Unexplained Behavior of Phylogeny Estimation Methods

Achutha Balaji, Yuheng Cai, Ruoran Huang, Megan Lolling, Mengwei Sun
Max Bacharach, Sebastien Roch

December 2022

1 Introduction

A phylogenetic tree is a diagram that depicts the lines of evolutionary descent of different species that share a common ancestor. The goal of phylogeny estimation using Maximum Likelihood Estimation (MLE) is to recover the topology of a phylogenetic tree given datasets of genes, species, and other taxa. Long Branch Attraction (LBA) is a systemic error incurred when distantly related i.e. long-branched lineages, instead appear to be closely related.

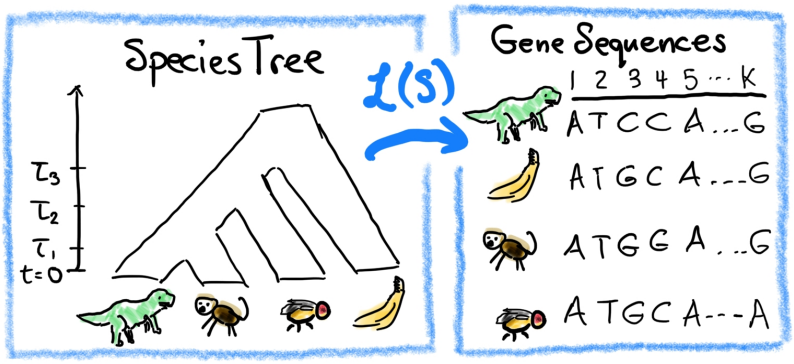

Figure 1: From gene sequence data to a phylogenetic tree.

This project aims to investigate whether Long Branch Attraction (LBA) introduces bias to MLE in phylogenetic tree reconstruction even for simple evolutionary models and small-scale trees. Specifically, we aim to compute the optimality conditions (i.e. the tree parameters that give the maximum likelihood) of all possible configurations for the unrooted four-species tree with two long branches through a combination of analytical solutions to the MLE equation and minimization software. After implementing a Homotopy Continuation approach and a Least Squares Optimization approach in Julia, we find that the results from the two methods only agrees at certain boundary cases. The homotopy continuation methods do not yield real solutions when there are two extremely long branches. Although we work out a case of the quartet 13|24 and find that homotopy continuation method only produce real solutions when the dataset satisfies some constraints, further research into the inconsistencies between two methods is needed.

1.1 The Data

In this project, we consider gene sequence data generated according to the CFN site substitution model, which models nucleotide evolution along an evolutionary tree. Under this model, there are two nucleotides, which we identify as -1 and $+1$. The parameters for this model consist of a species tree with associated edge lengths. We consider an unrooted species tree with four leaves $1, 2, 3, 4$. Each time this process is run, a nucleotide ± 1 is obtained at each of the leaves. The output is a random vector X with four entries, corresponding to the nucleotide observed at the four leaves. Running the process k times independently, we

obtain k observations $\sigma^{(1)}, \ldots, \sigma^{(k)}$. Each observation is an element in the following 16-element set S:

$$\begin{aligned} S &\coloneqq \left\{(\sigma_1, \sigma_2, \sigma_3, \sigma_4)^\top : \sigma_1, \sigma_2, \sigma_3, \sigma_4 \in \{-1, +1\}\right\} \\ &= \left\{(-1,-1,-1,-1)^\top, (-1,-1,-1,+1)^\top, (-1,-1,+1,-1)^\top, \ldots, (+1,+1,+1,+1)^\top\right\}. \end{aligned}$$

The elements of S are called as **site patterns**; these are the possible outcomes of running the site substitution process once on the tree, which generates a single binary nucleotide (either $+1$ or -1) for each of the four leaves of the species tree. For example, the site pattern $(-1,-1,+1,-1)$ represents the event that nucleotide -1 is observed in species 1,2, and 4, and that the nucleotide $+1$ is observed in species 3.

Example 1. *Suppose we observe a segment of DNA consisting of $k = 5$ sites, from each of four species.*

$$\begin{aligned} &\textit{Species 1}: -1,-1,+1,-1,+1 \\ &\textit{Species 2}: -1,-1,+1,-1,+1 \\ &\textit{Species 3}: -1,+1,+1,+1,+1 \\ &\textit{Species 4}: -1,-1,+1,+1,+1 \end{aligned}$$

We observe that $\sigma^{(1)} = (-1,-1,-1,-1)^\top$, $\sigma^{(2)} = (-1,-1,+1,-1)^\top$, $\sigma^{(3)} = \sigma^{(5)} = (+1,+1,+1,+1)$, and $\sigma^{(4)} = (-1,-1,+1,+1)^\top$.

1.2 Site Pattern Probabilities

The probability of observing a given site pattern depends on the *topology* of the tree as well as its *branch lengths*. In our situation, we will consider three main cases, which are shown in Figure 2, and which correspond to the three possible unrooted tree topologies for a tree with four leaves. In addition, the tree edges have associated edge parameters $\theta_1, \ldots, \theta_5 \in [0,1]$, which measure the amount of correlation between two nucleotides at the ends of the edge: if $\theta_i = 0$ then the nucleotides are independent, and if $\theta_i = 1$ then they are always the same.

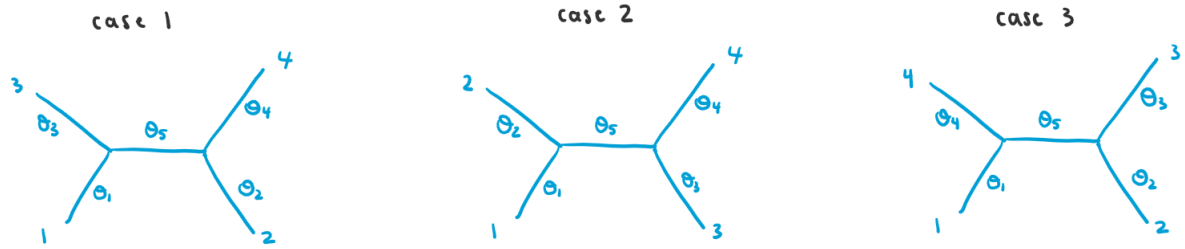

Figure 2: The 3 main cases considered in this project.

Using a general formula from [3], we obtain the formulas for the probability mass function of X in each of the three cases in Figure 2. For each $\sigma = (\sigma_1, \sigma_2, \sigma_3, \sigma_4)^\top \in S$, we have the following probabilities:

Case 1:

$$\begin{aligned} \mathbb{P}[X = \sigma] = \frac{1}{16}(1 &+ \sigma_1\sigma_2\theta_1\theta_5\theta_2 + \sigma_1\sigma_3\theta_1\theta_3 + \sigma_1\sigma_4\theta_1\theta_5\theta_4 + \sigma_2\sigma_3\theta_2\theta_5\theta_3 + \sigma_2\sigma_4\theta_2\theta_4 \\ &+ \sigma_3\sigma_4\theta_3\theta_5\theta_4 + \sigma_1\sigma_2\sigma_3\sigma_4\theta_1\theta_3\theta_2\theta_4) \end{aligned}$$

Case 2:

$$\begin{aligned} \mathbb{P}[X = \sigma] = \frac{1}{16}(1 &+ \sigma_1\sigma_2\theta_1\theta_2 + \sigma_1\sigma_3\theta_1\theta_5\theta_3 + \sigma_1\sigma_4\theta_1\theta_5\theta_4 + \sigma_2\sigma_3\theta_2\theta_5\theta_3 + \sigma_2\sigma_4\theta_2\theta_5\theta_4 \\ &+ \sigma_3\sigma_4\theta_3\theta_4 + \sigma_1\sigma_2\sigma_3\sigma_4\theta_1\theta_2\theta_3\theta_4) \end{aligned}$$

Case 3:

$$\mathbb{P}[X=\sigma] = \frac{1}{16}(1+\sigma_1\sigma_2\theta_1\theta_5\theta_2+\sigma_1\sigma_3\theta_1\theta_5\theta_3+\sigma_1\sigma_4\theta_1\theta_4+\sigma_2\sigma_3\theta_2\theta_3+\sigma_2\sigma_4\theta_2\theta_5\theta_4 \\ +\sigma_3\sigma_4\theta_3\theta_5\theta_4+\sigma_1\sigma_2\sigma_3\sigma_4\theta_1\theta_4\theta_2\theta_3)$$

We also make the simplifying assumption that the 'distance' between leaves 1 and 2 is known to be a fixed $\tau = 1/2$. In particular, we assume that $\theta_1\theta_2\theta_5 = 1/2$ in Cases 1 and 3, and that $\theta_1\theta_2 = 1/2$ in Case 2.

1.2.1 True Probabilities

We will assume that the data is generated under a model in which the true species tree has configuration 13|24 and has the following parameters: $\theta_3 = \theta_4 = 0$, and $\theta_1\theta_2\theta_5 = \tau$. Throughout this project, we assume $\tau = 1/2$.

Definition 1 (True Probabilities)**.** *For each site pattern $\sigma \in S$, define $f_\sigma \coloneqq \mathbb{P}[X=\sigma]$ be the true probability of observing σ under the model parameters described above; that is,*

$$f_\sigma = \frac{1}{16}(1+\tau\sigma_1\sigma_2) = \frac{1}{16}\left(1+\frac{1}{2}\sigma_1\sigma_2\right) \tag{1}$$

1.3 Approximating Data with the Multivariate Normal Distribution

In this section, we show how we approximate large-k data using the Central Limit Theorem. For this section, it will be convenient to denote the 16 site patterns in S as $s_1, \ldots, s_{16}$, and to denote the probability of the site patterns as $q_1, \ldots, q_{16}$ respectively. Then, for each $i = 1, \ldots, k$, define the random vector

$$Y_i = e_j \in \mathbb{R}^{16} \text{ if and only if } \sigma^{(i)} = s_j$$

where $e_1, \ldots, e_{16}$ are the standard basis vectors of $\mathbb{R}^{16}$. Therefore, Y_i indicates which site pattern was observed at the i^{th} site. From here we define the random vector

$$F_k = Y_1 + \cdots + Y_k \in \mathbb{R}^{16}$$

to be the **frequency vector**. The j-th component of F_k tells us how many times we observed site pattern s_j out of k total observations in our data set. It is important to note that $Y_1, \ldots, Y_k$ are i.i.d categorical random variables with the probability mass function $\mathbb{P}[Y_i = e_j] = q_j$. Thus, we can think of each Y_i as a special case of the multinomial distribution with the number of trials $n = 1$ (and it follows that F_k follows the multinomial distribution with the number of trials $n = k$ and outcome probabilities $q_1, \ldots, q_{16}$). By the properties of the multinomial distribution, we can immediately obtain the mean vector $\underline{\mu} = \mathbb{E}[Y_i] = (q_1, q_2, \ldots, q_{16})^T$ of Y_i, and the covariance matrix Σ of Y_i is formed by the diagonal entries $\text{Var}((Y_i)_j) = q_j(1-q_j)$ for $j = 1, \ldots, 16$ and the off-diagonal entries $\text{Cov}((Y_i)_j, (Y_i)_k) = -q_jq_k$ for $j, k = 1, \ldots, 16$ with $j \neq k$. Equivalently, $\underline{\mu} = \underline{q}$ and $\Sigma = \text{diag}(\underline{q}) - \underline{q}\underline{q}^T$ where $\underline{q} = (q_1, q_2, \ldots, q_{16})^\top$. Both $\underline{\mu}$ and Σ will be important when we want to generate gene sequence data. We will use the Central Limit Theorem for this purpose.

Theorem 1 (Central Limit Theorem)**.** *Suppose we have n i.i.d random vectors $X_1, \ldots, X_n \in \mathbb{R}^k$ each with mean vector $\mu = \mathbb{E}[X_i]$ and covariance matrix Σ. The multivariate Central Limit Theorem states*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) \xrightarrow{d} \mathcal{N}_k(0, \Sigma)$$

In other words, the multidimensional Central Limit Theorem states that the scaled component-wise summation of i.i.d random vectors converges to a multivariate normal distribution.

As mentioned previously, our data F_k takes the form of a frequency vector which species how often each site pattern is observed in the data. In other words, we can generate data by sampling an F_k random vector. As we will see in this section, we can sample F_k just by sampling a point from a multivariate Gaussian distribution via the multidimensional Central Limit Theorem (if we choose a large enough k). Let $\underline{\mu}$ and Σ be the mean vector and covariance matrix of Y_i, respectively (defined above). By Theorem 1, we have that

$$\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left(Y_i-\underline{\mu}\right)\xrightarrow{d}\mathcal{N}_{16}(0,\Sigma)$$

Recalling that $F_k = Y_1 + \cdots + Y_k$, we have

$$\frac{1}{\sqrt{k}}\left(F_k-k\underline{\mu}\right)\xrightarrow{d}\mathcal{N}_{16}(0,\Sigma)$$

and so

$$F_k\xrightarrow{d}\mathcal{N}_{16}(k\underline{\mu},k\Sigma)$$

where

$$k\underline{\mu}=k\underline{q}\in\mathbb{R}^{16}\text{ and }k\Sigma=k\left(\operatorname{diag}(\underline{q})-\underline{q}\underline{q}^T\right)\in\mathbb{R}^{16\times16}$$

Therefore, we can generate frequency vectors by sampling points from a multivariate Gaussian distribution with mean vector $k\underline{\mu}$ and covariance matrix $k\Sigma$.

2 From *Maximum Likelihood* to *Least Squares*

One of the ways to find the tree that best fit the given data is through maximum likelihood estimation. For convenience, we use the negative log likelihood function. Given k independent observations $\sigma^{(1)},\ldots,\sigma^{(k)}$ the negative log likelihood function is

$$-\log(L)=-\sum_{i=1}^{k}\log\mathbb{P}(X_i=\sigma^{(i)})=-\sum_{i=1}^{k}\log\hat{g}_{\sigma^{(i)}}.$$

Here, $\hat{g}_{\sigma^{(i)}}$ denotes the estimated probability of observing $\sigma^{(i)}$ for a given choice of tree and parameters. We define the observed frequencies as $\hat{f}_\sigma$ the fraction of our k samples of F_k (defined in Section 1.3) for which site pattern σ is observed. By Theorem 1, we assume that

$$\hat{f}_\sigma\approx f_\sigma+\frac{z_\sigma}{\sqrt{k}}\tag{2}$$

where z_σ is a component of the data vector $z=\sqrt{k}\left(\frac{F_k}{k}-\underline{q}\right)$.

In addition to the approximation assumption of Eq. (2), we also make the assumption that for each σ, the estimated probability $\hat{g}_\sigma$ is close to the true probability f_σ:

$$\hat{g}_\sigma\approx f_\sigma+\frac{\hat{\epsilon}_\sigma}{\sqrt{k}}\tag{3}$$

so that

$$\hat{\epsilon}_\sigma=(\hat{g}_\sigma-f_\sigma)/\sqrt{k}.$$

The idea justifying Eq. (3), which we do not prove, is that if the estimated probabilities are far away from the true probabilities, then the likelihood of the estimate will be very low, so such estimates can be ignored. Using this assumption, next lemma gives approximations for $\hat{\epsilon}_\sigma$ when k is large. See Appendix A for poof.

Lemma 1 (Rescaling: Formulas for $\hat{\epsilon}_\sigma$)**.** *For every $\sigma\in S$, we have the following approximations for $\hat{\epsilon_\sigma}$:*

$$\hat{\epsilon}_\sigma=\begin{cases}\sigma_1\sigma_3\hat{\theta_1}\hat{\alpha_3}+\sigma_2\sigma_4\hat{\theta_2}\hat{\alpha_4}+\sigma_1\sigma_4\hat{\theta_1}\hat{\theta_5}\hat{\alpha_4}+\sigma_2\sigma_3\hat{\theta_2}\hat{\theta_5}\hat{\alpha_3} & \textit{in Case 1.}\\ \sigma_1\sigma_3\hat{\theta_1}\hat{\theta_5}\hat{\alpha_3}+\sigma_2\sigma_4\hat{\theta_2}\hat{\theta_5}\hat{\alpha_4}+\sigma_1\sigma_4\hat{\theta_1}\hat{\theta_5}\hat{\alpha_4}+\sigma_2\sigma_3\hat{\theta_2}\hat{\theta_5}\hat{\alpha_3} & \textit{in Case 2.}\\ \sigma_1\sigma_4\hat{\theta_1}\hat{\alpha_4}+\sigma_2\sigma_3\hat{\theta_2}\hat{\alpha_3}+\sigma_1\sigma_3\hat{\theta_1}\hat{\theta_5}\hat{\alpha_3}+\sigma_2\sigma_4\hat{\theta_2}\hat{\theta_5}\hat{\alpha_4} & \textit{in Case 3.}\end{cases}$$

where $\alpha_i\coloneqq\sqrt{k}\hat{\theta}_i$ for $i=3,4$.

We then use Taylor Expansion and Eqs. (2) and (3) to simplify the maximum likelihood expression into a least square equation were only higher order terms of z_σ and $\hat{\epsilon}_\sigma$ are involved. The full details of this calculation are lengthy and are provided in Appendix B, but the key result is the following:

$$\begin{aligned} -\log(L) &\approx -\sum_\sigma k\left(f_\sigma + \frac{z_\sigma}{\sqrt{k}}\right)\ln\left(f_\sigma + \frac{\hat{\epsilon}_\sigma}{\sqrt{k}}\right) \\ &= -\sum_\sigma ka\ln a + \sum_\sigma \sqrt{k}z_\sigma + \frac{1}{2}\sum_\sigma \frac{1}{a}\left(\hat{\epsilon}_\sigma - z_\sigma\right)^2 + O(\frac{1}{\sqrt{k}}) \end{aligned}$$

where $a = f_\sigma + \frac{z_\sigma}{\sqrt{k}}$. A second Taylor Expansion for $a = f_\sigma + \frac{z_\sigma}{\sqrt{k}}$ then gives

$$-\log(L) = G(k, f, z) + \frac{1}{2}\sum_\sigma \frac{1}{f_\sigma}(\hat{\epsilon}_\sigma - z_\sigma)^2 + O(\frac{1}{\sqrt{k}})$$

where the first term G does not depend on $\hat{\epsilon}$. The significance of this equation is that the second term on the right-hand side is the only non-negligible term that depends on $\hat{\epsilon}$. So for large k, maximizing the likelihood is equivalent to minimizing the *least squares* term:

$$\frac{1}{2}\sum_{\sigma\in S}\frac{1}{f_\sigma}\left(\hat{\epsilon}_\sigma - z_\sigma\right)^2. \tag{4}$$

Using Lemma 1, we get an explicit formula for this in each of our three cases. This allows us to consider only optimizing the least square problems in Section 3.

3 Software Approaches to Optimization

3.1 Nonlinear Optimization

We wish to minimize the objective function given in Eq. (4), so we have the following nonlinear least squares optimization problem for each of the 3 main cases in Fig. 2 (see also Appendix C for more precise specification):

$$\min_{\hat{\theta_1},\hat{\theta_2},\hat{\alpha_3},\hat{\alpha_4},\hat{\alpha_5}} \quad \frac{1}{2}\sum_{\sigma\in S}\frac{1}{f_\sigma}\left(\hat{\epsilon}_\sigma - z_\sigma\right)^2.$$

Of the three main cases shown in Figure 2, each has several **boundary cases**, which are subcases in which one or more of the edge parameters $\theta_1, \ldots, \theta_5$ are equal to either 0 or 1. (Not all choices are valid, however. For example, due to the constraint that $\theta_1\theta_2\theta_5 = 1/2$ in Case 1, it is not possible for $\theta_i = 0$, $i = 1, 2, 5$, and it is also not possible for $\theta_1 = \theta_2 = \theta_5 = 1$). Also, since we make the change of variable $\alpha_i = \sqrt{k}\theta_i$ for $i = 3, 4$ the boundary values for those two parameters are 0 and $\sqrt{k}$.

The procedure was first to generate data z, and then to minimize Eq. (4) over all basic models for option 1,2,3, and all the boundary cases for each option. We used the nonlinear optimization packages JuMP and Ipopt in Julia. And we keep track of the cumulative times that each option is selected. And we also show which special cases are preferred given this configuration.

3.2 Homotopy Continuation

As above, our goal is to minimize the least squares formula in Eq. (4). The general approach here was as follows. First, define a constraint, g, for cases 1 and 3 to be $\theta_1\theta_2\theta_5 = \frac{1}{2}$, and our constraint, g, for case 2 to be $\theta_1\theta_2 = \frac{1}{2}$. Then, applying the Lagrange Multiplier method, we set up our Lagrangian objective function as:

$$\frac{1}{2}\sum_\sigma \frac{1}{f_\sigma}(z_\sigma - \epsilon_\sigma)^2 + \lambda\cdot g$$

After differentiating concerning our branch lengths, we get a system of equations of length 6 in terms of our variables $\theta_1, \theta_2, \theta_5, \alpha_3, \alpha_4$, and λ. Given a set of simulated data, we solve the system of equations using the

homotopyContinuation package [1] in Julia and output real solutions of our branch length parameters as sets of critical points to compare with the least squares optimization software. To account for our different boundary cases, we add constraints (ex. $\theta_1 = 1$) and a corresponding number of λ_i's, then proceed with the same Lagrange multipliers procedure.

We also attempted to apply this approach by combining Langrange multipliers with homotopy in several other ways as well. For example, we used it to maximize the likelihood function directly subject to certain algebraic constraints of the CFN model found in [2].

4 Discussion/Conclusion

In this project, we aimed to find whether Long Branch Attraction (LBA) will introduce bias to the Maximum Likelihood method used for the phylogeny estimation of a quartet tree. First, through Taylor Expansion, we were allowed to reduce our likelihood function to a least square polynomial for the numerical approximation methods, which are the Homotopy Continuation method and the Nonlinear Optimization approach. We also found that we can approximate large-k data using the Central Limit Theorem. Then, we implemented a Homotopy Continuation approach and used the Nonlinear Optimization Package in Julia to find the analytical solutions to our MLE phylogeny estimation problem. The two minimization software do not always yield consistent results. For certain boundary cases, e.g., when branches 3 and 4 are infinitely long, Homotopy Continuation produces no real solutions, while the Least Squares Optimization software still finds a result. Even when starting from the given solution, Homotopy Continuation cannot compute a start system due to the added constraint variables. After working out a case with $\theta_3 = 0$ and $\theta_1 = 1$ analytically, we verified that there are two constraints on the data for the Homotopy Continuation Method to yield real solutions (See Lemma 2 in Appendix D). And at the same time, the nonlinear package always seems to find a "better" solution with a smaller objective function value than the Homotopy method. However, further research is needed to determine why this occurs.

A Proof of Lemma 1

- Case 1:

  $\hat{\epsilon}_\sigma = \sqrt{k}(\hat{g}_\sigma - f_\sigma)$
  $= \sqrt{k}(-\frac{1}{16}(1 + \sigma_1\sigma_2\tau + \sigma_1\sigma_3\hat{\theta_1}\hat{\theta_3} + \sigma_2\sigma_4\hat{\theta_2}\hat{\theta_4} + \sigma_1\sigma_4\hat{\theta_1}\hat{\theta_5}\hat{\theta_4} + \sigma_2\sigma_3\hat{\theta_2}\hat{\theta_5}\hat{\theta_3} + \sigma_3\sigma_4\hat{\theta_3}\hat{\theta_5}\hat{\theta_4} + \sigma_1\sigma_2\sigma_3\sigma_4\hat{\theta_3}\hat{\theta_1}\hat{\theta_4}\hat{\theta_2})).$

  We can re-scale $\hat{\theta_3}$ and $\hat{\theta_4}$ by substituting them into $\frac{\hat{\alpha_3}}{\sqrt{k}}$ $\frac{\hat{\alpha_4}}{\sqrt{k}}$, $\sigma_3\sigma_4\hat{\theta_3}\hat{\theta_5}\hat{\theta_4} = O(\frac{1}{k})$, and $\sigma_1\sigma_2\sigma_3\sigma_4\hat{\theta_3}\hat{\theta_1}\hat{\theta_4}\hat{\theta_2} = O(\frac{1}{k})$ as below:

  $$\hat{\epsilon}_\sigma = \sigma_1\sigma_3\hat{\theta_1}\hat{\alpha_3} + \sigma_2\sigma_4\hat{\theta_2}\hat{\alpha_4} + \sigma_1\sigma_4\hat{\theta_1}\hat{\theta_5}\hat{\alpha_4} + \sigma_2\sigma_3\hat{\theta_2}\hat{\theta_5}\hat{\alpha_3}$$

  with $\hat{\theta_1}\hat{\theta_5}\hat{\theta_2} = \tau$

- Case 2: option 1

  $\hat{\epsilon}_\sigma = \sqrt{k}(\hat{g}_\sigma - f_\sigma)$
  $= \sqrt{k}(-\frac{1}{16}(1 + \sigma_1\sigma_2\tau + \sigma_3\sigma_4\hat{\theta_3}\hat{\theta_4} + \sigma_1\sigma_3\hat{\theta_1}\hat{\theta_5}\hat{\theta_3} + \sigma_2\sigma_4\hat{\theta_2}\hat{\theta_5}\hat{\theta_4} + \sigma_1\sigma_4\hat{\theta_1}\hat{\theta_5}\hat{\theta_4} + \sigma_2\sigma_3\hat{\theta_2}\hat{\theta_5}\hat{\theta_3} + \sigma_1\sigma_2\sigma_3\sigma_4\hat{\theta_2}\hat{\theta_1}\hat{\theta_4}\hat{\theta_3})).$

  We can re-scale $\hat{\theta_3}$ and $\hat{\theta_4}$ by substituting them into $\frac{\hat{\alpha_3}}{\sqrt{k}}$ $\frac{\hat{\alpha_4}}{\sqrt{k}}$, $\sigma_3\sigma_4\hat{\theta_3}\hat{\theta_4} = O(\frac{1}{k})$, and $\sigma_1\sigma_2\sigma_3\sigma_4\hat{\theta_2}\hat{\theta_1}\hat{\theta_4}\hat{\theta_3} = O(\frac{1}{k})$ as below:

  $$\hat{\epsilon}_\sigma = \sigma_1\sigma_3\hat{\theta_1}\hat{\theta_5}\hat{\alpha_3} + \sigma_2\sigma_4\hat{\theta_2}\hat{\theta_5}\hat{\alpha_4} + \sigma_1\sigma_4\hat{\theta_1}\hat{\theta_5}\hat{\alpha_4} + \sigma_2\sigma_3\hat{\theta_2}\hat{\theta_5}\hat{\alpha_3}$$

  with $\hat{\theta_1}\hat{\theta_2} = \tau$

- Case 2: option 2

  $\hat{\epsilon}_\sigma = \sqrt{k}(\hat{g}_\sigma - f_\sigma)$
  $= \sqrt{k}(-\frac{1}{16}(1 + \sigma_1\sigma_2\tau + \sigma_3\sigma_4\hat{\theta_3}\hat{\theta_4} + \sigma_1\sigma_3\hat{\theta_1}\hat{\theta_5}\hat{\theta_3} + \sigma_2\sigma_4\hat{\theta_2}\hat{\theta_5}\hat{\theta_4} + \sigma_1\sigma_4\hat{\theta_1}\hat{\theta_5}\hat{\theta_4} + \sigma_2\sigma_3\hat{\theta_2}\hat{\theta_5}\hat{\theta_3} + \sigma_1\sigma_2\sigma_3\sigma_4\hat{\theta_2}\hat{\theta_1}\hat{\theta_4}\hat{\theta_3})).$

  We could also re-scale $\hat{\theta_3}$, $\hat{\theta_4}$, and $\hat{\theta_5}$ by substituting them into $\frac{\hat{\alpha_3}}{k^{1/4}}$, $\frac{\hat{\alpha_4}}{k^{1/4}}$, $\frac{\hat{\alpha_5}}{k^{1/4}}$, $\sigma_3\sigma_4\hat{\theta_3}\hat{\theta_4} = O(\frac{1}{\sqrt{k}})$ as below:

  $$\hat{\epsilon}_\sigma = \sigma_3\sigma_4\hat{\alpha_3}\hat{\alpha_4} + \sigma_1\sigma_3\hat{\theta_1}\hat{\alpha_5}\hat{\alpha_3} + \sigma_2\sigma_4\hat{\theta_2}\hat{\alpha_5}\hat{\alpha_4} + \sigma_1\sigma_4\hat{\theta_1}\hat{\alpha_5}\hat{\alpha_4} + \sigma_2\sigma_3\hat{\theta_2}\hat{\alpha_5}\hat{\alpha_3} + \sigma_1\sigma_2\sigma_3\sigma_4\hat{\theta_2}\hat{\theta_1}\hat{\alpha_4}\hat{\alpha_3}$$

  with $\hat{\theta_1}\hat{\theta_2} = \tau$

- Case 3:

  $\hat{\epsilon}_\sigma = \sqrt{k}(\hat{g}_\sigma - f_\sigma)$
  $= \sqrt{k}(-\frac{1}{16}(1 + \sigma_1\sigma_2\tau + \sigma_1\sigma_4\hat{\theta_1}\hat{\theta_4} + \sigma_2\sigma_3\hat{\theta_2}\hat{\theta_3} + \sigma_1\sigma_3\hat{\theta_1}\hat{\theta_5}\hat{\theta_3} + \sigma_2\sigma_4\hat{\theta_2}\hat{\theta_5}\hat{\theta_4} + \sigma_3\sigma_4\hat{\theta_3}\hat{\theta_5}\hat{\theta_4} + \sigma_1\sigma_2\sigma_3\sigma_4\hat{\theta_4}\hat{\theta_1}\hat{\theta_3}\hat{\theta_2})).$

  We can re-scale $\hat{\theta_3}$ and $\hat{\theta_4}$ by substituting them into $\frac{\hat{\alpha_3}}{\sqrt{k}}$ $\frac{\hat{\alpha_4}}{\sqrt{k}}$, $\sigma_3\sigma_4\hat{\theta_3}\hat{\theta_5}\hat{\theta_4} = O(\frac{1}{k})$, and $\sigma_1\sigma_2\sigma_3\sigma_4\hat{\theta_3}\hat{\theta_1}\hat{\theta_4}\hat{\theta_2} = O(\frac{1}{k})$ as below:

  $$\hat{\epsilon}_\sigma = \sigma_1\sigma_4\hat{\theta_1}\hat{\alpha_4} + \sigma_2\sigma_3\hat{\theta_2}\hat{\alpha_3} + \sigma_1\sigma_3\hat{\theta_1}\hat{\theta_5}\hat{\alpha_3} + \sigma_2\sigma_4\hat{\theta_2}\hat{\theta_5}\hat{\alpha_4}$$

  with $\hat{\theta_1}\hat{\theta_5}\hat{\theta_2} = \tau$

B Calculation for Maximum Likelihood

$$
\begin{aligned}
L(\hat{g}_{\sigma^{(i)}}|x) &= -\ln\left\{\prod_{i=1}^{k}\hat{g}_{\sigma^{(i)}}\right\} \\
&= -\sum_{i=1}^{k}\ln\hat{g}_{\sigma^{(i)}} \\
&= -\sum_{\sigma\in S}\left[\sum_{i:\sigma^{(i)}=\sigma}\ln\hat{g}_{\sigma^{(i)}}\right] \\
&= -\sum_{\sigma\in S}\left[\sum_{i:\sigma^{(i)}=\sigma}\ln\hat{g}_{\sigma}\right] \\
&= -\sum_{\sigma\in S} n_\sigma \ln\hat{g}_\sigma \\
&= -\sum_{\sigma} k\hat{f}_\sigma \ln\hat{g}_\sigma \\
&\approx -\sum_{\sigma} k(f_\sigma+\frac{z_\sigma}{\sqrt{k}})\ln(f_\sigma+\frac{\hat{\epsilon}_\sigma}{\sqrt{k}}) \\
&= -\sum_{\sigma} k(f_\sigma+\frac{z_\sigma}{\sqrt{k}})\ln(f_\sigma+\frac{(\hat{\epsilon}_\sigma-z_\sigma)+z_\sigma}{\sqrt{k}}) \\
&= -\sum_{\sigma} k(f_\sigma+\frac{z_\sigma}{\sqrt{k}})\ln(f_\sigma+\frac{z_\sigma}{\sqrt{k}}+\frac{(\hat{\epsilon}_\sigma-z_\sigma)}{\sqrt{k}}).
\end{aligned}
$$

Let $f_\sigma+\frac{z_\sigma}{\sqrt{k}}=a$, then we can get:

$$
\begin{aligned}
&= -\sum_{\sigma} ka\ln(a+\frac{(\hat{\epsilon}_\sigma-z_\sigma)}{\sqrt{k}}) \\
&= -\sum_{\sigma} ka(\ln a+\ln(1+\frac{a^{-1}(\hat{\epsilon}_\sigma-z_\sigma)}{\sqrt{k}})) \\
&= -\sum_{\sigma} ka\ln a-\sum_{\sigma} ka\ln(1+\frac{a^{-1}(\hat{\epsilon}_\sigma-z_\sigma)}{\sqrt{k}}).
\end{aligned}
$$

By Taylor Expansion, we can expand $g(\hat{\epsilon}_\sigma)=\ln(1+\frac{a^{-1}(\hat{\epsilon}_\sigma-z_\sigma)}{\sqrt{k}})$ as:

$$
\ln(1+\frac{a^{-1}(\hat{\epsilon}_\sigma-z_\sigma)}{\sqrt{k}})=\frac{a^{-1}(\hat{\epsilon}_\sigma-z_\sigma)}{\sqrt{k}}-\frac{(a^{-1})^2(\hat{\epsilon}_\sigma-z_\sigma)^2}{2k}-\sum_{i=3}^{\infty}((-1)^i\frac{(a^{-1})^i(\hat{\epsilon}_\sigma-z_\sigma)^i}{i(\sqrt{k})^i}).
$$

Multiply both sides with ka, we have

$$
\begin{aligned}
ka\ln(1+\frac{a^{-1}(\hat{\epsilon}_\sigma-z_\sigma)}{\sqrt{k}}) &= ka\frac{a^{-1}(\hat{\epsilon}_\sigma-z_\sigma)}{\sqrt{k}}-ka\frac{(a^{-1})^2(\hat{\epsilon}_\sigma-z_\sigma)^2}{2k}-ka\sum_{i=3}^{\infty}((-1)^i\frac{(a^{-1})^i(\hat{\epsilon}_\sigma-z_\sigma)^i}{i(\sqrt{k})^i}) \\
&= \sqrt{k}(\hat{\epsilon}_\sigma-z_\sigma)-\frac{(a^{-1})(\hat{\epsilon}_\sigma-z_\sigma)^2}{2}-k\sum_{i=3}^{\infty}((-1)^i\frac{(a^{-1})^{i-1}(\hat{\epsilon}_\sigma-z_\sigma)^i}{i(\sqrt{k})^i}) \\
&= \sqrt{k}(\hat{\epsilon}_\sigma-z_\sigma)-\frac{(a^{-1})(\hat{\epsilon}_\sigma-z_\sigma)^2}{2}-O(\frac{1}{\sqrt{k}}).
\end{aligned}
$$

Now, plug the above back to the equation,

$$\begin{aligned}
-\sum_{\sigma} k(f_\sigma + \frac{z_\sigma}{\sqrt{k}}) \ln(f_\sigma + \frac{\hat{\epsilon}_\sigma}{\sqrt{k}}) &= -\sum_{\sigma} ka \ln a - \sum_{\sigma} ka \ln(1 + \frac{a^{-1}(\hat{\epsilon}_\sigma - z_\sigma)}{\sqrt{k}}) \\
&\approx -\sum_{\sigma} ka \ln a - \sum_{\sigma} \sqrt{k}(\hat{\epsilon}_\sigma - z_\sigma) + \sum_{\sigma} \frac{1}{2}(a^{-1})(\hat{\epsilon}_\sigma - z_\sigma)^2 + O(\frac{1}{\sqrt{k}}) \\
&= -\sum_{\sigma} ka \ln a - \sum_{\sigma} \sqrt{k}(\hat{\epsilon}_\sigma - z_\sigma) + \frac{1}{2}\sum_{\sigma}(a^{-1})(\hat{\epsilon}_\sigma - z_\sigma)^2 + O(\frac{1}{\sqrt{k}}).
\end{aligned}$$

By Central Limit Theorem, we have

$$\hat{\epsilon}_\sigma = \sqrt{k}(\hat{g}_\sigma - f_\sigma)$$

By definition, we know

$$\sum_{\sigma} \hat{g}_\sigma = 1 = \sum_{\sigma} f_\sigma$$

So, we have

$$\sum_{\sigma} \hat{\epsilon}_\sigma = \sum_{\sigma} \sqrt{k}(\hat{g}_\sigma - f_\sigma) = 0$$

So we can rewrite the previous equation as below:

$$L(\hat{g}_{\sigma^{(i)}}|x) = -\sum_{\sigma} ka \ln a + \sum_{\sigma} \sqrt{k} z_\sigma + \frac{1}{2}\sum_{\sigma}(a^{-1})(\hat{\epsilon}_\sigma - z_\sigma)^2 + O(\frac{1}{\sqrt{k}})$$

Now we are going to do the Taylor Expansion for $a = f_\sigma + \frac{z_\sigma}{\sqrt{k}}$
Then we have

$$\begin{aligned}
\frac{1}{a} &= \frac{1}{f_\sigma + \frac{z_\sigma}{\sqrt{k}}} \\
&= \frac{f_\sigma^{-1}}{1 + \frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}}}
\end{aligned}$$

As we know $\frac{1}{1-r} = 1 + r + r^2 + ...$, we can replace r with $-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}}$, then we have

$$\begin{aligned}
\frac{1}{a} &= f_\sigma^{-1}(\frac{1}{1 - (-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})}) \\
&= f_\sigma^{-1}(1 + (-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}}) + (-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^2 + (-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^3 + ...) \\
&= f_\sigma^{-1} \sum_{n=0}^{\infty} (-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^n \\
\Rightarrow a &= \frac{f_\sigma}{\sum_{n=0}^{\infty}(-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^n}
\end{aligned}$$

we then substitute a^{-1} into our previous equation to get

$$\begin{aligned}
L(\hat{g}_{\sigma^{(i)}}|x) &= -\sum_{\sigma} k \left(\frac{f_\sigma}{\sum_{n=0}^{\infty}(-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^n} \right) \ln \left(\frac{f_\sigma}{\sum_{n=0}^{\infty}(-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^n} \right) + \sum_{\sigma} \sqrt{k} z_\sigma \\
&\quad + \frac{1}{2} \sum_{\sigma} \left(f_\sigma^{-1} \sum_{n=0}^{\infty} \left(-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}} \right)^n \right) (\hat{\epsilon}_\sigma - z_\sigma)^2 + O(\frac{1}{\sqrt{k}}) \\
&= -\sum_{\sigma} k \left(\frac{f_\sigma}{\sum_{n=0}^{\infty}(-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^n} \right) \ln \left(\frac{f_\sigma}{\sum_{n=0}^{\infty}(-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^n} \right) + \sum_{\sigma} \sqrt{k} z_\sigma \\
&\quad + \frac{1}{2} \sum_{\sigma} \left(f_\sigma^{-1} \left(1 + \sum_{n=1}^{\infty} (-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^n \right) \right) (\hat{\epsilon}_\sigma - z_\sigma)^2 + O(\frac{1}{\sqrt{k}}) \\
&= -\sum_{\sigma} k \left(\frac{f_\sigma}{\sum_{n=0}^{\infty}(-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^n} \right) \ln \left(\frac{f_\sigma}{\sum_{n=0}^{\infty}(-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^n} \right) + \sum_{\sigma} \sqrt{k} z_\sigma \\
&\quad + \frac{1}{2} \sum_{\sigma} f_\sigma^{-1} (\hat{\epsilon}_\sigma - z_\sigma)^2 + \frac{1}{2} \sum_{\sigma} \left(f_\sigma^{-1} \sum_{n=1}^{\infty} (-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^n \right) (\hat{\epsilon}_\sigma - z_\sigma)^2 + O(\frac{1}{\sqrt{k}}) \\
&= -\sum_{\sigma} k \left(\frac{f_\sigma}{\sum_{n=0}^{\infty}(-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^n} \right) \ln \left(\frac{f_\sigma}{\sum_{n=0}^{\infty}(-\frac{z_\sigma f_\sigma^{-1}}{\sqrt{k}})^n} \right) + \sum_{\sigma} \sqrt{k} z_\sigma \\
&\quad + \frac{1}{2} \sum_{\sigma} f_\sigma^{-1} (\hat{\epsilon}_\sigma - z_\sigma)^2 + O(\frac{1}{\sqrt{k}})
\end{aligned}$$

Note that observe that the first two terms do not depend on $\hat{\epsilon}_\sigma$.

C Case-by-Case Least Square Optimization Problems

This section includes case-by-case setups for the least square optimization problems under constraints of $\theta_1\theta_2\theta_5 = \tau$ or $\theta_1\theta_2 = \tau$ for our implementation of the homotopy continuation method and nonlinear optimization package in Julia.

Least Square Optimization: Case 1. $\hat{\theta_1}\hat{\theta_5}\hat{\theta_2} = \tau$

$$\hat{\epsilon}_j = \sigma_1^{(j)}\sigma_3^{(j)}\hat{\theta_1}\hat{\alpha_3} + \sigma_2^{(j)}\sigma_4^{(j)}\hat{\theta_2}\hat{\alpha_4} + \sigma_1^{(j)}\sigma_4^{(j)}\hat{\theta_1}\hat{\theta_5}\hat{\alpha_4} + \sigma_2^{(j)}\sigma_3^{(j)}\hat{\theta_2}\hat{\theta_5}\hat{\alpha_3} \tag{5}$$

where $\hat{\theta_1}, \hat{\theta_2}, \hat{\alpha_3}, \hat{\alpha_4}$, and $\hat{\theta_5}$ are parameters of the tree. Recall that our case 1 assumes $\hat{\theta_1}\hat{\theta_5}\hat{\theta_2} = \tau$ is a known parameter. We are left with the optimization problem

$$\begin{aligned}
\min_{\hat{\theta_1},\hat{\theta_2},\hat{\alpha_3},\hat{\alpha_4},\hat{\theta_5}} \quad & \frac{1}{2}\sum_{j=1}^{16} q_j^{-1}(z_j - \hat{\epsilon}_j)^2 \\
\text{s.t.} \quad & \hat{\theta_1}\hat{\theta_5}\hat{\theta_2} = \tau \\
& 0 \le \hat{\theta_1} \le 1 \\
& 0 \le \hat{\theta_2} \le 1 \\
& 0 \le \hat{\alpha_3} \\
& 0 \le \hat{\alpha_4} \\
& 0 \le \hat{\theta_5} \le 1
\end{aligned}$$

Least Square Optimization: Case 2.1 $\hat{\theta_1}\hat{\theta_2} = \tau$

$$\hat{\epsilon}_j = \sigma_1^{(j)}\sigma_3^{(j)}\hat{\theta_1}\hat{\theta_5}\hat{\alpha_3} + \sigma_2^{(j)}\sigma_4^{(j)}\hat{\theta_2}\hat{\theta_5}\hat{\alpha_4} + \sigma_1^{(j)}\sigma_4^{(j)}\hat{\theta_1}\hat{\theta_5}\hat{\alpha_4} + \sigma_2^{(j)}\sigma_3^{(j)}\hat{\theta_2}\hat{\theta_5}\hat{\alpha_3}$$

where $\hat{\theta_1}, \hat{\theta_2}, \hat{\alpha_3}, \hat{\alpha_4}$, and $\hat{\theta_5}$ are parameters of the tree. Recall that our case 2 assumes $\hat{\theta_1}\hat{\theta_5}\hat{\theta_2} = \tau$ is a known parameter. We are left with the optimization problem

$$\begin{aligned} \min_{\hat{\theta_1},\hat{\theta_2},\hat{\alpha_3},\hat{\alpha_4},\hat{\theta_5}} \quad & \frac{1}{2}\sum_{j=1}^{16} q_j^{-1}(z_j - \hat{\epsilon}_j)^2 \\ \text{s.t.} \quad & \hat{\theta_1}\hat{\theta_2} = \tau \\ & 0 \le \hat{\theta_1} \le 1 \\ & 0 \le \hat{\theta_2} \le 1 \\ & 0 \le \hat{\alpha_3} \\ & 0 \le \hat{\alpha_4} \\ & 0 \le \hat{\theta_5} \le 1 \end{aligned}$$

Least Square Optimization: Case 2.2 $\hat{\theta_1}\hat{\theta_2} = \tau$

$$\hat{\epsilon}_j = \sigma_1^{(j)}\sigma_3^{(j)}\hat{\theta_1}\hat{\alpha_5}\hat{\alpha_3} + \sigma_2^{(j)}\sigma_4^{(j)}\hat{\theta_2}\hat{\alpha_5}\hat{\alpha_4} + \sigma_1^{(j)}\sigma_4^{(j)}\hat{\theta_1}\hat{\alpha_5}\hat{\alpha_4} + \sigma_2^{(j)}\sigma_3^{(j)}\hat{\theta_2}\hat{\alpha_5}\hat{\alpha_3} + \sigma_1^{(j)}\sigma_2^{(j)}\sigma_3^{(j)}\sigma_4^{(j)}\hat{\theta_2}\hat{\theta_1}\hat{\alpha_3}\hat{\alpha_4}$$

where $\hat{\theta_1}, \hat{\theta_2}, \hat{\alpha_3}, \hat{\alpha_4}$, and $\hat{\alpha_5}$ are parameters of the tree. Recall that our case 2 assumes $\hat{\theta_1}\hat{\alpha_5}\hat{\theta_2} = \tau$ is a known parameter. We are left with the optimization problem

$$\begin{aligned} \min_{\hat{\theta_1},\hat{\theta_2},\hat{\alpha_3},\hat{\alpha_4},\hat{\alpha_5}} \quad & \frac{1}{2}\sum_{j=1}^{16} q_j^{-1}(z_j - \hat{\epsilon}_j)^2 \\ \text{s.t.} \quad & \hat{\theta_1}\hat{\theta_2} = \tau \\ & 0 \le \hat{\theta_1} \le 1 \\ & 0 \le \hat{\theta_2} \le 1 \\ & 0 \le \hat{\alpha_3} \\ & 0 \le \hat{\alpha_4} \\ & 0 \le \hat{\alpha_5} \le 1 \end{aligned}$$

Least Square Optimization: Case 3. $\hat{\theta_1}\hat{\theta_5}\hat{\theta_2} = \tau$

$$\hat{\epsilon}_j = \sigma_1^{(j)}\sigma_4^{(j)}\hat{\theta_1}\hat{\alpha_4} + \sigma_2^{(j)}\sigma_3^{(j)}\hat{\theta_2}\hat{\alpha_3} + \sigma_1^{(j)}\sigma_3^{(j)}\hat{\theta_1}\hat{\theta_5}\hat{\alpha_3} + \sigma_2^{(j)}\sigma_4^{(j)}\hat{\theta_2}\hat{\theta_5}\hat{\alpha_4}$$

where $\hat{\theta_1}, \hat{\theta_2}, \hat{\alpha_3}, \hat{\alpha_4}$, and $\hat{\theta_5}$ are parameters of the tree. Recall that our case 3 assumes $\hat{\theta_1}\hat{\theta_5}\hat{\theta_2} = \tau$ is a known parameter. We are left with the optimization problem

$$\begin{aligned} \min_{\hat{\theta_1},\hat{\theta_2},\hat{\alpha_3},\hat{\alpha_4},\hat{\theta_5}} \quad & \frac{1}{2}\sum_{j=1}^{16} q_j^{-1}(z_j - \hat{\epsilon}_j)^2 \\ \text{s.t.} \quad & \hat{\theta_1}\hat{\theta_5}\hat{\theta_2} = \tau \\ & 0 \le \hat{\theta_1} \le 1 \\ & 0 \le \hat{\theta_2} \le 1 \\ & 0 \le \hat{\alpha_3} \\ & 0 \le \hat{\alpha_4} \\ & 0 \le \hat{\theta_5} \le 1 \end{aligned}$$

D Analytic Solution to Simple Boundary Case

Here we consider the case of the quartet 13|24 with $\theta_3 = 0$ and $\theta_1 = 1$, and $\theta_1\theta_2\theta_5 = \frac{1}{2}$. In this section we will compute the solution of the following **least squares minimization problem:**

$$\operatorname{argmin}\left\{\sum_\sigma \frac{1}{f_\sigma}\left(Z_\sigma - \hat{\epsilon}_\sigma\right)^2 : 0 < \theta_2, \theta_5 < 1 \text{ and } 0 < \alpha_4 < \sqrt{k}, \text{ and } \theta_2\theta_5 = \frac{1}{2}\right\}$$

where $\hat{\epsilon}_\sigma$ is defined by Eq. (5) as

$$\hat{\epsilon}_\sigma = \sigma_2\sigma_4\theta_2\alpha_4 + \sigma_1\sigma_4\theta_5\alpha_4. \tag{6}$$

and recalling Eq. (1), we have

$$f_\sigma = \frac{1}{16}\left(1 + \frac{1}{2}\sigma_1\sigma_2\right). \tag{7}$$

We emphasize that in this case we only consider solutions with $\theta_2, \theta_5 \in (0,1)$, and $\alpha_4 \in (0, \sqrt{k})$. We will show the following lemma:

Lemma 2 (Special Boundary Case)**.** *Given a tree with quartet topology* 13|24 *and constraints* $\theta_3 = 0$, $\theta_1 = 1$, *and* $\theta_1\theta_2\theta_5 = \frac{1}{2}$, *the solution to the least squares minimization problem must satisfy*

$$\theta_2 = \sqrt{\frac{B_2 - B_1}{2(B_1 + B_2)}} \quad \textit{and} \quad \alpha_4 = \frac{B_2 + 3B_1}{16\left(\left|\frac{B_2 - B_1}{2(B_1+B_2)}\right| - 1\right)}$$

where

$$\begin{aligned} B_1 &= z_{-+-+} + z_{-+++} + z_{+---} + z_{+-+-} - z_{-+--} - z_{-++-} - z_{+--+} - z_{+-++} \\ B_2 &= z_{---+} + z_{--++} + z_{++--} + z_{+++-} - z_{----} - z_{--+-} - z_{++-+} - z_{++++}. \end{aligned}$$

for $z = (z_{----}, \ldots, z_{++++})$ *as defined in* **??**.

Remark 1. *It is clear from the statement of Lemma 2 that there is at most one positive real solution, and possibly none. The necessary and sufficient conditions for a solution to exist are* $\frac{B_2 - B_1}{2(B_1+B_2)} \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$ *and* $B_2 + 3B_1 < 0$. *If there is no solution, then the least squares objective function is minimized* ***on the boundary;*** *that is, when* $\theta_2 \in \left\{\frac{1}{2}, 1\right\}$ *or* $\alpha_4 \in \left\{0, \sqrt{k}\right\}$. *[Note:* $\theta_2 = 1/2$ *is a boundary case because the other constraints then imply that* $\theta_5 = 1$.*] These results are consistent with the results obtained using homotopy—as in those cases, we found either 1 solution or a solution on the boundary. Also, in the proof, we well see that there are two critical points (i.e. corresponding to* $\theta = \pm\sqrt{\frac{B_2 - B_1}{2(B_1+B_2)}}$*), which based on my memory is also consistent with the sorts of critical points we obtained from homotopy.*

Proof of Lemma 2. By Eqs. (6) and (7), we have

$$\begin{aligned} \sum_\sigma \frac{1}{f_\sigma}\left(Z_\sigma - \hat{\epsilon}_\sigma\right)^2 = &\sum_{\sigma:\sigma_1\sigma_2=+1} \frac{32}{3}\left(Z_\sigma - \sigma_2\sigma_4\theta_2\alpha_4 - \sigma_1\sigma_4\theta_5\alpha_4\right)^2 \\ &+ \sum_{\sigma:\sigma_1\sigma_2=-1} 32\left(Z_\sigma - \sigma_2\sigma_4\theta_2\alpha_4 - \sigma_1\sigma_4\theta_5\alpha_4\right)^2. \end{aligned}$$

Next we make the substitution $\theta_5 = \frac{1}{2\theta_2}$ in order to eliminate the variable θ_5. For convenience, we will also write $x = \theta_2$ and $y = \alpha_4$. Making these substitutions gives:

$$\begin{aligned} \sum_\sigma \frac{1}{f_\sigma}\left(Z_\sigma - \hat{\epsilon}_\sigma\right)^2 &= \frac{32}{3}\sum_{\sigma:\sigma_1\sigma_2=+1}\left(Z_\sigma - \sigma_2\sigma_4 xy - \frac{\sigma_1\sigma_4 y}{2x}\right)^2 \\ &\quad + 32\sum_{\sigma:\sigma_1\sigma_2=-1}\left(Z_\sigma - \sigma_2\sigma_4 xy - \frac{\sigma_1\sigma_4 y}{2x}\right)^2 \\ &=: M(x,y). \end{aligned}$$

Next, we will assume that M has a critical point on the interior of its domain, i.e. the set

$$\left\{(x,y) \in \mathbb{R}^2 : 0 < x < 1, 0 < y < \sqrt{k}\right\},$$

and the following calculations will ascertain what the critical point(s) in that case must be. Assuming $x, y \neq 0$, the critical points of M are exactly the solutions of the critical equations

$$\begin{cases} M_y(x,y) = 0 \\ M_x(x,y) = 0 \end{cases} \tag{8}$$

which has the same solutions as the system

$$\begin{cases} \frac{3x^2}{32} M_y(x,y) = 0 \\ \frac{3x^3}{32y} M_x(x,y) = 0 \end{cases} \tag{9}$$

Using the following Julia code, we obtain expressions for Eq. (9):

```
# Initialization:
using HomotopyContinuation
@var x y z[1:16]

# Make an ordered list s of site patterns (-1,-1,-1,-1),(-1,-1,-1,+1),
# (-1,-1,+1,-1), (-1,-1,+1,+1),...,(+1,+1,+1,+1):
index_set=Iterators.product(fill([1;-1],4)...)|>collect # Define 2x2x2x2 array
s=reverse(map(i->reverse(index_set[i]),1:16))

# Define least squares cost function:
f = [(1/16) * (1 + (1/2) * s[i][1] * s[i][2]) for i in 1:16]
X(i) = (1/f[i])*(z[i] - s[i][2]*s[i][4]*x*y - s[i][1]*s[i][4]*y/(2*x))^2
LS = sum([X(i) for i in 1:16])

# Obtain the critical equations:
M_x = differentiate(LS, x)
M_y = differentiate(LS, y)

# Expand the equations:
julia> HomotopyContinuation.expand((3/32)*x^2*M_y)

    16.0*y - 1.0*x*z₁ - 3.0*x*z₁₀ + 3.0*x*z₁₁ - 3.0*x*z₁₂ + 1.0*x*z₁₃ - 1.0*x*z₁₄ +
    1.0*x*z₁₅ - 1.0*x*z₁₆ + 1.0*x*z₂ - 1.0*x*z₃ + 1.0*x*z₄ - 3.0*x*z₅ + 3.0*x*z₆ -
    3.0*x*z₇ + 3.0*x*z₈ + 3.0*x*z₉ - 32.0*x^2*y - 2.0*x^3*z₁ + 6.0*x^3*z₁₀ -
    6.0*x^3*z₁₁ + 6.0*x^3*z₁₂ + 2.0*x^3*z₁₃ - 2.0*x^3*z₁₄ + 2.0*x^3*z₁₅ -
    2.0*x^3*z₁₆ + 2.0*x^3*z₂ - 2.0*x^3*z₃ + 2.0*x^3*z₄ + 6.0*x^3*z₅ - 6.0*x^3*z₆ +
    6.0*x^3*z₇ - 6.0*x^3*z₈ - 6.0*x^3*z₉ + 64.0*x^4*y

julia> HomotopyContinuation.expand((3/32)*x^3*(1/y)*M_x)

    -16.0*y + 1.0*x*z₁ + 3.0*x*z₁₀ - 3.0*x*z₁₁ + 3.0*x*z₁₂ - 1.0*x*z₁₃ + 1.0*x*z₁₄ -
    1.0*x*z₁₅ + 1.0*x*z₁₆ - 1.0*x*z₂ + 1.0*x*z₃ - 1.0*x*z₄ + 3.0*x*z₅ - 3.0*x*z₆ +
    3.0*x*z₇ - 3.0*x*z₈ - 3.0*x*z₉ - 2.0*x^3*z₁ + 6.0*x^3*z₁₀ - 6.0*x^3*z₁₁ +
    6.0*x^3*z₁₂ + 2.0*x^3*z₁₃ - 2.0*x^3*z₁₄ + 2.0*x^3*z₁₅ - 2.0*x^3*z₁₆ +
    2.0*x^3*z₂ - 2.0*x^3*z₃ + 2.0*x^3*z₄ + 6.0*x^3*z₅ - 6.0*x^3*z₆ + 6.0*x^3*z₇ -
    6.0*x^3*z₈ - 6.0*x^3*z₉ + 64.0*x^4*y
```

Regrouping terms in the results obtained from the Julia code, we obtain the following two equations:

$$\frac{3x^2}{32}M_y(x,y) = \left(64x^4 - 32x^2 + 16\right)y + (B_2 + 3B_1)\,x + (2B_2 - 6B_1)\,x^3 \tag{10}$$

$$\frac{3x^3}{32y}M_x(x,y) = (64x^4 - 16)y - (B_2 + 3B_1)x + (2B_2 - 6B_1)x^3 \tag{11}$$

where

$$B_1 = z_6 + z_8 + z_9 + z_{11} - z_5 - z_7 - z_{10} - z_{12}$$
$$B_2 = z_2 + z_4 + z_{13} + z_{15} - z_1 - z_3 - z_{14} - z_{16}$$

(Note that since $(z_1, z_2, \ldots, z_{16}) = (z_{----}, z_{---+}, \ldots, z_{+++})$, it is easily checked that B_1 and B_2 can be written in the forms given in the statement of this lemma.) By Eqs. (10) and (11), any critical points to M lying in the interior of its domain must be solutions to the following critical equations:

$$\left(64x^4 - 32x^2 + 16\right)y + (B_2 + 3B_1)\,x + (2B_2 - 6B_1)\,x^3 = 0 \tag{12}$$

$$(64x^4 - 16)y - (B_2 + 3B_1)x + (2B_2 - 6B_1)x^3 = 0 \tag{13}$$

Subtracting Eq. (13) from Eq. (12) gives

$$32(1 - x^2)y + 2(B_2 + 3B_1)x = 0,$$

and since $0 < x < 1$, this implies

$$y = \frac{B_2 + 3B_1}{16(x^2 - 1)}. \tag{14}$$

Adding Eq. (12) and Eq. (13) gives

$$32(4x^4 - x^2)y + 4(B_2 - 3B_1)x^3 = 0.$$

Dividing both sides by $2x^2$ gives

$$16(4x^2 - 1)y + 2(B_2 - 3B_1)x = 0.$$

Plugging in the formula for y from Eq. (14) gives

$$\frac{(4x^2 - 1)(B_2 + 3B_1)}{(x^2 - 1)} + 2(B_2 - 3B_1)x = 0.$$

Multiplying both sides by $x^2 - 1$, and then rearranging terms gives

$$3(B_2 - B_1) - 6(B_1 + B_2)x^2 = 0$$

so that

$$x = \pm\sqrt{\frac{B_2 - B_1}{2(B_1 + B_2)}}$$

and hence, by Eq. (14),

$$y = \frac{B_2 + 3B_1}{16\left(\left|\frac{B_2 - B_1}{2(B_1 + B_2)}\right| - 1\right)}.$$

□

E Julia Nonlinear Code

```
obs1 = [-1 -1 -1 -1 -1 -1 -1 -1 +1 +1 +1 +1 +1 +1 +1 +1;
        -1 -1 -1 -1 +1 +1 +1 +1 -1 -1 -1 -1 +1 +1 +1 +1;
        -1 -1 +1 +1 -1 -1 +1 +1 -1 -1 +1 +1 -1 -1 +1 +1;
        -1 +1 -1 +1 -1 +1 -1 +1 -1 +1 -1 +1 -1 +1 -1 +1]
#represents the σ

k = 1000000000

# the next way is the way that we generate data based on customized τ
tau = 0.5
# τ is set to 0.5 in the representation but you can change the τ in the start of the
# generated part and get different configuration based on different τ
q = zeros(16)
for j = 1:16
    q[j] = (1 / 16) * (1 + obs1[1, j] * obs1[2, j] * tau)
end
# q will be the true probability of fσ

mean = k * q
# mean is the mean of the multivariable normal distribution given the k

meandia = diagm(q)
b = transpose(q)
c = q * b
d1 = meandia -c
cov = k * d1
# coveriance of the distribution

d = MvNormal(mean, cov)
# generate the Multivariate normal distribution from the mean and covariance
test123 = rand(d)
# test123 will be the g(x)

freqtest = (test123-mean)/ sqrt(k)
# freqtest is the ε since ε is the (g(x) - f(x))/ sqrt(k)

# the next one will be a basic of telling about
# how will the model work on case 1 and the most basic problem without any
# boundary cases.

model = Model(Ipopt.Optimizer)
@variable(model,x1>=0)
# x1 is the θ1
@variable(model,x2>=0)
# x2 is the θ2
@variable(model,x3>=0)
# x3 is the α3
@variable(model,x4>=0)
# x4 is the α4
@variable(model,x5>=0)
# x5 is the θ5
#set each variable to be larger than 0
```

```
@constraint(model,x1<=1)
# set θ1 to be less or equal to 1
@constraint(model,x2<=1)
# set θ2 to be less or equal to 1
@constraint(model,x3<=sqrt(k))
# set α3 to be less or equal to sqrt(k)
@constraint(model,x4<=sqrt(k))
# set α4 to be less or equal to sqrt(k)
@constraint(model,x5<=1)
# set θ5 to be less or equal to 1
@NLconstraint(model,x1*x2*x5 == tau)
# set θ1*θ2*θ5 equal to 1 as a nonlinear constrain
@NLobjective(model, Min, 1/2*sum(((freqtest[i]-obs1[1,i]*obs1[3,i]*x1*x3
                -obs1[2,i]*obs1[4,i]*x2*x4
                -obs1[1,i]*obs1[4,i]*x1*x5*x4
                -obs1[2,i]*obs1[3,i]*x2*x5*x3)^2)/q[i] for i in 1:16))
                # set the objective function
                # the configuration of the objective function is the same as the
                # Case 1 on page 10 with εj directedly replaced by θ1,θ2,θ5,α3,α4
optimize!(model)
# make the optimization packages run
obj =objective_value(model)
# obj will be the objective function's value after the optimization
# and it will also be returned by the function

# here are representations of how to handle the boundary cases
# for the special case that θ2,θ5 equal to 1 and α4 equal to 0
# I directly replaced the variable with the boundary value that we assigned to
# that variable and keeps other variables unchanged.


model = Model(Ipopt.Optimizer)
@variable(model,x1>=0)
# x1 is the θ1
@variable(model,x4>=0)
# x4 is the α4
#set each variable to be larger than 0
@constraint(model,x1<=1)
# θ1 less or equal to  1
@constraint(model,x4<=sqrt(k))
# α4 less or equal to  sqrt(k)
@NLconstraint(model,x1 == tau)
# θ1 equal to 1 since the θ2,θ5 are fixed
@NLobjective(model, Min, 1/2*sum(((freqtest[i]-obs1[1,i]*obs1[3,i]*0
                -obs1[2,i]*obs1[4,i]*x4
                -obs1[1,i]*obs1[4,i]*x1*x4
                -obs1[2,i]*obs1[3,i]*0)^2)/q[i] for i in 1:16))
                # set α3 equal to 0
                # set θ2,θ5 equal to 1
optimize!(model)
obj =objective_value(model)
# return the objected value
```

F Paragraph for Website

A phylogenetic tree is a diagram that depicts the lines of evolutionary descent of different species that share a common ancestor. The goal of phylogeny estimation using maximum-likelihood is to recover the topology of a phylogenetic tree from datasets of genes for some set of species or other taxa. Long branch attraction is a systemic error incurred when distantly related, i.e. long-branched lineages, instead appear to be closely related. In this project, we investigated whether long branch attraction introduces bias to maximum-likelihood estimation in phylogenetic tree reconstruction even for simple evolutionary models and small-scale trees. Specifically, we translated the maximum-likelihood problem into a least-squares problem, with the aim of computing the optimality conditions (i.e. the tree parameters that give the maximum likelihood) of all possible configurations of the unrooted four-species tree with two long branches through a combination of analytical solutions and minimization software. After implementing a Homotopy Continuation approach and a Nonlinear Optimization approach in Julia, we found that the results from the two methods agree in certain cases, but questions remain.

References

[1] Paul Breiding and Sascha Timme. Homotopycontinuation. jl: A package for homotopy continuation in julia. In *International Congress on Mathematical Software*, pages 458–465. Springer, 2018.

[2] Joseph Cummings, Benjamin Hollering, and Christopher Manon. Invariants for level-1 phylogenetic networks under the cavendar-farris-neyman model, 2021.

[3] C. Semple, M. Steel, and B.D.M.S.M. Steel. *Phylogenetics*. Oxford Lecture Mathematics and. Oxford University Press, 2003.