Length Stretching

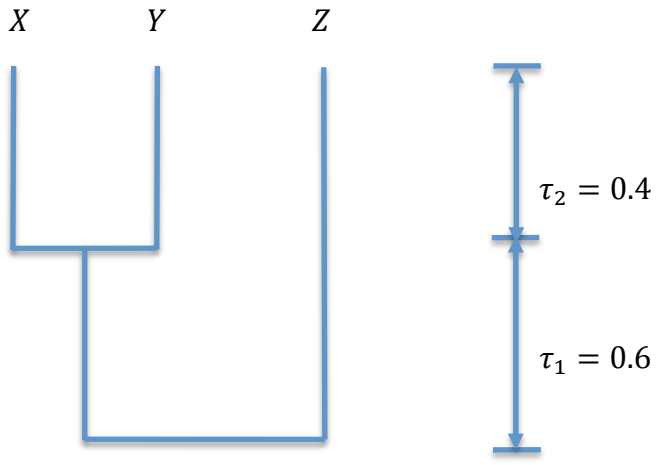
We can obtain a new tree by stretching the branch lengths of the raw tree without changing its topology. The C matrix and its condition number κ for the new tree will change accordingly.

For an unltrametric tree, let $\boldsymbol{\tau}$ be the tree height from the root to the tip. Without loss of generality, we can first scale $\boldsymbol{\tau}$ into a unit and then decompose $\boldsymbol{\tau}$ into d components. That is, $1 = \boldsymbol{\tau} = \tau_1 + \tau_2 \cdots + \tau_d$ where $\tau_i > 0, i = 1,2,\cdots,d$ 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 length for the tip species (with minimum tip length) evolved from its most recent common ancestor.

Next, consider the matrix 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 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 following equation

$$\tau_{i-1} = c_i - c_{i-1},\ i = 2, \cdots, d+1.$$

An example: given a 3 taxa tree with $\tau_1 = 0.6$ and $\tau_2 = 0.4$ as following



Then the C matrix is

$$C = \begin{array}{c} \\ X \\ Y \\ X \end{array} \begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{pmatrix} c_3 & c_2 & c_1 \\ c_2 & c_3 & c_2 \\ c_1 & c_2 & c_3 \end{pmatrix} \end{array} = \begin{pmatrix} 1 & 0.6 & 0 \\ 0.6 & 1 & 0.6 \\ 0 & 0.6 & 1 \end{pmatrix}$$

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 $\sum_{i=1}^{d} \tau_i = 1$ we can treat $(\tau_1, \tau_2, \cdots, \tau_d)$ as a d-dimensional random vector from a Dirichlet distribution. $(\tau_1, \tau_2, \cdots, \tau_d)$ can be generated by first drawing d independent gamma random variables, $T_i, i = 1,2, \cdots, d,$ each with different shape parameters $k\tau_i, i = 1,2, \cdots, d$ and rate parameter 1 where k is an arbitrary but positive constant, then the d-tuple vector $(t_1^*, t_2^*, \cdots, t_d^*) = \frac{1}{\sum_{i=1}^{d} T_i}(T_1, T_2, \cdots, T_d)$ is a Dirichlet random vector with $t_i^* \in (0,1)$, $\sum_{i=1}^{d} t_i^* = 1$, and concentration parameters $k\tau_i, i = 1,2, \cdots, d$. Here the positive constant k is an arbitrary scaling variable that always preserves the correct mean. By the property of Dirichlet distribution, we have

$$E(t_i^*) = \frac{k\tau_i}{\sum_{i=1}^{d} k\tau_i} = \frac{k\tau_i}{k\sum_{i=1}^{d} \tau_i} = \tau_i,$$

and the mode given by

$$M_{t_i^*} = \frac{k\tau_i - 1}{(\sum_{i=1}^{d} k\tau_i) - d} = \frac{k\tau_i - 1}{k - d} \text{ where } k\tau_i > 1, i = 1,2, \cdots, d.$$

The choice of k is thus determined by $\min_{1 \le i \le d} k\tau_i > 1$. We can choose a positive integer k satisfies

$$k = \left\lceil \frac{1}{\min_{1 \le i \le d} \tau_i} \right\rceil$$

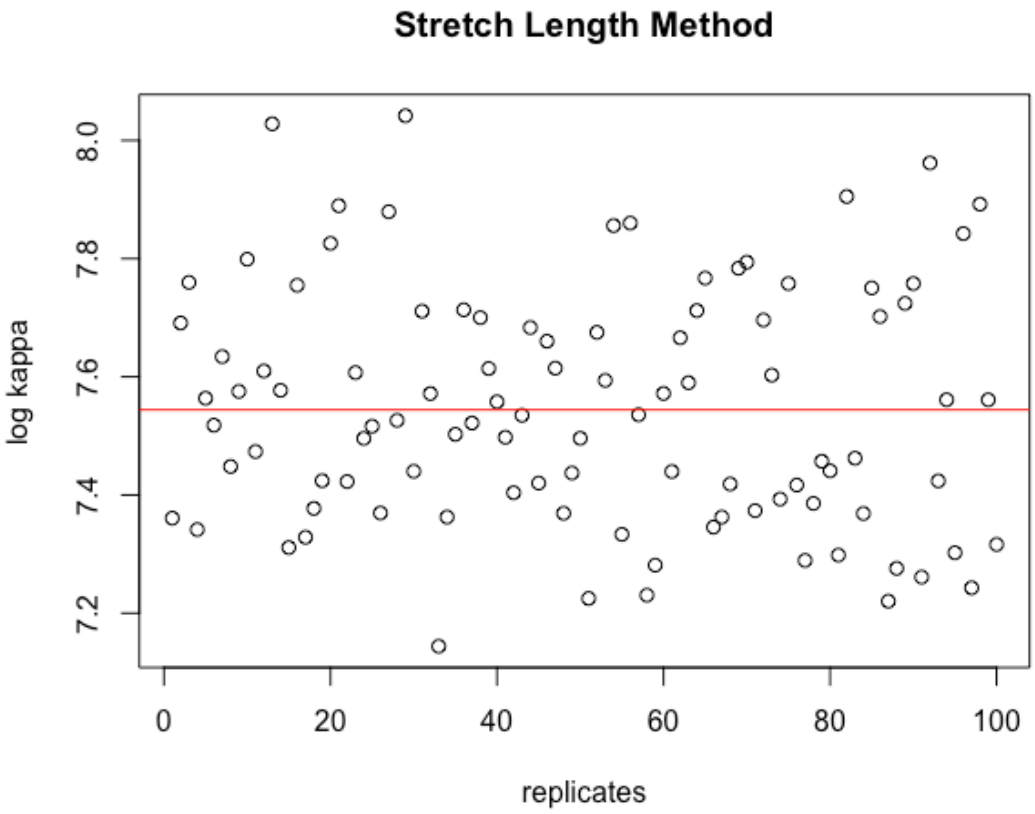
where $\lceil a \rceil$ returns the integer closest but less than the real number a.

Because the mode $M_{t_i^*}$ of t_i^* is not equal to the expectation of t_i^*, such choice of k does not guarantee that the distribution of t_i^* is centered or symmetric around its mean τ_i. The mode converges to the expected mean when k approach to infinity (i.e. $\frac{k\tau_i - 1}{k - d} \longrightarrow \tau_i$ as $k \longrightarrow \infty$). However, although choosing

larger k helps to center the distribution around τ_i, picking k too large will cause the samples t_i^* to be tightly centered around the given estimates $(\hat{\tau}_1, \hat{\tau}_2, \cdots, \hat{\tau}_d)$. The choice of k 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.

To test this method, we first simulate a tree, and then calculate its condition number κ. We then stretch the length of the tree and calculate the corresponding condition number κ'. The following is a comparison of the two trees of 10 taxa.

We perturb the raw tree 100 times to get 100 new trees where their condition numbers are plotted in the following figure. The red line is the condition number of the raw tree.



In short, this method could give us a better or worse condition number. Maybe we can choose a good tree with lower κ from them?