1 Supplemental Documentation

This documentation is provided in the hopes that it will be useful for understanding the output of our estimation functions, which are somewhat unusual in that they are designed to return maximally informative output even in cases where the MLE would not traditionally be said to exist.

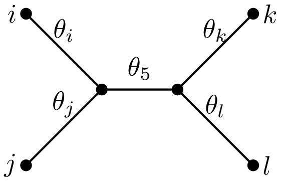
If you have quesetions or concerns about any of this, please feel free to reach out to me.

–Max Hill

(Last updated: June 8, 2024)

1.1 The parameters

A 4-leaf binary tree has five Hadamard edge parameters $\theta_1, \theta_2, \ldots, \theta_5 \in [0,1]$, as shown:



Here, i, j, k, l are pairwise distinct elements of $\{1,2,3,4\}$. Recall that for each edge e, the Hadamard parameter for e is defined as

$$\theta_e := e^{-2d_e}$$

where d_e is the branch length of edge e measured in expected number of mutations per site. Thus, a Hadamard edge parameter of 0 corresponds to an infinitely long branch; an edge parameter of 1 corresponds to a branch length of zero. The above tree has binary tree toplogy $ij|kl$, which depending on the assignment of leaf labels may be 12|34, 13|24, or 14|23. Within the code itself, these three topologies are always represented by the variable τ, which takes values $1, 2$ or 3 respectively.

1.2 Interior vs boundary cases

If the maximum likelihood estimate is a binary tree such that $\theta_1, \ldots, \theta_5 \in (0,1)$, then the MLE is said to occur *in the interior* of the parameter space. On the other hand, the maximum likelihood estimate is said to *occur on the boundary* if one or more of the estimated branch lengths is 0 or 1. `FourLeafMLE.jl` computes the global MLE by maximizing the likelihood of the data over all trees in the interior of the parameter space as well as over all possible boundary cases.

To simplify the analysis, our estimation software requires that the data be *generic*, which includes the requirement that each possible site pattern $(xxxx, xxxy, xxyx, xxyy, xyxx, xyxy, xyyx, xyyy)$ be observed at least once. For data satisfying this requirement, the analyses of many boundary cases becomes trivial, as they can immediately be seen to have likelihood zero. For example, if the data exhibits any site pattern other than $aaaa$, then the boundary case in which $\theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 1$ always has likelihood zero. This is the simplest case, but there are many other cases like this as well.

After excluding trivial cases, we organize the remaining nontrival cases into 10 categories, denoted

$$R_1, R_2, \ldots, R_{10}.$$

A complete description of these classes is provided in the section below.

1.3 Output of the functions `fourLeafMLE()` and `listMaxima()`

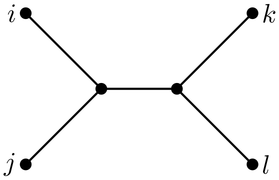
The maximum likelihood estimation functions in `FourLeafMLE.jl` will always identify which of these classes $R_1, \ldots, R_{10}$ the MLE belongs to, along with an appropriate number of edge parameters. If the MLE is a binary 4-leaf tree with positive and finite branch lengths, then the number of edge parameters will be 5, corresponding to the 4 leaves $(\theta_1, \ldots, \theta_4)$ and 1 internal branch (θ_5). In other cases, the number of edge parameters may be fewer, since some of the classes in $R_1, \ldots, R_{10}$ have fewer than 5 edges. The reduction in the number of edges is necessary in cases where not all five edge parameters $\theta_1, \dots, \theta_5$ are independently identifiable.

More information about the output is provided in the doc string for `fourLeafMLE()`, which can be shown by running

```
julia> @doc fourLeafMLE
```

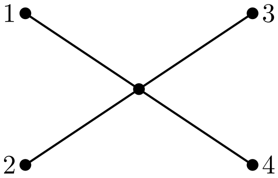
1.4 Description of $R_1, R_2, \dots, R_{10}$

(R$_1$) Binary quartets:



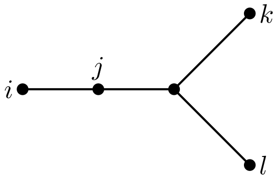
There are three distinct elements of R_1, corresponding to the three unrooted binary 4-leaf tree topologies 12|34, 13|23, and 14|23. All edge parameters $\theta_1, \dots, \theta_5 \in (0, 1)$.

(R$_2$) Star tree:



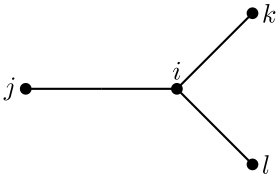
R_2 has only one element, the 4-leaf star toplogy. The R_2 category corresponds to the case where the tree's internal edge parameter $\theta_5 = 1$ and the parameters for the leaf edges are $\theta_1, \theta_2, \theta_3, \theta_4 \in (0, 1)$. If the MLE takes this form, then output of `FourLeafMLE.jl` will return 4 edge paramters, corresponding to the four edges of the star tree.

(R$_3$) Y-shaped tree:



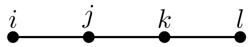
This boundary case corresponds to the situation where $\theta_j = 1$ and all other edge parameters are in $(0, 1)$. There are 12 distinct boundary cases with this reduced topology. If the MLE takes the form of an R_3 tree, then `FourLeafMLE.jl` returns

(R$_4$) Claw:



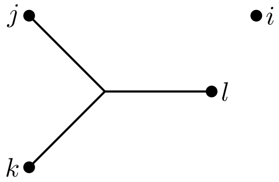
A 4-leaf tree belongs to class R_4 if and only if there is some choice of pairwise distinct $i, j, k, l \in [4]$ such that $\theta_i = \theta_5 = 1$ and $\theta_j \in (0,1)$ for all $j \in [4]\backslash\{i,5\}$. There are 4 distinct cases with reduced topology R_4.

(R_5) Line:



R_5 corresponds to the case in which there is some choice of pairwise distinct $i, j, k, l \in [4]$, such that T has topology $ij|kl$ with $\theta_j, \theta_k = 1$ and $\theta_5, \theta_i, \theta_l \in (0,1)$. Note that R_5 has 12 distinct cases.

(R_6) Infinite $\{i\}$-branch model:

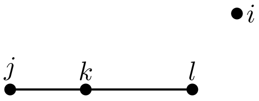


R_6 consists of four elements, depending on whether $i = 1, 2, 3$ or 4. A tree T corresponds to case R_6 if and only if there exists a $i \in [4]$ such that $\theta_i = 0$ and

$$\left(\prod_{e \in \mathcal{P}(T,A)} \theta_e \right) \in (0,1)$$

for all $A \subseteq [4]\backslash\{i\}$ with $|A|= 2$, where $\mathcal{P}(T, A)$ is the path on T connecting the two vertices in A.

(R_7) Degenerate infinite $\{i\}$-branch model:



A tree T corresponds to R_7 if and only if one of the following conditions holds:

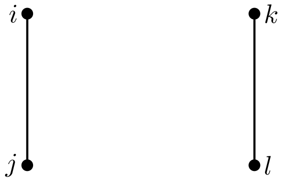
(a) T has topology $ij|kl$, with $\theta_i = 0$, $\theta_k = 1$, $\theta_j\theta_k \in (0,1)$ and $\theta_l \in (0,1)$.

(b) T has topology $ik|jl$, with $\theta_i = 0$, $\theta_k = \theta_5 = 1$, and $\theta_j, \theta_l \in (0,1)$.

(c) T has topology $il|jk$, with $\theta_i = 0$, $\theta_k = 1$, $\theta_j \in (0,1)$, and $\theta_l\theta_5 \in (0,1)$.

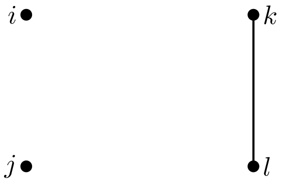
There are 12 distinct assignments of labels $i, j, k, l \in \{1, 2, 3, 4\}$ corresponding to R_7.

(R_8) Infinite $\{i, j\}$-branch model



A tree has reduced topology R_8 if and only if for some choice of pairwise distinct $i, j, k, l \in [4]$, T has topology $ij|kl$ with $\theta_5 = 0$ and $\theta_i\theta_j, \theta_k\theta_l \in (0, 1)$. There are three distinct cases for R_8: $(i, j, k, l) = (1, 2, 3, 4), (1, 3, 2, 4)$, or $(1, 4, 2, 3)$.

(R$_9$) Infinite $\{i\}, \{j\}$-branch model



R_9 consists of 6 elements. A tree T exhibits reduced topology R_9 if and only if there exists a pair i, j such that for all $A \subseteq [4]$ with $|A| = 2$, the following condition holds:

$$\prod_{e\in\mathcal{P}(T,A)} \theta_e = 0 \text{ if and only if } \{i, j\} \cap A \neq \emptyset.$$

(R$_{10}$) Full independence model:

A tree correspond to R_{10} if and only if $\prod_{e\in\mathcal{P}(T,A)} \theta_e = 0$ for all $A \subseteq [4]$. In this case, the observations at all four leaves are independent.