Joint subtree distributions under two evolutionary tree models

Symposium in Memory of Charles Stein [1920 - 2016]

June 19, 2019

Kwok-Pui Choi

Joint work with
Taoyang Wu
OUTLINE

1. INTRODUCTION

2. JOINT SUBTREE DISTRIBUTIONS UNDER TWO EVOLUTIONARY MODELS

3. DISCRIMINATION EFFICIENCY

4. WORK IN PROGRESS
1. Introduction

- Trees provide visual representation of the evolutionary relatedness among organisms, species, or genes.
- Phylogenetic tree has become a powerful tool in studying evolution and diversification in Biology.
- It finds wide applications in evolutionary biology, developmental biology (cell lineages) and epidemiology. Such as
  - Inferring evolutionary process from phylogenetic tree shape\(^1\);
  - Studying diversification\(^2\);
  - Study of pathogens\(^3,3a\).
- Tree shape is the signature of the forces that produce biodiversity\(^4\)
- Fitting stochastic models to tree data helps infer macro-evolutionary processes such as speciation and extinction rates.

---

1. Introduction

- $T$ is a phylogenetic tree on $\mathcal{X} = \{1, \ldots, 5\}$, the set of leaves.
- All edges are directed away from the root, $a$ (arrows not drawn).
- Binary tree $(V(T), E(T))$, set of pendant edges denoted by $E^*(T)$.
- Each node is either of degree 1 (for leaves); degree 2 (root); or degree 3 (the rest of internal nodes).
Our objectives

- Given a tree $T$, to infer if this tree is more likely generated by YHK model than by PDA model.

- Given a “tree feature” (index), to assess how good this feature in discriminating one generative model from another.
  
  - YHK trees are observed to be more symmetric. Many tree indices have been proposed to measure tree balance.
  
  - Examples: Colless index and Sackin index
1.2 Two random generative tree models: YHK & PDA

Two common generative tree models: YHK (Yule-Harding-Kingsman) model and PDA (Proportional-to-Distinguishable-Arrangements) model

- YHK: Choose a pendant edge at random in $T$ to form $T'$.
- PDA: Choose an edge at random from all edges in $T$ to form $T'$.
- In this figure, the PDA model: $T' = T[e_9; x_7]$ is obtained from $T$ by attaching the leaf labelled $x_7$ to edge $e_9$. 
1.2 Two generative tree models: YHK & PDA

Starting with a tree of 2 leaves, iteratively attach (manner to be specified) one leaf at a time to grow the tree to its desired size.

- YHK, also known as equal-rates-Markov model: all extant lineages have equal probabilities of speciating at an instant.
  - Uniformly sample a pendant edge from the tree’s pendant edges, and attach a new leaf to it.

- PDA: Each “possible arrangement” of $n$ leaves into a tree is equally likely.
  - Uniformly sample an edge from the present tree’s edges, and attach a new leaf to it.
2.1 Subtree

A cherry—a subtree with two leaves.
For examples,
\(\{x_1, x_5\}\) (together with their parent) forms a cherry, and we think of three edges: \(e_7, e_1, e_5\)
\(\{x_2, x_4\}\) forms another cherry with edges: \(e_9, e_2, e_4\).

A pitchfork—a subtree with three leaves.
E.g., \(\{x_1, x_5, x_3\}\) with their most recent common ancestor; 5-edge:
\(e_8, e_7, e_3, e_1, e_5\).
2.2 Subtrees of sizes 2 and 3 (Cherry and Pitchfork)

- It has been observed that YHK generated trees are more “symmetric”: and hence contain more cherries.

**Definitions**

- $A_n$: Number of pitchforks in a tree with $n$ leaves
- $C_n$: Number of cherries in a tree with $n$ leaves
- Let $h_n$ denote the joint pmf of $A_n$ and $C_n$ under the YHK model. That is,

$$h_n(a, c) = P_Y[A_n = a, C_n = c].$$

- Let $f_n$ be the pmf of $A_n$; and $g_n$ pmf of $C_n$ under the YHK model.
- Denote $\tilde{h}_n, \tilde{f}_n, \tilde{g}_n$ the counterparts under the PDA model.
**Thm 1** We have the following recurrence relations.

(a) YHK Model

For $n \geq 3$,

$$
\begin{align*}
    h_{n+1}(a,c) &= \frac{2a}{n} h_n(a,c) + \frac{a+1}{n} h_n(a+1,c-1) \\
    &\quad + \frac{2(c-a+1)}{n} h_n(a-1,c) + \frac{n-a-2c+2}{n} h_n(a,c-1);
\end{align*}
$$

and $h_3(1,1) = 1$, and $h_3(a,c) = 0$ if $(a,c) \neq (1,1)$.

(b) PDA Model

For $n \geq 3$,

$$
\begin{align*}
    \tilde{h}_{n+1}(a,c) &= \frac{n+3a-c-1}{2n-1} \tilde{h}_n(a,c) + \frac{a+1}{2n-1} \tilde{h}_n(a+1,c-1) \\
    &\quad + \frac{3(c-a+1)}{2n-1} \tilde{h}_n(a-1,c) + \frac{n-a-2c+2}{2n-1} \tilde{h}_n(a,c-1);
\end{align*}
$$

and $\tilde{h}_3(1,1) = 1$, and $\tilde{h}_3(a,c) = 0$ if $(a,c) \neq (1,1)$. 

\[ \text{p10} \]
Key step in the proof of Thm 1
Edge decomposition for a PDA tree: Partition $E(T)$ into
$E_1(T)$: pendant edges that are contained in a pitchfork but not in a cherry;
$E_2(T)$: edges that are contained in a cherry but not in a pitchfork;
$E_3(T)$: pendant edges that are contained in neither a pitchfork nor a cherry;
$E_4(T) = E(T) \setminus (E_1(T) \cup E_2(T) \cup E_3(T))$.

\[ E_1(T) = \{e_3\}, \quad E_2(T) = \{e_2, e_4, e_9\}, \quad E_3(T) = \{e_6\}, \]
\[ E_4(T) = \{e_0, e_1, e_5, e_7, e_8, e_{10}\}. \]
**Key step in the proof of Thm 1**

**Edge decomposition for a PDA tree**

- Track the changes in the numbers of pitchforks and cherries when a new leaf joins the tree $T$ at $E_i(T)$:

\[
(A(T'), C(T')) = \begin{cases} 
(A(T) - 1, C(T) + 1), & e \in E_1(T), \\
(A(T) + 1, C(T)), & e \in E_2(T), \\
(A(T), C(T) + 1), & e \in E_3(T), \\
(A(T), C(T)), & e \in E_4(T),
\end{cases}
\]

and

\[
\begin{align*}
|E_1(T)| &= A(T), \\
|E_2(T)| &= 3(A(T) - C(T)), \\
|E_3(T)| &= n - A(T) - 2C(T), \\
|E_4(T)| &= n - 1 + 3A(T) - C(T).
\end{align*}
\]
Thm 2  Denote expectation under YHK (resp., PDA) model by $E_Y$ (resp., $E_U$). Let $\phi : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$. Then, for $n \geq 4$,

$$nE_Y[\phi(A_{n+1}, C_{n+1})] = 2E_Y[A_n\phi(A_n, C_n)]$$

$$+ E_Y[A_n\phi(A_n - 1, C_n + 1)]$$

$$+ 2E_Y[(C_n - A_n)\phi(A_n + 1, C_n)]$$

$$+ E_Y[(n - A_n - 2C_n)\phi(A_n, C_n + 1)];$$

$$(2n - 1)E_U[\phi(A_{n+1}, C_{n+1})] = E_U[(n + 3A_n - C_n - 1)\phi(A_n, C_n)]$$

$$+ E_U[A_n\phi(A_n - 1, C_n + 1)]$$

$$+ 3E_U[(C_n - A_n)\phi(A_n + 1, C_n)]$$

$$+ E_U[(n - A_n - 2C_n)\phi(A_n, C_n + 1)].$$
Special cases

(1) Let $\psi : \mathbb{R} \rightarrow \mathbb{R}$ be any function.

(a) Define $\phi(x, y) = \psi(y)$, we have

$$
\begin{align*}
    n \mathbb{E}_Y[\psi(C_{n+1})] &= \mathbb{E}_Y[2C_n \psi(C_n) + (n - 2C_n) \psi(C_{n+1})]; \\
    (2n-1) \mathbb{E}_U[\psi(C_{n+1})] &= \mathbb{E}_U[(n + 2C_n - 1) \psi(C_n) \\
    &\quad + (n - 2C_n) \psi(C_{n+1})].
\end{align*}
$$

(b) Define $\phi(x, y) = I_k(y)$, we have

$$
\begin{align*}
    ng_{n+1}(k) &= 2k g_n(k) + (n - 2k + 2) g_n(k - 1); \\
    (2n-1) \tilde{g}_{n+1}(k) &= (n + 2k - 1) \tilde{g}_n(k) + (n - 2k + 2) \tilde{g}_n(k - 1),
\end{align*}
$$

for $n \geq 3$ and $k \geq 1$.

(c) Take $\phi(x, y) = y$ to derive $\mathbb{E}_Y[C_n]$ and $\mathbb{E}_U[C_n]$;

(d) $\phi(x, y) = y^2$ to derive $\text{Var}_Y[C_n]$ and $\text{Var}_U[C_n]$. 
On the average

- more cherries with less variation in a YHK tree than in a PDA tree:
  \[
  \mathbb{E}_Y[C_n] = \frac{n}{3} > \mathbb{E}_U[C_n] = \frac{n(n-1)}{2(2n-3)} \sim \frac{n}{4};
  \]
  \[
  \text{Var}_Y[C_n] = \frac{2n}{45} < \text{Var}_U[C_n] = \frac{n(n-1)(n-2)(n-3)}{2(2n-3)^2(2n-5)} \sim \frac{n}{16};
  \]

- more pitchforks but also with more variation in a YHK tree than in a PDA tree:
  \[
  \mathbb{E}_Y[A_n] = \frac{n}{6} > \mathbb{E}_U[A_n] = \frac{n(n-1)(n-2)}{2(2n-3)(2n-5)} \sim \frac{n}{8};
  \]
  \[
  \text{Var}_Y[A_n] = \frac{23n}{420} > \text{Var}_U[A_n] = \frac{3n(n-1)(n-2)(n-3)(4n^3-40n^2+123n-110)}{4(2n-3)^2(2n-5)^2(2n-7)(2n-9)} \sim \frac{3n}{64};
  \]

- Correlation of numbers of cherries and pitchforks is constant (independent of \( n \)) under YHK; whereas correlation = \( O(1/n) \) under PDA:
  \[
  \text{Cor}_Y(A_n, C_n) = -\sqrt{\frac{14}{69}}; \quad \text{Cor}_U(A_n, C_n) \sim \frac{-1}{\sqrt{3}n}.
  \]
We examine the likelihood ratio of the number of cherries under YHK versus PDA: $f_n(k)/\tilde{f}_n(k)$. It is monotone in $k$ for fixed $n$. 
3.1 Model-based total variation distance between YHK and PDA

Definition We define a total variation distance between YHK model and PDA model on the collection of trees with \( n \) leaves as

\[
mtv_n(YHK, PDA) := \frac{1}{2} \sum_{T \in \mathcal{T}_n} |P_Y(T) - P_U(T)|.
\]

For \( T \in \mathcal{T}_n \), it is known

\[
P_U(T) = \frac{1}{(2n-3)!!} \quad \text{and} \quad P_Y(T) = \frac{2^{n-1}}{n! \pi(T)}
\]

where

\[
\pi(T) := \prod_{v \in \overset{\circ}{V}} (\Lambda(v) - 1)
\]

where \( \overset{\circ}{V} \) denotes the set of internal nodes of \( T \), and \( \Lambda(v) \) is the number of descendants of \( v \) (Semple & Steel, 2003).

An application: \( H_0 : YHK \) vs \( H_1 : PDA \),

\[
\text{power} \leq \alpha + mtv_n(YHK, PDA).
\]
By exhaustive enumeration, we computed the numerical values of \( mtv_n(YHK, PDA) \) for \( n = 3, \ldots, 20 \).

For larger \( n \), we estimate \( mtv_n(YHK, PDA) \) by simulation.

Rewrite

\[
mtv_n(YHK, PDA) = \sum_{T \in \mathcal{T}_n} [P_Y(Y) - P_U(T)]_+ \\
= \sum_{T \in \mathcal{T}_n} \left[ \frac{P_Y(Y)}{P_U(T)} - 1 \right]_+ P_U(T) = \mathbb{E}_U[R]
\]

where \( R : \mathcal{T}_n \to \mathbb{R} \) defined as \( R(T) = \left[ \frac{P_Y(Y)}{P_U(T)} - 1 \right]_+ \).

Sample \( T_1, T_2, \ldots, T_m \) from \( \mathcal{T}_n \) according to the PDA model. Compute \( R(T_1), R(T_2), \ldots, R(T_m) \). And

\[
\frac{1}{m} \sum_{i=1}^{m} R(T_i) \approx \mathbb{E}_U[R] = mtv_n(YHK, PDA).
\]
Simulation is conducted up to $n = 80$.

We plot $-\log(1 - mtv_n(YHK, PDA))$ as a function of $n$, suggesting

$$mtv_n(YHK, PDA) \approx 1 - e^{-\alpha - \beta n}.$$
3.2 Discrimination Efficiency

- For a chosen feature of a tree $T \in \mathcal{T}_n$, we propose a score, discrimination efficiency (DE), to assess the extent this feature is able to discriminate $YHK$ from $PDA$.

- For examples,
  - Colless index of balanced tree
    \[
    \text{Colless}(T) = \sum_{v \in \mathring{V}} |R(v) - L(v)|
    \]
    where $\mathring{V}$ denotes the set of interior vertices; $R(v)$ (resp., $L(v)$) denotes the number of descendants by the right (left) child of $v$.
  - Number of cherries in $T$, $C_n(T)$
  - Joint numbers of cherries and pitchforks in $T$, $(A_n(T), C_n(T))$
3.2 Discrimination Efficiency

- Suppose we consider a tree index $I_n$ on $\mathcal{T}_n$

\[
mtv_{I_n}(YHK, PDA) := \frac{1}{2} \sum_a |g_n(a) - \tilde{g}_n(a)|
\]

where $g_n(a) = \sum P_y(T)$ where the sum is taken over all $T \in \mathcal{T}_n$ with $I_n(T) = a$.

- Triangle inequality implies

\[
0 \leq mtv_{I_n}(YHK, PDA) \leq mtv_n(YHK, PDA).
\]

- Define discrimination efficiency of index $I_n$:

\[
DE(C_n) := \frac{mtv_{C_n}(YHK, PDA)}{mtv_n(YHK, PDA)} \in [0, 1].
\]
3.2 Discrimination Efficiency

- Write $\mathcal{T}_n(a) = \{ T \in \mathcal{T}_n : l_n(T) = a \}$.

- Since

$$mtv_{l_n}(YHK, PDA) = \frac{1}{2} \sum_a \left| \sum_{T \in \mathcal{T}_n(k)} [P_y(T) - P_U(T)] \right|,$$

index $l_n$ is optimal

$\iff$ for each $a$, $P_y(T) - P_U(T)$ is of the same sign for all $T \in \mathcal{T}_n(a)$

$\iff$ for each $a$, $P_y(T)$ is a constant

$\iff$ for each $a$, $\pi(T) = \prod_{v \in \mathring{V}} (\Lambda(v) - 1)$ is a constant

- This leads to the tree index: $\sum_{v \in \mathring{V}} \log(\Lambda(v) - 1)$ one that is proposed by Blum and Francois (2006)\(^5\).

---

4 Work in progress

4.1 Extend the results to unrooted trees

- Studying unrooted trees is of interest: many inference methods return unrooted trees first. Then some ways to root this tree if a rooted tree is desired.

- C + Thompson + Wu observed the edge decomposition step can be modified to handle unrooted tree.
4 Work in progress

4.2 We are interested in the correlation of the number of nodes with $a$-descendants and the number of nodes of $b$-descendants

Number of cherries = Number of nodes with 2-descendants;  
Number of pitchforks = Number of nodes with 3-descendants.

▶ Recall YHK: $\rho_{n,YHK}(2,3)$ is a constant in $n$.

C + Wu proved that the same is true for general $a$ and $b$. 
Under PDA model: correlation can be positive!

Left panel: $a = 2, b = 3$, the correlation is negative for all $n$.

Middle panel: $a = 3, b = 4$, the correlation changes sign from $-ve$ to $+ve$.

Right panel: $a = 4, b = 5$, the correlation is positive for all $n$.

C + Wu gave a characterization for which pair $(a, b)$ will lead to left panel and right panel.
4 Work in progress

4.3 Two one-parameter families of evolutionary models

4.3.a. Aldous’ $\beta$-branch split model (1996, 2001)

Prob of left sister clade and right sister clade obtain $i$ leaves and $n-i$ leaves respectively is

$$p_\beta(i|n) = \frac{1}{a_n(\beta)} \frac{\Gamma(\beta + i + 1)\Gamma(\beta + n-i + 1)}{\Gamma(i+1)\Gamma(n-i+1)}, \quad 1 \leq i \leq n-1.$$  

- $\beta = 0$: YHK
- $\beta = -\frac{3}{2}$: PDA

- Yule model did not fit the empirical tree data well.
- Many empirical trees are consistent with Aldous’ branch split model: $\beta \approx -1$.\(^6\)

\(^6\)AB model in previous plot corresponds to $\beta = -1$
4 Work in progress

4.3 Two one-parameter families of generative models

4.3.b. Ford’s $\alpha$ model (2005)

- Starting with a small tree, grow it to the desired size by adding one leaf at a time.
- Fix $\alpha \in [0, 1]$.
  - Attach a new leaf to the current tree’s internal edge with probability $\alpha$; and to a pendant edge $1 - \alpha$.
- YHK: $\alpha = 0$;
  - PDA: $\alpha = \frac{1}{2}$. 

p29