((a+b-\lambda)^2-a^2)=0$.Then we have 3 eigenvalues $\lambda = 2a+b, a+b, b$. As a, b are both nonnegative numbers, the smallest eigenvalue for $\boldsymbol{C}$ is b which is the shortest tip length on the tree in Fig. 7. For a general case, the proof of this property under tree of arbitrary taxa is provided in Lemma 1.

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

Proof: Given an ultrametric bifurcated tree $\mathbb{T}$ of n tips, there exists a unique strictly ultrametric matrix (Nabben and Varga, 1994) $\boldsymbol{C}$ for representing the relatedness among the group of species. If b is the smallest tip length, by the property of the structure of the ultrametric tree, $\boldsymbol{C}-b\boldsymbol{I}$ has at two identical columns as well as two identical rows. This implies that $\det(\boldsymbol{C}-b\boldsymbol{I})=0$. Therefore, b is an eigenvalue of $\boldsymbol{C}$.■

The next step is to show that b is the smallest eigenvalue in the eigenvalue set of $\boldsymbol{C}$. We claim that for all $\lambda_0 < b$, then λ_0 is not an eigenvalue. Consider the matrix $\boldsymbol{C}_0 = \boldsymbol{C}-\lambda_0\boldsymbol{I}$, then $\boldsymbol{C}_0$ is still a strictly ultrametric matrix which is always invertible (see Nabben and Varga (1994), and Corollary 6.2.27 in (Horn and Johnson, 1986)). Then we have $\det(\boldsymbol{C}_0)\neq 0$ which implies $\det(\boldsymbol{C}-\lambda_0\boldsymbol{I})\neq 0$. This consequence indicates that λ_0 is not an eigenvalue of $\boldsymbol{C}$. Therefore, b must be the smallest eigenvalue of $\boldsymbol{C}$.

The general proof in Lemma 1 shores up the pruning approach from a theoretical perspective, even if some researcher may object to losing a taxon in their analysis and may lead to a biased estimate. We next show 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 $\boldsymbol{C}_1$ be the matrix obtained by dropping the shortest tip from the tree and κ_1 be the condition number of $\boldsymbol{C}_1$. Then $\kappa \geq \kappa_1$.*

Proof: Let $0<\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_n$ be the eigenvalues of $\boldsymbol{C}$. Since $\boldsymbol{C}_1$ is still a strictly ultrametic matrix of size $(n-1)$ by $(n-1)$, we can assume that $\boldsymbol{C}_1$ has eigenvalues $0<\tau_1\leq\tau_2\leq\cdots\leq\tau_{n-1}$. By a special case of the Cauchy's interlacing theorem (Ch. 10.1 in (Parlett, 1987)), we have $0<\lambda_1\leq\tau_1\leq$

$\lambda_2 \le \tau_2 \le \cdots \le \lambda_{n-1} \le \tau_{n-1} \le \lambda_n$. The condition number, defined as the ratio of the largest eigenvalue to the smallest eigenvalue are computed as $\kappa = \lambda_n/\lambda_1$ and $\kappa_1 = \tau_{n-1}/\tau_1$ for $\boldsymbol{C}$ and $\boldsymbol{C}_1$, respectively. From direct algebraic calculation, we have $\kappa - \kappa_1 = \frac{\lambda_n}{\lambda_1} - \frac{\tau_{n-1}}{\tau_1} = \frac{\tau_1\lambda_n - \tau_{n-1}\lambda_1}{\lambda_1\tau_1} \ge \frac{\tau_1\tau_{n-1} - \tau_{n-1}\lambda_1}{\lambda_1\tau_1} = \frac{\tau_{n-1}(\tau_1 - \lambda_1)}{\lambda_1\tau_1} \ge 0$. Hence $\kappa \ge \kappa_1$ which concludes that a lower κ can always be achieved by dropping the shortest tip. ■

Remark: Above two lemmas have a link with the result obtained in Ane (2008) where the whole spectrum of the matrix $\boldsymbol{C}$ is derived for the special case of a symmetric treee and has been extended in Ho and Ané (2014) for an OU model. It would be interesting to explore whole spectrum for arbitrary ultrametic tree for genrealization.

Length Stretching

We also consider to adopt the method in Jhwueng (2010, 2013) which stretch the branch lengths of the raw tree without changing its topology. We call it the Gamma method and illustrating the procedure as following: For an ultrametric tree, let $\tau = 1$ be the scaled tree height from the root to the tip. Without loss of generality, we can first scale τ into a unit and then decompose τ into d components. That is, $1 = \tau = \tau_1 + \tau_2 + \cdots + \tau_d$ where $\tau_i > 0, i = 1,2,\cdots,d$ represents the length between the i^{th} 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 length for the tip species (with minimum tip length) evolved from its most recent common ancestor. Next, consider the matrix $\boldsymbol{C}$ obtained from the raw tree. Let the $(d+1)$-tuple elements $(c_1, c_2, \cdots, c_d, c_{d+1})$ be the distinct entries in $\boldsymbol{C}$ satisfying $1 = c_{d+1} > c_d > c_{d-1} > \cdots > c_1 = 0$. We can represent the relation between τ_i and c_i by the equation $\tau_i = c_{i+1} - c_i, i = 1,\cdots,d$. The following lemma helps us to well confirm the relationship between the $c_{i=1}^{d+1}$ and $\boldsymbol{C}$.

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

Proof: Without lose of generality, suppose $\{c_{d+1} > c_d > \cdots > c_1\}$ are $d+1$ distinct elements in C where $c_{d+1} = 1$ and $c_1 = 0$. By definition of the C, $d_i, i = 1,2,\cdots,d+1$ is equivalent to the node height from the root. By taking the difference between two successive node heights and define $\tau_i = c_{i+1} - c_i, i = 1,\cdots,d$, then τ_i represent the branch segment between the ith and the $i+1$th speciation and $\{\tau_i\}_{i=1}^d$ has exactly d elements.

On the othre hand, let $\tau_1, \tau_2, \cdots, \tau_d$ be the branch segments defined by taking the difference between the two successive node heights. Let $c_1 = 0$, and define $c_{i+1} = \tau_i + c_i, i = 1,2,\cdots,d$. (i.e. $c_2 = \tau_1 + c_1 = \tau_1, c_3 = \tau_2 + c_2 = \tau_2 + \tau_1$, $c_4 = \tau_3 + c_3 = \tau_3 + \tau_2 + \tau_1, \cdots, c_{d+1} = \tau_d + c_d = \tau_d + \tau_{d-1} + \cdots + \tau_1 = 1$.) Then each c_i represents the node height of the ith speciation events (could be more than one if speciations occur at the same times) and c_i is stacked up by accumulating the branch segment between two successive speciation events. Note that c_i measures the affinity between a pair of species and the definition of τ_i make $\{c_i\}_{i=1}^{d+1}$ eligible to represent the affinity between arbitrary pair of species at tip. So $\{c_i\}_{i=1}^{d+1}$ filled out the C matrix. This explans that there are exactly $d+1$ distinct element in C. ■

We use Figure 7 (left panel) as an example. Let $\tau_1 = a = 0.6$ and $\tau_2 = b = 0.4$, then the $\boldsymbol{C}$ matrix is

$$\boldsymbol{C} = \begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{pmatrix} a+b & a & 0 \\ a & a+b & 0 \\ 0 & 0 & a+b \end{pmatrix} \end{array} = \begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{pmatrix} 1 & 0.6 & 0 \\ 0.6 & 1 & 0.6 \\ 0 & 0.6 & 1 \end{pmatrix} \end{array}.$$

Observing that $c_3 = 1, c_2 = 0.6, c_1 = 0$, we have $c_2 - c_1 = 0.6 - 0 = 0.6 = \tau_1$, and $c_3 - c_2 = 1 - 0.6 = 0.4 = \tau_2$. To stretch the lengths but retain the topology of the raw tree, since $\tau_i > 0$ and $\Sigma_{i=1}^d \tau_i = 1$, we can treat $(\tau_1, \tau_2, \cdots, \tau_d)$ as a d-dimensional random vector from a Dirichlet distribution. Then $(\tau_1, \tau_2, \cdots, \tau_d)$ can be generated by first drawing d independent gamma random variables, $T_i, i = 1,2,\cdots,d$, each with different shape parameters $m\tau_i, i = 1,2,\cdots,d$ and rate parameter 1 where m is an arbitrary but positive constant, then the d-tuple vector $(t_1^*, t_2^*, \cdots, t_d^*) = \frac{1}{\Sigma_{i=1}^d T_i}(T_1, T_2, \cdots, T_d)$ is a Dirichlet random vector with $t_i^* \in (0,1)$, $\Sigma_{i=1}^d t_i^* = 1$, and concentration parameters $m\tau_i, i = 1,2,\cdots,d$. 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 given by $M_{t_i^*} = (m\tau_i - 1)/[(\Sigma_{i=1}^d m\tau_i) - d] = (m\tau_i - 1)/(m-d)$ where $m\tau_i > 1, i = 1,2,\cdots,d$. The choice of m is thus determined by $\min_{1\le i\le d} m\tau_i > 1$. We choose a positive integer m satisfies $m = \lceil \frac{1}{\min_{1\le i\le d}\tau_i} \rceil$

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 (Jhwueng, 2010).

Another efficient algorithm (Gelman et al., 2003) relies on the univariate marginal and conditional distributions being Beta distribution. We call it the Beta method and illustrating the procedure as following: : simulate t_1^* from a Beta$(\tau_1, \Sigma_{i=2}^{d}\tau_i)$ distribution. Then simulate $t_2^*, t_3^*, \cdots, t_{d-1}^*$ in order, as follows. For $j = 2, \cdots, d-1$, simulate s_j from a Beta$(\tau_j, \Sigma_{i=j+1}^{d}\tau_i)$ distribution, and let $t_j^* = (1 - \Sigma_{i=1}^{j-1}\tau_i)s_j$. Finally, set $t_d^* = 1 - \Sigma_{i=1}^{d-1}\tau_i$. Figure 9 shows a comparison of the three trees of 10 taxa.

To test whether both Gamma and Beta methods could reduce the condition numbers, we first simulate a tree, and then calculate its condition number κ. We then stretch the lengths of the trees and calculate the corresponding condition number κ. We perturb the raw tree 100 times to get 100 new trees where their condition numbers are plotted in the Figure 9. Figure 9 shows that the length stretching method could lead to a better or worse condition number.

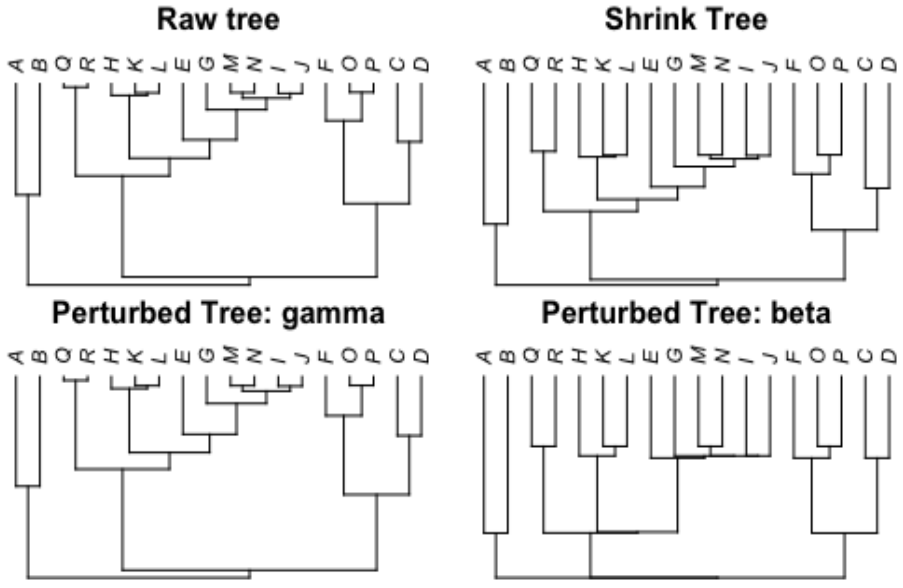

Figure 8. Tree length stretching. The upper left tree is the raw tree, the upper right is the shrink tree, the lower left tree is the stretched tree stretch under the gamma method and the lower right tree is the stretched tree under the beta method.

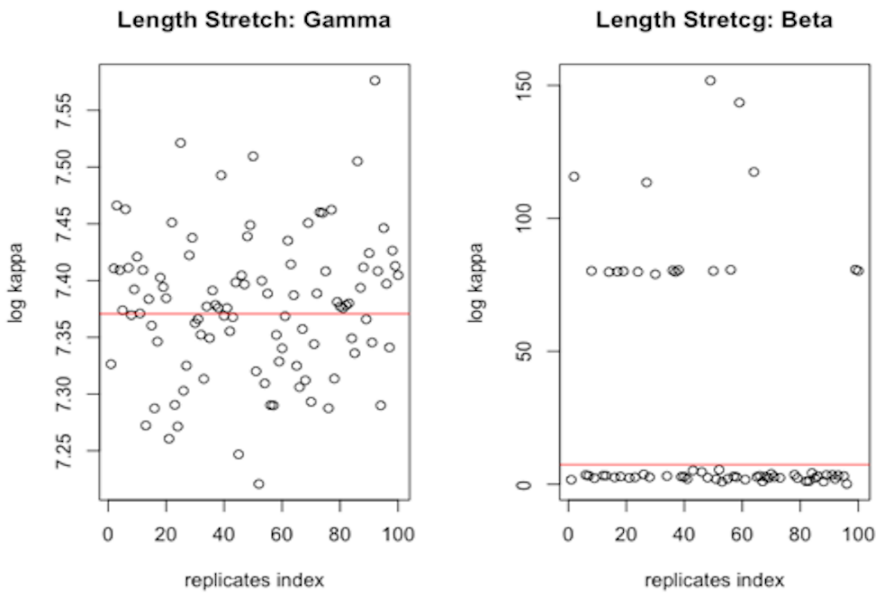

Figure 9. Condition number of the stretch trees under Gamma method and Beta method. The red line is the condition number of the raw tree.

DISCUSSION

In this work, we study the condition number for the phylogenetic tree. Although methods proposed here provide improvement for the matrix condition, it remains to explore how much the improvement can those methods make, in particular, for the parameter estimation using the update matrix.

Our simulations are similar in spirit to what was performed in the Fig. 5 in Adams and Collyer (2018) where the condition number of phylogenetic covariance matrices at differing level of sample size was shown. Their work was interested in the condition number relative to type I error of comparative methods while our examinations use different aspects of the effect of tree condition.

Notice that one common approach is to add a small constant to the eigenvalues of the phylogenetic similarity matrix and re-estimate the phylogenetic similarity matrix from the eigenvectors and adjusted eigenvalues. These are based on the observation that ill-conditioned matrices tend to have too much spread in their eigenvalues: the largest are too large and the smallest are too small. One therefore finds a way of shrinking the variation in these to obtain a more valid similarity matrix. Ledoit and Wolf (2004) outlined the conceptual and theoretical issues and provides one implementation. However, this method alter the topology of the phylogenies. So we do not consider to implement it here.

One possible extension is to look at the condition number of the phylogenetic variance covariance matrix for more general trait models. For instance, if one assumes Ornstein-Uhlenbeck process for trait evolution (Hansen et al., 2008), then the phylogenetic variance covariance matrix is $\mathbf{\Sigma} = \sigma^2 \mathbf{\Sigma}_\alpha[i,j] = \sigma^2 \exp(-\alpha t_{ij})(1-\exp(-2\alpha t_a))/(2\alpha)$ 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 $\boldsymbol{Y} \sim \mathcal{N}(\theta \mathbf{1}, \sigma^2 \mathbf{\Sigma}_\alpha)$. In particular, when $\alpha = 0$ we have $\mathbf{\Sigma}_\alpha = \boldsymbol{C}$ is the phylogenetic similarity matrix and the process for trait evolution is considered as under Brownian motion. There are many trait models developed under different assumptions based on this OU process, the corresponding phylogenetic variance covariance matrix $\mathbf{\Sigma}$ would have different and more complicated elements with more parameters and assumptions on the evolutionary process (Butler and King, 2004; Beaulieu et al., 2012; O'Meara et al., 2006; Hansen et al., 2008; Jhwueng and Maroulas, 2014, 2016).

In addition, (Bastide et al., 2018) 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.

ASSESSING THE STABILITY OF PARAMETER ESTIMATION

Olivos-Trujillo et al. (2015) code: StabilityOfEstimate.r

CONCLUSION

In this paper, we explore the condition number κ of the phylogenetic similarity matrix $\boldsymbol{C}$ transformed from a phylogenetic tree. For practical application, we searched database and analyzed 126 empirical trees from literatures and looked into their condition numbers of the phylogenetic similarity matrix. We then proposed three possible methods (shrinkage matrix regularization, tree pruning and length stretching) to alleviate the ill-condition matrix issue arising from the phylogenetic tree. We also performed simulation via generating trees under Brownian motion and compared their condition numbers with the modified tree using our proposed methods. Results showed that our methods (in particular for shrinkage matrix regularization and tree pruning methods) help to reduces the conditioned number to make the subsequent operation more stable. We hope our work can provide useful toolkits for the phylogenetic community. The R scripts can be accessed at Github: `https://github.com/djhwueng/kappapcm`.

ACKNOWLEDGMENTS

I thank Brian O'Meara and Dean Adams for their tremendous helps and communications for conceiving the problem and providing suggestion for analysis on this manuscripts.

REFERENCES

Adams, D. C. and Collyer, M. L. (2018). Multivariate phylogenetic comparative methods: evaluations, comparisons, and recommendations. *Systematic biology*, 67(1):14–31.

Ane, C. (2008). Analysis of comparative data with hierarchical autocorrelation. *Evolution*, 2(3):1078–1102.

Bartoń, K. (2013). *MuMIn: multi-model inference, R package version 1.9.13*. CRAN http://CRAN.R-project.org/package=MuMIn.

Bastide, P., Solis-Lemus, C., Kriebel, R., Sparks, K. W., and Ané, C. (2018). Phylogenetic comparative methods on phylogenetic networks with reticulations. *Systematic Biology*.

Beaulieu, J., Jhwueng, D.-C., Boettiger, C., and O'Meara, B. (2012). Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution*, 66(8):2369–2383.

Bininda-Emonds, O. R., Cardillo, M., Jones, K. E., MacPhee, R. D., Beck, R. M., Grenyer, R., Price, S. A., Vos, R. A., Gittleman, J. L., and Purvis, A. (2007). The delayed rise of present-day mammals. *Nature*, 446(7135):507.

Boettiger, C. and Temple Lang, D. (2012). Treebase: an r package for discovery, access and manipulation of online phylogenies. *Methods in Ecology and Evolution*, 3(6):1060–1066.

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

Colijn, C. and Plazzotta, G. (2018). A metric on phylogenetic tree shapes. *Systematic Biology*, 67(1):113–126.

Cybis, G. B., Sinsheimer, J. S., Bedford, T., Mather, A. E., Lemey, P., and Suchard, M. A. (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *The annals of applied statistics*, 9(2):969.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American journal of human genetics*, 25(5):471.

Felsenstein, J. (1985). Phylogeny and the comparative method. *America Naturalist*, 125(1):1–15.

FitzJohn, R. G. (2012). Diversitree: comparative phylogenetic analyses of diversification in r. *Methods in Ecology and Evolution*, 3(6):1084–1092.

Freckleton, R. P. (2012). Fast likelihood calculations for comparative analyses. *Methods in Ecology and Evolution*, 3(5):940–947.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC.

Goolsby, E. W., Bruggeman, J., and Ané, C. (2017). Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution*, 8(1):22–27.

Hadfield, J. and Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of evolutionary biology*, 23(3):494–508.

Hadfield, J. D. (2010). Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33(2):1–22.

Hansen, T., Pienaar, J., and Orzack, S. (2008). A comparative method for studying adaptation to a randomly evolving environment. 62:1965–77.

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Molecular biology and evolution*, 32(4):835–845.

Higham, N. J. (2002). *Accuracy and stability of numerical algorithms*, volume 80. Siam.

Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., Crandall, K. A., Deng, J., Drew, B. T., Gazis, R., et al. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, 112(41):12764–12769.

Ho, L. S. T. and Ané, C. (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology*, 63(3):397–408.

Horn, R. A. and Johnson, C. R., editors (1986). *Matrix Analysis*. Cambridge University Press, New York, NY, USA.

Horvilleur, B. and Lartillot, N. (2014). Monte carlo algorithms for brownian phylogenetic models. *Bioinformatics*, 30(21):3020–3028.

Jhwueng, D.-C. (2010). *Some problems in phylogenetic comparative method*. PhD thesis, Indiana University Bloomington.

Jhwueng, D.-C. (2013). Assessing the goodness of fit of phylogenetic comparative methods: A meta-analysis and simulation study. *PLOS ONE*, 8(6):1–12.

Jhwueng, D.-C. and Maroulas, V. (2014). Phylogenetic ornstein-uhlenbeck regression curves. *Statistics and Probability Letters*, 89:110–117.

Jhwueng, D.-C. and Maroulas, V. (2016). Adaptive trait evolution in random environment. *Journal of Applied Statistics*, 43(12):2310–2324.

Ledoit, O. and Wolf, M. (2004). Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119.

Maddison, D. and Schulz, K.-S. (2007). The tree of life web project.

Michonneau, F., Brown, J. W., and Winter, D. J. (2016). rotl: an r package to interact with the open tree of life data. *Methods in Ecology and Evolution*, 7(12):1476–1481.

Mir, A., Rosselló, F., and Rotger, L. (2013). A new balance index for phylogenetic trees. *Mathematical Biosciences*, 241(1):125–136.

Nabben, R. and Varga, R. S. (1994). A linear algebra proof that the inverse of a strictly ultrametric matrix is a strictly diagonally dominant stieltjes matrix. *SIAM Journal on Matrix Analysis and Applications*, 15(1):107–113.

Olivos-Trujillo, M., Gajardo, H. A., Salvo, S., González, A., and Muñoz, C. (2015). Assessing the stability of parameters estimation and prediction accuracy in regression methods for estimating seed oil content in brassica napus l. using nir spectroscopy. In *Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), 2015 CHILEAN Conference on*, pages 25–30. IEEE.

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

O'Meara, B., Heath, T., Midford, P. E., Chamberlain, S., Brown, J. W., and Schliep, K. (2016). datelife: 0.2.3.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877.

Paradis, E., Claude, J., and Strimmer, K. (2004). Ape: Analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290.

Parlett, B. N. (1987). *The Symmetric Eigenvalue Problem (Classics in Applied Mathematics)*. Society for Industrial and Applied Mathematics.

Piel, W. H., Donoghue, M. J., and Sanderson, M. J. (2002). Treebase: a database of phylogenetic knowledge. (No. 171):41–47.

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., Gray, R. R., Arinaminpathy, N., Stramer, S. L., Busch, M. P., et al. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences*, page 201206598.

Sanderson, M. J., Donoghue, M. J., Piel, W. H., and Eriksson, T. (1994). TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *American Journal*

of Botany, 81(6):183+.

Schafer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. 4:Article32.

Sokal, R., Michener, C., and of Kansas, U. (1958). *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas science bulletin. University of Kansas.

Stadler, T. (2009). On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, 261(1):58 – 66.

Stadler, T. (2011). Simulating trees with a fixed number of extant species. *Systematic Biology*, 60(5):676–684.

Stoltzfus, A., Lapp, H., Matasci, N., Deus, H., Sidlauskas, B., Zmasek, C. M., Vaidya, G., Pontelli, E., Cranston, K., Vos, R., et al. (2013). Phylotastic! making tree-of-life knowledge accessible, reusable and convenient. *BMC bioinformatics*, 14(1):158.

Theiler, J. (2012). The incredible shrinking covariance estimator.

Tung Ho, L. s. and Ané, C. (2014). A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Systematic biology*, 63(3):397–408.