PHYLOGENETIC MIXTURES AND LINEAR INVARIANTS FOR EQUAL INPUT MODELS

MARTA CASANELLAS AND MIKE STEEL

Abstract. The reconstruction of phylogenetic trees from molecular sequence data relies on modelling site substitutions by a Markov process, or a mixture of such processes. In general, allowing mixed processes can result in different tree topologies becoming indistinguishable from the data, even for infinitely long sequences. However, when the underlying Markov process supports linear phylogenetic invariants, then provided these are sufficiently informative, the identifiability of the tree topology can be restored. In this paper, we investigate a class of processes that support linear invariants once the stationary distribution is fixed, the ‘equal input model’. This model generalizes the ‘Felsenstein 1981’ model (and thereby the Jukes–Cantor model) from four states to an arbitrary number of states (finite or infinite), and it can also be described by a ‘random cluster’ process. We describe the structure and dimension of the vector space of phylogenetic mixtures (and the complementary space of linear invariants) for any fixed phylogenetic tree (and for all trees – the so called ‘model invariants’), on any number of leaves. We also provide a precise description of the space of mixtures and linear invariants for the special case of 4 leaves. By combining techniques from discrete random processes and (multi-)linear algebra, our results build on a classic result that was first established by James Lake in 1987.

1. Introduction

Tree–based Markov processes on a discrete state space play a central role in molecular systematics. These processes allow biologists to model the evolution of characters thereby to develop techniques for inferring a phylogenetic tree for a group of species from a sequence of characters (such as the sites at aligned DNA or amino acid sequences [5]). Under the assumption that each character evolves independently on the same underlying tree, according to a fixed Markov process, the tree topology can be inferred in a statistically consistent way (i.e. with an accuracy approaching 1 as the number of characters grow) by methods such as maximum likelihood estimation (MLE) [4] and techniques based on phylogenetic invariants [6]. This holds even though one may not know the values of the other (continuous) parameters associated with the model, which typically relate to the length of the edges, and relative rates of different substitution types.

The assumption that all characters evolve under the same Markov process is a very strong one, and biologists generally allow the underlying process to vary in some way between the characters. For example, a common strategy is to allow characters to evolve at different rates (i.e. the edge lengths are all scaled up or down in equal proportion at each site by a factor sampled randomly from some simple parameterized distribution). In that case, provided the rate distribution is sufficiently constrained, the tree topology can still be inferred in a statistically consistent manner [1] [3], and by using MLE, or related methods.

However, when this distribution is not tightly constrained, or when edge lengths are free to vary in a more general fashion from character to character then different trees can lead to identical probability distributions on characters [1] [6]. In that case, it can be impossible to decide which of two (or more) trees generated the given data, even when the number of characters tends to infinity. In statistical terminology, identifiability of the tree topology parameter is lost. For certain types of Markov models, however, identifiability of the tree topology is possible, even
in these general settings. These are models for which (i) linear relationships (called ‘linear phylogenetic invariants’) exist between the probabilities of different characters, and which are invariant to other continuous parameters associated with the model (such as edge lengths) and (ii) these invariants can be used to determine the tree topology (at least for \( n = 4 \) leaves) \[18, 19\]. The first such linear invariants were discovered by James Lake in a landmark paper in 1987 \[12\] for the Kimura 2ST model, and the Jukes–Cantor submodel.

In this paper we extend Lake–type invariants to a more general setting and for another type of process, the ‘equal input’ model (defined shortly, but it can be regarded as the simplest Markov process that allows different states to have different stationary probabilities). By building also on the approach of \[13\] (which dealt just with the 2-state setting) we investigate the vector space of linear invariants, and describe the space of phylogenetic mixtures on a tree (or trees) under the equal input model once the stationary distribution is fixed. In brief summary, our main results describe the vector space (and its dimension) of the space of phylogenetic mixtures of the equal input models for any number \( n \) of leaves:

- across all trees (Theorem 16) by classifying the linear ‘model invariants’;
- for a fixed tree (Theorem 25); and
- for an infinite state version of the equal input model, known as Kimura’s infinite allele model (Proposition 38).

Theorem 36 and Corollary 37 provide a more detailed description for trees with \( n = 4 \) leaves. The case \( n = 4 \) is of particular interest, since the existence of a set of linear phylogenetic invariants for this case and which, collectively, suffice to identify the tree topology means that there also exist informative linear phylogenetic invariants that can identify any fully-resolved (binary) tree topology on any number of leaves. This follows from the well-known fact that any binary tree topology is fully determined by its induced quartet trees (for details and references, see \[15\]).

We also establish various other results along the way, including a ‘separability condition’ from which a more general description of Lake–type invariants follows (Proposition 8). We begin with some standard definitions, first for Markov processes on trees, and then for the equal input model, which we show is formally equivalent to a random cluster process on a tree (Proposition 6). We then develop a series of preliminary results and lemmas that will lead to the main results described above.

2. Markov processes on trees

Given a tree \( T = (V, E) \) with leaf set \( X \), a Markov process on \( T \) with state space \( S \) is a collection of random variables \( (Y_v : v \in V) \) taking values in \( S \), and which satisfies the following property. For each interior vertex \( v \) in \( T \), if \( V_1, \ldots, V_m \) are the sets of vertices in the connected components of \( T - v \) then the \( m \) random variables \( W_i = (Y_v : v \in V_i) \) are conditionally independent given \( Y_v \).

Equivalently, if we were to direct all the edges away from some (root) vertex, \( v_0 \), then this condition says that conditional on \( Y_v \) (for an interior vertex \( v \) of \( T \)) the states in the subtrees descended from \( v \) are independent of each other, and are also independent of the states in the rest of the tree.

A Markov process on \( T \) is determined entirely by the probability distribution \( \pi \) at a root vertex \( v_0 \), and the assignment \( e \mapsto P(e) \), that associates a transition matrix with each edge \( e = (u, v) \) of \( T \) (the edge is directed away from \( v_0 \)). Matrix \( P(e) \) has row \( \alpha \) and column \( \beta \) entry equal to \( P_\alpha(\beta) := P(Y(v) = \beta | Y(u) = \alpha) \), and so each row sums to 1. If stochastic vector \( \pi \) has the property that \( \pi \pi P(e) \) for every edge \( e \) of \( T \), then \( \pi \) is said to be a stationary distribution for the process. A phylogenetic model is a Markov process on a tree where the transition matrices are required to belong to a particular class \( \mathcal{M} \).
In this paper we will be concerned with trees in which the set \( X \) of leaves are labelled, and all non-leaf (interior) vertices are unlabelled and have degree at least three; these are called phylogenetic \( X \)-trees [15]. A tree with a single interior vertex is called a star, while a tree for which every interior vertex has degree three is said to be binary. We will write \( abcd \) for the binary tree on four leaves (a quartet tree) that has an edge separating leaves \( a,b \) from \( c,d \). A function \( \chi : X \to S \) is called a character and any Markov process on a tree with state space \( S \) induces a (marginal) probability distribution on these characters. An important algebraic feature of this distribution is that the probability of a character \( P(\chi) \) under a Markov process on \( T \) is a polynomial function of the entries in the transition matrices.

### 2.1. The equal input model.

The equal input model (EI) for a set \( S \) of \( \kappa \) states is a particular type of Markov process on a tree, defined as follows. Given a root vertex \( v_0 \) let \( \pi \) be a distribution of states at \( v_0 \) and for each (directed) edge \( e = (u,v) \) (directed away from \( v_0 \)). In the EI model, each transition matrix \( P(e) \) has the property that for some value \( \theta_e \in [0,1] \) and all states \( \alpha, \beta \in S \) with \( \alpha \neq \beta \) we have:

\[
P_{\alpha \beta} = \pi_\beta \cdot \theta_e.
\]

We shall assume that the distribution \( \pi \) is strictly positive throughout the paper.

This model generalizes the familiar fully symmetric model of \( \kappa \) states (such as the ‘Jukes-Cantor model’, when \( \kappa = 4 \)) to allow each state to have its own stationary probability. In the case \( \kappa = 4 \) with \( S \) equal to the four nucleotide bases, the model is known as the Felsenstein 1981 model. The defining property of the model is that the probability of a transition from \( \alpha \) to \( \beta \) (two distinct states) is the same, regardless of the initial state \( \alpha(\neq \beta) \).

**Lemma 1.** The following properties hold for the equal input model.

(i) \( P_{\alpha \alpha}(e) = 1 - \theta_e + \pi_\alpha \theta_e \).

(ii) \( \pi \) is a stationary distribution for each vertex \( v \) of the \( T \) (i.e. \( P(Y(v) = \alpha) = \pi_\alpha \)).

(iii) The process is time-reversible (i.e. for each edge \( e, \pi_\alpha P_{\alpha \beta}(e) = \pi_\beta P_{\beta \alpha}(e) \)).

(iv) If \( p \) is the probability that the ends of \( e \) receive different states under the EI model, then \( p = (1 - \sum_\alpha \pi_\alpha \theta_e) \).

(v) The process is multiplicatively closed. In other words, \( (P(e)P(e'))_{\alpha \beta} = \pi_\beta \theta_e, \) where \( \theta = 1 - (1 - \theta_e)(1 - \theta_e) \).

**Proof.** For (i), \( P_{\alpha \alpha}(e) = 1 - \sum_{\beta \neq \alpha} P_{\alpha \beta}(e) = 1 - \theta_e \sum_{\beta \neq \alpha} \pi_\beta = 1 - \theta_e(1 - \pi_\alpha) \). For (ii), it suffices to show that if \( (u,v) \) is a directed edge and \( u \) has stationary distribution \( \pi \) then \( v \) does too. But

\[
P(Y(v) = \beta) = \sum_\gamma \pi_\gamma P_{\gamma \beta}(e) = \pi_\beta P_{\beta \beta}(e) + \sum_{\gamma \neq \beta} \pi_\gamma P_{\gamma \beta}(e).
\]

For (iii), the result clearly holds if \( \alpha = \beta \) so suppose \( \alpha \neq \beta \). Then

\[
\pi_\alpha P_{\alpha \beta} = \pi_\alpha (\pi_\beta \theta_e) = \pi_\beta (\pi_\alpha \theta_e) = \pi_\beta P_{\beta \alpha}(e).
\]

For (iv),

\[
p = \sum_\alpha \pi_\alpha \sum_{\beta \neq \alpha} P_{\alpha \beta}(e) = \sum_\alpha \pi_\alpha \sum_{\beta \neq \alpha} \pi_\beta \theta_e,
\]

which simplifies for the expression in (iv). Property (v) is left as an exercise. ♦

For an equal input model, the transition matrix \( P(e) \) has eigenvalue \( 1 - \theta_e \), with multiplicity \( k - 1 \) (and eigenvalue 1 with multiplicity 1). Also, for fixed \( \pi \) the matrices \( P(e) \) commute, as they can be simultaneously diagonalized by a fixed matrix (which depends on \( \pi \)). Equal input models
with also have a continuous realisation with rate matrix $Q$ defined by its off-diagonal entries as follows:

$$Q_{\alpha\beta} = \pi_\beta, \text{ for all } \alpha, \beta \in S, \alpha \neq \beta$$

(the diagonal entries are determined by the requirement that each row of $Q$ sums to 0). Then $P^r = e^{\mathsf{Q} t}$ for $t = -\ln(1 - \theta_e)$, and so $\theta_e = 1 - e^{-t}$. In the case where $\pi$ is uniform, the $EI$ model reduces to the fully symmetric model in which all substitution events have equal probability.

One feature of the $EI$ model, that fails for most other Markov processes on trees, is the following. Let $\sigma$ be any partition of the state space $S$, and for a state $s \in S$ let $[s]$ denote the corresponding block of $\sigma$ containing $s$. Then for an $EI$ process $Y$ on the set $V$ of vertices of a phylogenetic tree $T$, let $\tilde{Y}$ be the induced stochastic process on $V$, defined by $\tilde{Y}(v) = [Y(v)]$ for all vertices $v$ of $T$.

**Proposition 2.** For any $EI$ model with parameters $\pi$ and $\{\theta_e\}$, and any partition $\sigma$ of $S$, $\tilde{Y}$ is also an $EI$ Markov process on $T$, with parameters $\tilde{\pi}$ and $\{\tilde{\theta}_e\}$, where for each block $B$ of $\sigma$,

$$\tilde{\pi}_B := \sum_{\beta \in B} \pi_\beta.$$

**Proof.** By Theorem 6.3.2 of [11], the condition for $\tilde{Y}$ to be a Markov process is that it satisfies ‘lumpability’ criterion that for any two choices $\alpha, \alpha' \in A \in \sigma$, and block $B \in \sigma$,

$$P(\tilde{Y}(v) \in B|Y(u) = \alpha) = P(\tilde{Y}(v) \in B|Y(u) = \alpha').$$

For each $B \neq A$, this last equality is clear from [11], and since $P(\tilde{Y}(v) \in A|Y(v) = \alpha) = 1 - \sum_{B \in \sigma, B \neq A} P(\tilde{Y}(v) \in B|Y(u) = \alpha)$ the criterion also holds for the case $B = A$. Finally, for $B \neq A$, $P(\tilde{Y}(v) = B|Y(u) = A) = \sum_{\beta \in B} (\pi_\beta \theta_e) = \tilde{\pi}_B \theta_e$. \qed

### 2.2. A Useful Lemma

For results to come the following lemma, and its corollary will be helpful.

**Lemma 3.** For variables $x_1, x_2, \ldots, x_r$, consider any polynomial $h(x) \in \mathbb{R}[x_1, \ldots, x_r]$ of the form

$$h(x) = \sum_{A \subseteq [r]} c_A \prod_{j \in A} x_j,$$

where $c_A \in \mathbb{R}$. Then $c_A = 0$ for all $A \subseteq [r]$ (i.e. $h \equiv 0$) if and only if for any $t \neq 0$, $h(x) = 0$ for all $x \in \{0, t\}^r$.

**Proof.** The ‘only if’ part holds automatically; for the ‘if’ direction, given any subset $B$ of $[r]$, let $h(B) = h(x^B)$ where $x_i^B = t$ if $i \in B$ and $x_i^B = 0$ otherwise. Then $h(B) = 0$ by hypothesis, and $h(B) = \sum_{A \subseteq B} c_A \prod_{\alpha \in A} x_\alpha^B$, by definition. Applying the (generalized) principle of inclusion and exclusion it follows that, for each $A \subseteq [n]$, $c_A l^{|A|} = \sum_{B \subseteq A} (-1)^{|A - B|} h(B) = 0$, so $c_A = 0$. \qed

The following corollary is an immediate consequence of the previous lemma.

**Corollary 4.** Consider variables $x_1, x_2, \ldots, x_r$ and for each $i = 1, \ldots, M$ let $f_i$ be a polynomial of the form

$$\sum_{A \subseteq [r]} c_A^{(i)} \prod_{j \in A} x_j,$$

Let $f = (f_1, \ldots, f_M) : \mathbb{R}^r \to \mathbb{R}^M$ and let $L : \mathbb{R}^M \to \mathbb{R}$ be a linear map. Define an equivalence relation among the elements of $\{0, 1\}^r$ by $x \sim x'$ if $f(x) = f(x')$, and let $x_1, \ldots, x_s$ be representatives of these equivalence classes. We call $q_i = f(x_i)$, $i = 1, \ldots, s$. Then $L(f(x)) = 0$ for all $x \in \mathbb{R}^r$ if and only if $L(q_j) = 0$ for $j = 1, \ldots, s$. 


Proof. The map \( h = L \circ f \) satisfies the hypotheses of the previous lemma, hence \( L(f(x)) = 0 \) for all \( x \) if and only if \( L(f(x)) = 0 \) for all \( x \in \{0, 1\}^r \). Then the statement follows immediately due to the definition of the equivalence relation. \( \square \)

In what follows we will use this corollary to check linear relations among the character probabilities.

In the \( EI \) model, once we fix \( \pi \), the probability \( \mathbb{P}_T(\chi|\Theta) \) of observing a character at the leaves of \( T \) satisfies the hypotheses of the corollary with \( r \) equal to the number of edges and variables in \( \Theta = \{ \theta_e \}_{e \in E(T)} \). Indeed, by Lemma 1 (i), any entry of the transition matrix \( P(e) \) is a linear function of \( \theta_e \) and hence the expression

\[
\mathbb{P}_T(\chi|\Theta) = \sum_{(s_v) \in S^{|\text{Int}(T)|}} \prod_{(u, w) \in E(T)} P_{s_u, s_w}(e) \pi_{s_v} \quad (\text{where the sum is over the states at the set } \text{Int}(T) \text{ of interior vertices of } T \text{ and subject to the convention that } s_w = \chi(l) \text{ if } w \text{ is the leaf } l) \]

satisfies the hypotheses of Lemma 3 and its corollary.

Remark 5. Lemma 3 and its corollary can be slightly modified to accommodate substitution matrices with more parameters as it was done in [7].

2.3. The equal input model as a random cluster model. Our alternative description of the \( EI \) model is as an instance of the (finite) random cluster model (briefly \( RC \)) on trees (this phrase is also used to study processes on graphs, such as the ‘Ising model’ in physics). For an unrooted phylogenetic tree with leaf set \([n]\), each edge \( e \) of \( T \) is cut independently with probability \( p_e \). The leaves in each connected component of the resulting disconnected graph are then all assigned the same state \( s \) with probability \( \pi_s \), independently of assignments to the other components (see Fig. 1). More precisely, for any binary function \( g : E(T) \to \{0, 1\} \), define \( C(g) \) to be the set of connected components in \( T \{ e \in E(T) | g(e) = 1 \} \). Then the probability \( \mathbb{P}_T(\chi|\{p_e\}_e) \) of observing a character \( \chi \) at the leaves of \( T \) under the \( RC \) model is

\[
\mathbb{P}_T(\chi|\{p_e\}_e) = \sum_{g : E(T) \to \{0, 1\}} \mathbb{P}(\chi|g)p_g(e)^{g(e)}(1 - p_e)^{1-g(e)}
\]

where \( \mathbb{P}(\chi|g) = 0 \) if \( \chi(i) \neq \chi(j) \) for some leaves \( i, j \) in the same connected component in \( C(g) \) and is equal to \( \prod_{e \in C(g)} \pi_{e \chi} \) otherwise (where \( \chi_e \) denotes the value of \( \chi \) at the leaves of \( T \) that are in \( e \)). In particular, the \( RC \) model also satisfies the hypotheses of Lemma 3 and its corollary.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{tree.png}
\caption{Cutting the three edges marked * in the tree on the left leads to the partition of \( X \) shown at right. Under the random cluster model these four blocks are independently assigned states from the distribution \( \pi \).}
\end{figure}
Proposition 6. The EI model with parameters $\pi$ and $\{\theta_e\}$ produces an identical probability distribution on characters as the random cluster model in which $p_e = \theta_e$ for each edge $e$ of $T$.

Proof. For the two models the probability of a given character (given by Equations (2) and (3)) satisfies the conditions required by Lemma 3 and so we can use Corollary 4 with $M = 2$ and $L$ the difference between the probability of a given character by the two models. Therefore, it suffices to show that the two models produce the same probability distribution on characters whenever $\theta_e = 1$ for all $e \in F$ and $\theta_e = 0$ of all edges $e$ of $T$ not in $F$ (for all possible choices of subset $F \in E(T)$). Given $F$, notice that if $\theta_e = 1$ for a directed edge $e = (u, v)$ of $T$ in the EI model, then $P^{(e)}_{\alpha\beta} = \pi_{\beta}$ for all $\beta \in S$, including $\beta = \alpha$. In other words, when $\theta_e = 1$ for $e = (u, v)$, the state at $v$ is completely independent of the state at $u$. This is equivalent to cutting the edge and assigning a random state according to the distribution $\pi$ to $v$, and thereby to all the other vertices of $T$ for which there is a path to $v$ that does not cross another edge in $F$ (since $P^{(e)}$ is the identity matrix on those edges); this is just the process described by the random cluster model. \hfill \Box

3. LINEAR PHYLOGENETIC INVARIANTS IN PHYLOGENETICS

Definition 7. Consider a phylogenetic model $M$ with state space $S$ on a phylogenetic tree $T$ with $n$ leaves. A phylogenetic invariant of a tree $T$ under the model $M$ is a polynomial $f$ in $S^n$ indeterminates that vanishes on any distribution $P_{T,\Theta}$ that arises under the phylogenetic model $M$ (that is, $f(p) = 0$ if $p = P_{T,\Theta}$, for any set $\Theta$ of transition matrices and distribution at the root vertex).

We say that a polynomial in $S^n$ coordinates is a model invariant if it is a phylogenetic invariant for any tree on $n$ leaves under the phylogenetic model $M$. A phylogenetic invariant of a tree $T$ that is not a model invariant is called a topology invariant.

A phylogenetic invariant is a linear phylogenetic invariant (resp. linear model invariant, linear topology invariant) if each monomial involves exactly one indeterminate and has degree 1. Note that this implies that the polynomial is homogeneous (the independent term is 0). There are phylogenetic invariants of degree 1 that are not homogeneous, for example the trivial phylogenetic invariant that arises from the observation that in a distribution all coordinates must sum to one. However, taking this trivial invariant into account, any other phylogenetic invariant of degree 1 can be rewritten as a homogeneous phylogenetic invariant of degree 1 (indeed, $\sum_i a_i x_i + a$ is a phylogenetic invariant if and only if $\sum_i (a_i + a) x_i$ is a phylogenetic invariant). This is why we only call linear phylogenetic invariants those that are homogeneous of degree 1. The sets of linear model invariants and linear phylogenetic invariants of a tree $T$ are vector spaces.

Linear phylogenetic invariants are of particular interest since they hold even if the process changes from character to character (provide it stays within the model for which the invariant is valid). An important early example of linear phylogenetic invariants were discovered by James Lake in 1987 [12] and we now describe a more general version of this result.

For any phylogenetic $X$–tree, $T$ (not necessarily binary), and an interior vertex $v$ of $T$ consider the disconnected graph $T - v$. Let $t$ and $t'$ be two of the trees incident with $v$.

Suppose that a Markov process $Y$ on $T$ takes values in state space $S$. For any state $s$ of $S$ write $Y(t) = s$ if all the leaves of $T$ that are in $t$ are in state $s$ (similarly for $t'$). Consider the following property.

(PS) Partial separability. For some interior vertex $v$, and for some subset $\{a_1, a_2, b_1, b_2\}$ of four distinct elements of $S$ one has

$$P(Y(t) = a_i | Y(v) = s) = \pi(a_i)c, \text{ when } s \in S - \{a_1, a_2\}, i = 1, 2;$$
and
\[ \mathbb{P}(Y(t') = b_j | Y(v) = s) = \pi(b_j) d, \] when \( s \in S - \{b_1, b_2\}, j = 1, 2. \]

Here \( c \) and \( d \) are arbitrary functions dependent on the tree and associated parameters (but not the states) and \( \pi \) is an arbitrary function of the states such that \( \pi(a_i) \neq 0, \pi(b_i) \neq 0, i = 1, 2. \)

Partial separability is satisfied by various models. For example, when \( |S| = 4 \), it holds for the Kimura 2-ST model (and hence the Jukes-Cantor model) by taking \( \{a_1, a_2\} = \{A, G\} \) (purines) and \( \{b_1, b_2\} = \{C, T\} \) (pyrimidines), in which case \( \pi(a_i) = \pi(b_i) = \frac{1}{4} \) for \( i = 1, 2. \) The property also holds for the fully symmetric model on any number of states. Moreover, the property holds for the \( EI/RC \) model on any number of states if \( t \) and \( t' \) are single leaves (in this case \( \pi \) can be taken to be the stationary distribution).

Let \( \mathcal{E} \) be any event that involves the states at the leaves of \( T \) not in \( t \) or \( t' \). For example, if \( y \) and \( y' \) are leaves of \( T \) not in \( t \) or \( t' \) then \( \mathcal{E} \) might be the event that \( Y(y) = s \) and \( Y(y') = s' \) for some particular \( s, s' \in S. \)

Let us write \( \tilde{p}_{\mathcal{E}ij} \) for the probability of the three–way conjunction \( \mathcal{E} \land \{Y(t) = a_i\} \land \{Y(t') = b_j\} \). Notice that \( p_{\mathcal{E}ij} \) is a sum of probabilities of various characters (i.e. a marginal distribution). Let

\[ \tilde{p}_{\mathcal{E}ij} = \frac{1}{\pi(a_i)\pi(b_j)} \cdot p_{\mathcal{E}ij} \text{ and let } \Delta := \tilde{p}_{\mathcal{E}11} + \tilde{p}_{\mathcal{E}22} - \tilde{p}_{\mathcal{E}12} - \tilde{p}_{\mathcal{E}21}. \]

**Proposition 8** (Lake–type invariants). *If a Markov process on \( T \) satisfies the partial separability condition \((PS)\), then \( \Delta = 0. \)*

**Proof.** By the Markov property,
\[ \tilde{p}_{\mathcal{E}ij} = \sum_s \mathbb{P}(Y(v) = s) \cdot \mathbb{P}(\mathcal{E}|Y(v) = s) \cdot \mathbb{P}(Y(t) = a_i|Y(v) = s) \cdot \mathbb{P}(Y(t') = b_j|Y(v) = s). \]

Let \( r_{ij} = \pi(a_i) \cdot \pi(b_j) \), and let

\[ \Delta_s = r_{22}p_1p'_1 + r_{11}p_2p'_2 - r_{12}p_2p'_1 - r_{21}p_1p'_2, \]

where \( p_i = \mathbb{P}(Y(t) = a_i|Y(v) = s) \), and \( p'_j = \mathbb{P}(Y(t') = b_j|Y(v) = s) \). Then we can write

\[ \Delta = \frac{1}{\pi(a_1)\pi(a_2)\pi(b_1)\pi(b_2)} \sum_s \mathbb{P}(Y(v) = s) \cdot \mathbb{P}(\mathcal{E}|Y(v) = s) \cdot \Delta_s. \]

Thus it suffices to show that \( \Delta_s = 0 \) for all \( s. \)

We consider three cases: (i) \( s \in \{a_1, a_2\} \), (ii) \( s \in \{b_1, b_2\} \) and (iii) \( s \in S - \{a_1, a_2, b_1, b_2\}. \)

In Case (i), suppose \( s = a_i. \) Then \( p'_1 = \pi(b_1) d \) and \( p'_2 = \pi(b_2) d, \) and so

\[ \Delta_s = d[p_1r_{22}\pi(b_1) + p_2r_{11}\pi(b_2) - p_2r_{12}\pi(b_1) - p_1r_{21}\pi(b_2)] = dp_1[r_{22}\pi(b_1) - r_{21}\pi(b_2)] + dp_2[r_{11}\pi(b_2) - r_{12}\pi(b_1)] = 0 + 0 = 0. \]

Case (ii) is similar. In Case (iii), \( p_i p'_j = r_{ij} \) and so

\[ \Delta_s = cd[r_{22}r_{11} + r_{11}r_{22} - r_{12}r_{21} - r_{21}r_{12}] = 0. \]

\( \square \)

**Example 9.** When we take \( t \) and \( t' \) single leaves, the \( EI/RC \) model satisfies the \((PS)\) property and Lemma 8 can be applied. If the stationary distribution \( \pi \) is fixed, then \( \Delta \) gives rise to two linear phylogenetic invariants for the quartet tree 12|34,

\[
H_1 : \frac{x_{xy}}{\pi(x)\pi(y)} + \frac{x_{yz}}{\pi(z)\pi(w)} - \frac{x_{xy}}{\pi(z)\pi(y)} - \frac{x_{yz}}{\pi(x)\pi(w)}
\]

\[
H_2 : \frac{x_{yxy}}{\pi(x)\pi(y)} + \frac{x_{xyz}}{\pi(z)\pi(w)} - \frac{x_{yzy}}{\pi(z)\pi(y)} - \frac{x_{xzy}}{\pi(x)\pi(w)}
\]
Fibonacci number, starting with $F_1$ and $F_2$. A singleton block is a full subforest of $T$, that is, a collection of vertex disjoint trees $\{T_1, \ldots, T_r\}$ such that the only nodes of degree 1 in $T_i$ are leaves of $T$ (we allow $T_i$ to be formed by only one leaf and we allow $F = \{T\}$ also). We say that a subforest $F = \{T_1, \ldots, T_r\}$ is a full subforest if $\bigcup_i L(T_i) = X$; we let $\mathcal{F}_T$ be the set of full subforests of $T$. For a full subforest $F$, we define $\Theta_F$ to be the following collection of edge parameters under the RC model: $\theta_e = 0$ if $e \in E(T_i)$ for some $T_i \in F$ and $\theta_e = 1$ for all other edges $e$. We denote by $\sigma(F)$ the partition that $F$ describes on $[n]$, that is, two leaves are in the same block of $\sigma(F)$ if they lie in the same subtree of $F$. Note that the set of full subforests of a phylogenetic $X$–tree $T$ is the union of the full subforests of the star tree on $n$ leaves and the set of full subforests generated by deleting a collection of interior edges of $T$. The full subforest formed by singletons will be called the trivial subforest.

Given a character $\chi$, we define $\sigma(\chi)$ to be the partition $\{S_1, \ldots, S_l\}$ of the set of leaves defined according to “two leaves $i, j$ are at the same block of the partition if $\chi_i = \chi_j$”. Note that given a full subforest $F = \{T_1, \ldots, T_r\}$ of $T$ and a character $\chi$, $\mathbb{P}_T(\chi|\Theta)$ is zero if $\sigma(F)$ does not refine $\sigma(\chi)$ and is equal to $\prod_{i=1}^r \pi_{s_i}$ otherwise (here $s_i$ stands for the value of $\chi$ at the leaves of $T_i$).

For any partition $\sigma$ of $[n]$, and any phylogenetic tree $T$ on $[n]$, we say that $\sigma$ is convex on $T$ (or compatible with $T$) if the collection of induced subtrees $\{T[B] : B \in \sigma\}$ are vertex disjoint (here $T[B]$ is the minimal connected subgraph (subtree) of $T$ containing the leaves in $B$). Let $co(T)$ be the set of partitions of $[n]$ that are convex on $T$. There is a natural correspondence between full subforests of $T$ and convex partitions on $T$ that associates to each partition $\sigma$ in $co(T)$ the full subforest $\{T[B] : B \in \sigma\}$. Therefore, the number of full subforests of a tree $T$ is equal to $|co(T)|$, $|\mathcal{F}_T| = |co(T)|$. When $T$ is a binary tree, $|co(T)| = F_{2n-1}$ where $F_k$ is the $k$–th Fibonacci number, starting with $F(1) = F(2) = 1$ (see [17]). By contrast, for a star tree on $[n]$ $|co(T)| = 2^n - n$. A partition $\sigma = \{B_1, \ldots, B_k\}$ of $[n]$ is incompatible with $T$ if it is not convex on $T$, that is, there exist two blocks $B_i$ and $B_j$ from $\sigma$ for which $T[B_i]$ and $T[B_j]$ share at least one vertex. A singleton block $B$ of $\sigma$ is a block of size 1. The number of partitions of $[n]$ is known as the Bell number $B_n$.

4.2. Results.

**Lemma 10.** (a) Let $\Theta$ be a collection of parameters $(\theta_e)_{e \in E(T)}$ such that $\theta_e$ is either 0 or 1 for all $e \in E(T)$. Then there exists a unique full forest $F \in \mathcal{F}_T$ such that $\mathbb{P}_{T,\Theta} = \mathbb{P}_{T,\Theta_F}$.
(b) A degree 1 polynomial \( \sum_{\chi} \lambda_{\chi} x_{\chi} \) is a linear phylogenetic invariant for a tree \( T \) if and only if

\[
\sum_{\chi} \lambda_{\chi} \mathbb{P}_T(\chi|\Theta) = 0
\]

for any full subforest \( F \in \mathcal{F}_T \).

Proof. (a) We first prove that two full subforests \( F \) and \( G \) satisfy \( \mathbb{P}(\chi|\Theta) \neq \mathbb{P}(\chi|\Theta') \) for some \( \chi \) if \( F \neq G \). As \( F,G \) are full subforests, they are different if and only if they induce different partitions \( \sigma(F), \sigma(G) \) on the set of leaves. Then there exists an interior edge \( e_0 \) such that \( e_0 \) is compatible with \( \sigma(F) \) but is not compatible with \( \sigma(G) \). If \( \chi \) is the character that assigns state \( x \) to the leaves of one connected component of \( T - e_0 \) and state \( y \neq x \) at the leaves of the other component, then \( \mathbb{P}(\chi|\Theta_G) = 0 \) while \( \mathbb{P}(\chi|\Theta_T) \) is not zero.

Given \( \Theta \), let \( A \) be the set of edges \( e \) in \( T \) such that \( e \epsilon = 1 \). Let \( \sigma(T \setminus A) \) be the partition induced on \( X \) when removing all edges in \( A \) (if an edge in \( A \) is a pendant edge, then removing it means that we separate the corresponding leaf). Let \( F \) be the full subforest that induces the same partition as \( \sigma(T \setminus A) \) in the set of leaves. Then we have \( \mathbb{P}_{T,\Theta} = \mathbb{P}_{T,\Theta_F} \).

(b) This follows from part (a) and Corollary 4. \( \square \)

Let \( \Theta \) be a collection of edge parameters on a tree \( T \) evolving under the RC model. For a site character \( \chi \), we define

\[
\tilde{\pi}_T^T(\Theta) = \frac{\mathbb{P}_T(\chi|\Theta)}{\pi_{\chi_1} \pi_{\chi_2} \cdots \pi_{\chi_n}}.
\]

We call \( \tilde{\pi}_T \) the corresponding coordinates:

\[
\tilde{\pi}_T = \left( \frac{\mathbb{P}_T(\chi_1|\Theta)}{\pi_{\chi_1}}, \frac{\mathbb{P}_T(\chi_2|\Theta)}{\pi_{\chi_2}}, \cdots, \frac{\mathbb{P}_T(\chi_n|\Theta)}{\pi_{\chi_n}} \right).
\]

Lemma 11. We say that two characters \( \chi \) and \( \chi' \) are equivalent, \( \chi \equiv \chi' \), if \( \sigma(\chi) = \sigma(\chi') \) and \( \chi_i = \chi'_i \) for any leaf \( i \) that belongs to a block of the partition of cardinality greater than or equal to 2. Let \( \chi, \chi' \) be two characters on the set \( X = \{1, 2, \ldots, n\} \).

(a) If \( \chi \equiv \chi' \) then \( \tilde{\pi}_T - \tilde{\pi}_T' \) is a linear model invariant.

(b) If \( \pi \) is not invariant by any permutation of the set of states, then for any tree \( T \) the equality \( \tilde{\pi}_T^T(\Theta) = \tilde{\pi}_T'^T(\Theta) \) for every \( \Theta \) implies that \( \chi \equiv \chi' \) (i.e. in this case every linear phylogenetic invariant of type \( \tilde{\pi}_T \) satisfies \( \chi \equiv \chi' \)).

Proof. (a) Let \( \chi \) and \( \chi' \) be two equivalent characters, let \( \sigma \) be \( \sigma(\chi) = \sigma(\chi') \), and let \( T \) be any \( X \)-tree. According to Lemma 11 (b) we need to check that \( \tilde{\pi}_T(\Theta_F) = \tilde{\pi}_T(\Theta_F) \) for any \( F = \{T_1, \ldots, T_s\} \in \mathcal{F}_T \).

If \( \sigma(F) \) does not refine \( \sigma \), then \( \mathbb{P}_T(\chi|\Theta_F) \) and \( \mathbb{P}_T(\chi'|\Theta_F) \) are zero and we are done.

If \( \sigma(F) \geq \sigma \), then \( \mathbb{P}_T(\chi|\Theta_F) = \pi_{s_1} \cdots \pi_{s_n} \), where \( s_i \) is the value of \( \chi \) at the leaves of \( T_i \) (note that we may have \( s_i = s_j \)). Therefore \( \tilde{\pi}_T^T(\Theta_F) = \frac{\mathbb{P}_T(\chi|\Theta_F)}{\pi_{s_1} \cdots \pi_{s_n}} \) where \( n_i = |\mathcal{L}(T_i)| \). As \( \sigma(F) \geq \sigma(\chi) = \sigma(\chi') \) and the states of \( \chi \) and \( \chi' \) coincide for any block of \( \sigma \) of size \( \geq 2 \), the states of \( \chi \) and \( \chi' \) also coincide at the leaves of \( T_i \), if \( n_i \geq 2 \). Therefore, \( \tilde{\pi}_T^T(\Theta_F) = \tilde{\pi}_T'^T(\Theta_F) \).

As for (b), assume that \( \pi \) is not invariant by any permutation of the set of states (i.e. \( \pi_{s_i} = \pi_{s_j} \)) if and only if \( s_i = s_j \). Assume that for a tree \( T \) we have \( \tilde{\pi}_T^T(\Theta_F) = \tilde{\pi}_T'^T(\Theta_F) \) for any collection of edge parameters \( \Theta_F \). Then, for each block \( B_i \) of \( \sigma(\chi) \) of size \( b_i \) greater or equal than 2 consider the forest \( F_i = \{TB_i, \cup \notin B, \{1\}\} \), where \( TB_i \) is the smallest subtree of \( T \) joining the leaves of \( B_i \). Then \( \tilde{\pi}_T^T(\Theta_F) = \frac{\pi_{s_i}}{\pi_{s_i}'} \) if \( s_i \) is the state of \( \chi \) at the leaves of \( B_i \). By hypothesis this is equal to \( \tilde{\pi}_T'^T(\Theta_F) \). But \( \tilde{\pi}_T^T(\Theta_F) \) is zero if \( \sigma(\chi') \) does not contain the block \( B_i \). Performing the same argument for any block \( B_i \) of size \( b_i \geq 2 \) we obtain \( \sigma(\chi) = \sigma(\chi') \). Now for each such block \( B_i \) we have \( \tilde{\pi}_T^T(\Theta_F) = \tilde{\pi}_T^T'(\Theta_F) \) and hence \( \frac{\pi_{s_i}}{\pi_{s_i}'} = 1 \) if \( s_i' \) is the state of \( \chi' \) at the leaves of \( B_i \).

As \( b_i \geq 2 \), the assumption on \( \pi \) implies \( s_i = s_i' \). Thus, \( \chi \) and \( \chi' \) are equivalent characters. \( \square \)
Remark 12. If \( \pi \) is the uniform distribution (i.e. we consider the \( \kappa \)-state fully symmetric model), then we have \( \mathbb{P}_T(\chi|\Theta) = \mathbb{P}_T(\chi'|\Theta) \) if and only if \( \sigma(\chi) = \sigma(\chi') \). Indeed, in this case if we consider any permutation \( g \) of the set of states \( S \), the polynomials \( x_\chi - x_{g\cdot\chi} \) are linear phylogenetic invariants for any tree (see [3]), where \( g \cdot \chi \) stands the corresponding permutation of states at the leaves. But these polynomials can also be rewritten as \( x_\chi - x_{\chi'} \) for \( \sigma(\chi) = \sigma(\chi') \).

Examples: \( n = 3 \) and \( n = 4 \)

- For \( n = 3 \), Lemma 11 gives the following. If \( \kappa \geq 3 \) and we consider three different states \( x, y, z \) and another set of three different states \( x', y', z' \), the linear invariants obtained in Lemma 11 are:

  \[
  \tilde{x}_{xyz} - \tilde{x}_{x'y'z'} \\
  \tilde{x}_{xxz} - \tilde{x}_{xxx} \\
  \tilde{x}_{xyz} - \tilde{x}_{xxx} \\
  \tilde{x}_{yxx} - \tilde{x}_{zxx}
  \]

- For \( n = 4 \), Lemma 11 gives the following. If \( \kappa \geq 4 \) and we consider four different states \( x, y, z, w \) and another set of four different states \( x', y', z', w' \), the linear phylogenetic invariants of Lemma 11 are:

  \[
  \tilde{x}_{xyzw} - \tilde{x}_{x'y'z'w'} \\
  \tilde{x}_{xxyz} - \tilde{x}_{xxx'y'} \\
  \tilde{x}_{xxxy} - \tilde{x}_{xxxy'}
  \]

and the analogous invariants obtained for the other partitions of \([4]\) involving singletons.

There are several ways to construct linear invariants from smaller trees and a systematic way to find model invariants for certain models with stationary distribution has been described in [8]. The most immediate one is by marginalization, see the following lemma. If \( i \) is a leaf of \( T \), we call \( T_i \) the tree obtained by removing leaf \( i \) and its incident edge, and suppressing the resulting degree–2 vertex if the interior node had degree 3.

Lemma 13. Let \( i \) be a leaf of a phylogenetic \([n]\)-tree \( T \) and let \( T_i \) be the corresponding tree. Let \( l \) be a linear homogeneous map \( l: \mathbb{R}^{\kappa^n} \to \mathbb{R} \) and let \( M_i : \mathbb{R}^{\kappa^n} \to \mathbb{R}^{\kappa^n-1} \) be the marginalization map at leaf \( i \). If \( l(p_i) = 0 \) for any distribution \( p_i \) from a Markov process on the tree \( T_i \), then \( (l \circ M_i)(p) = 0 \) for any distribution \( p \) that comes from a Markov process on the tree \( T \).

Proof. To prove this lemma one just needs to observe that for any distribution \( p \) coming form a Markov process on \( T \), \( M_i(p) \) is a distribution on \( T_i \) that comes from the Markov process that at each edge \( e \) has the same transition matrices as \( e \) had on the tree \( T \).

Another way that is particular to RC/El model is described in the following lemma.

Lemma 14. [Extension lemma] Let \( \Delta = \sum_\chi a_\chi x_\chi \) be a linear invariant for an \([n]\)-tree \( T \) evolving under the RC model.
(a) Let $T'$ be the tree obtained by adding a new edge at the pendant edge of leaf $n$ of $T$. Let $s$ be a new state not involved in $\Delta$ (that is, $a_\chi = 0$ if $\chi$ contains $s$). Then,

$$
\sum_\chi a_\chi x_\chi
$$

is a linear invariant for $T'$.

(b) Let $T'$ be the tree obtained by attaching a tree $\bar{T}$ of $m + 1$ leaves to the pendant edge of leaf $n$. Let $\mu$ be a character on $m$ leaves for which $a_\chi = 0$ if $\chi$ contains some state in $\mu$ (that is, $\Delta$ does not involve the states of $\mu$ at any leaf). Then $\sum_\chi a_\chi x_\chi$ is a linear phylogenetic invariant for $T'$ (where $\mu$ stands for states $\chi$ at the first $n$ leaves and states $\mu$ at the other $m$ leaves).

(c) Let $T$ be the star tree, and let $\mu$ be a character on $m$ leaves for which $a_\chi = 0$ if $\chi$ contains some state in $\mu$. Then, for the star tree $T'$ with $n + m$ leaves evolving under the RC model, $\sum_\chi a_\chi x_{\chi|\mu}$ is a linear phylogenetic invariant.

**Proof.** (a) By Lemma[10], we only need to check that $\sum_\chi a_\chi x_\chi$ vanishes for the distributions generated with $\Theta = \Theta_F$ where $F$ is a full subforest of $T'$. Let $F$ be a full subforest of $T$. We denote by $\theta_{F|T}$ the corresponding probabilities at the edges of $T$ and we denote by $\Delta(\theta_{F|T})$ the value of $\Delta$ evaluated at $\mathbb{P}_{T,\Theta_F}$.

If $F$ contains a tree with the new edge $e'$ on it, then, for all $\chi$ involved in $\Delta$, we have $\mathbb{P}_T(\chi s|\Theta_F) = 0$ (because $s$ is a state not involved in $\Delta$) and the $\sum_\chi a_\chi x_\chi$ trivially vanishes. If $F$ does not contain the edge $e'$, then the new leaf is a singleton in $F$. In this case we have $\mathbb{P}_T(\chi s|\Theta_F) = \pi_s \mathbb{P}_T(\chi|\Theta_{F|T})$. Therefore $\sum_\chi a_\chi x_\chi$ evaluated at $\mathbb{P}_{T',\Theta_F}$ is $\Delta(\theta_{F|T})$ multiplied by $\pi_s$, so it vanishes as well.

(b) If $\bar{T}$ is binary, then the addition of $\bar{T}$ can be obtain by successively adding cherries to $T$. So, assume that we have added one cherry as in (a), so that we have assigned state $s$ to the new leaf $l_{n+1}$, and now we add a new cherry to the edge leading to $l_{n+1}$. Now the new state $s'$ that we consider for the new leaf now can be allowed to be equal to the state $s$ as soon as $s'$ differs from the states that appear in $\Delta$. Indeed, if $s' = s$, there might be forests containing the new cherry, but all of them give probability zero for the states appearing in the polynomial except if the forest is formed by the new cherry and other trees. For such a forest $F$ we have $\mathbb{P}(\chi ss'|\Theta_F) = \pi_s \mathbb{P}(\chi|\Theta_{F|T})$ and hence the polynomial evaluated at the parameters of this forest is $\Delta(\theta_{F|T})$ multiplied by $\pi_s$ which vanishes again.

If $\bar{T}$ is not a binary tree, then it can be also constructed from a binary tree by contracting edges. As for binary tree the polynomial is a phylogenetic invariant, so it is when we contract edges (note that if a polynomial is a phylogenetic invariant for a tree, then it is also a phylogenetic invariant for the tree $T_0$ obtained by contracting one edge $e_0$ because any collection of edge parameters at $T_0$ gives a collection of edge parameters for $\bar{T}$ by assigning $\theta_{e_0} = 0$).

(c) It follows from (b) by contracting edges. \(\square\)

5. Phylogenetic Mixtures

So far, we have found some linear polynomials that turn out to be either model invariants or topology invariants. But we were not able to say whether these invariants actually generate the space of linear phylogenetic invariants for a tree $T$. On the other hand, it would be interesting to know whether a distribution where all these linear invariants vanish is actually a linear combination from distributions on a tree or a mixture of trees. To this end, one defines the space of mixtures on a tree.[19]

**Definition 15.** Fix a distribution $\pi$ on the set of states. Given a particular tree $T$, we denote by $\mathbb{P}_{T,\Theta}$ the distribution of a RC model with parameters $\pi, \Theta$ on $T$. We define the *space of mixtures*
on $T$ as
\[
\mathcal{D}_T^\pi = \left\{ p = \sum \nu_i \mathbb{P}_{T_i, \Theta_i} \mid \sum \lambda_i = 1 \right\}.
\]

If $\mathcal{T}$ is the set of phylogenetic trees on $[n]$, we define the space of phylogenetic mixtures on $[n]$ as
\[
\mathcal{D}_T^\pi = \left\{ p = \sum \lambda_i \mathbb{P}_{T_i, \Theta_i} \mid \sum \lambda_i = 1, T_i \in \mathcal{T} \right\}.
\]

When $\{p_i\}_{i \in I}$ is a set of points in an affine linear space, we denote by $\langle p_i \mid i \in I \rangle_a$ the linear span of these points, that is, the set of points $q = \sum \lambda_i p_i$ with $\sum \lambda_i = 1$ (we put the subindex $a$ in order to distinguish this affine linear span from the usual linear span of vectors). Note that the space of phylogenetic mixtures is the affine linear variety
\[
\mathcal{D}_T^\pi = \left\{ p \mid p = \mathbb{P}_{T, \Theta}, T \in \mathcal{T} \right\}_a,
\]
and both lie inside the hyperplane
\[
H = \left\{ x = (x_\chi)_\chi \in \mathbb{R}^N \mid \sum_{\chi \in Ch([n, \kappa])} x_\chi = 1 \right\}.
\]

Strictly speaking, for applications in phylogenetics it is only relevant to consider points in $\mathcal{D}_T^\pi$ (or $\mathcal{D}_T^\pi$) that are actually distributions. In other words, one should be mainly interested in convex combinations of the points $\mathbb{P}_{T, \Theta}$:
\[
\left\{ p = \sum \lambda_i \mathbb{P}_{T_i, \Theta_i} \mid \lambda_i \geq 0, \sum \lambda_i = 1 \right\}
\text{ and }
\left\{ p = \sum \lambda_i \mathbb{P}_{T_i, \Theta_i} \mid \lambda_i \geq 0, \sum \lambda_i = 1, T_i \in \mathcal{T} \right\}.
\]

However, as the dimension of a polyhedron is the dimension of its affine hull, we focus on computing the dimension of $\mathcal{D}_T^\pi$ and $\mathcal{D}_T^\pi$.

For any distribution $\pi$, we denote by $L^\pi$ the vector space of linear model invariants and by $L_T^\pi$ the space of all linear phylogenetic invariants for a tree $T$. The orthogonal subspace of $L^\pi$ (respectively $L_T^\pi$) shall be denoted by $E^\pi$ (respectively $E_T^\pi$), that is, $E^\pi$ is the set of vectors in $\mathbb{R}^N$ where all the linear model invariants vanish and $E_T^\pi$ the set of vectors where all the linear phylogenetic invariants for $T$ vanish. In other words, $E_T^\pi$ and $E^\pi$ are spanned by the following vectors of distributions:
\[
E_T^\pi = \left\langle \mathbb{P} \mid p = \mathbb{P}_{T, \Theta} \right\rangle = \left\{ \sum \lambda_i \mathbb{P}_{i} \mid p_i = \mathbb{P}_{\Theta_i}, \lambda_i \in \mathbb{R} \right\},
\]
\[
E^\pi = \left\langle \mathbb{P} \mid p = \mathbb{P}_{T, \Theta}, T \in \mathcal{T} \right\rangle = \left\{ \sum \lambda_i \mathbb{P}_{i} \mid p_i = \mathbb{P}_{T_i, \Theta_i}, \lambda_i \in \mathbb{R}, T \in \mathcal{T} \right\}.
\]

Note that when we use $p \in \mathbb{R}^N$ as a vector, we use the notation $\bar{p}$ to distinguish it from its use as an affine point in $\mathbb{R}^N$. Then the following equalities are clear
\[
\mathcal{D}_T^\pi = E_T^\pi \cap H, \quad \mathcal{D}^\pi = E^\pi \cap H.
\]
Therefore, studying phylogenetic mixtures (on \([n]\) or on a tree) is equivalent to studying linear phylogenetic invariants (only model invariants or together with topology invariants). Note that due to Lemma 3, it is clear that
\[
\mathcal{D}^E_T = \{ p = P_{T, \Theta_F} | F \in \mathcal{F}_T \}, \quad \mathcal{D}^{\pi}_T = \{ p = P_{T, \Theta_F} | T \in \mathcal{T}, F \in \mathcal{F}_T \}
\]
(see also \[13\] Prop. 10).

In this section we compute the dimension of the spaces of phylogenetic mixtures.

5.1. Model invariants and phylogenetic mixtures. We fix \(n \geq 4\) throughout this section. We call \(\Sigma_k\) the set of partitions of \([n]\) of size at most \(k\) (note that if \(k \geq n\), this is the whole set of partitions of \([n]\)). If \(\sigma\) is a partition of \([n]\) compatible with a tree \(T\), we consider the minimal forest \(F_\sigma\) associated to \(\sigma\). The point \(q_\sigma\) will be briefly denoted as \(q_\sigma\). Note that if \(\sigma\) is compatible with two trees \(T, T'\), and \(F_\sigma, F_{\sigma'}\) are the corresponding minimal forests, then \(q_{\sigma'} = q_{\sigma'}\) (in particular \(q_\sigma\) is well defined).

**Theorem 16.** Let \(\pi\) be the uniform distribution or a generic distribution. For each partition \(\sigma\) of \([n]\) of size \(\leq \kappa\), consider the corresponding point \(q_\sigma\). Then
\[
\mathcal{D}^\pi = \{ q_\sigma \mid \sigma \in \Sigma_\kappa \}_{\mathbb{R}}.
\]
Moreover, \(\mathcal{D}^\pi\) has dimension \(|\Sigma_\kappa| - 1\) (which equals \(B_n - 1\) if \(\kappa \geq n\)).

The inclusion \(\{ q_\sigma \mid \sigma \in \Sigma_\kappa \}_{\mathbb{R}} \subseteq \mathcal{D}^\pi\) clearly holds (and if \(\kappa \geq n\), the other inclusion easily follows). The idea for the proof of the other inclusion is to use \(\mathcal{D}^\pi = E^\pi \cap H\), bound the dimension of \(E^\pi\) from above by a quantity \(d\) and then prove that the set of points \(q_\sigma\) span an affine linear variety of dimension \(d - 1\). So let us first prove that the points \(q_\sigma\) considered above are affinely independent.

**Lemma 17.** The affine linear variety \(\{ q_\sigma \mid \sigma \in \Sigma_\kappa \}_{\mathbb{R}}\) has dimension \(|\Sigma_\kappa| - 1\) for any distribution \(\pi\) (with positive entries).

We give the coordinates of the points \(q_\sigma\) for \(n = 4\) shortly, see Example 22.

**Proof.** We need to prove that the points \(q_\sigma\) in this set are affinely linearly independent, that is, if we have a linear combination
\[
\sum_{\sigma \in \Sigma_\kappa} \lambda_\sigma q_\sigma = 0
\]
with \(\sum_\sigma \lambda_\sigma = 0\), then we need to prove that the coefficients \(\lambda_\sigma\) are zero. We proceed by induction on \(m = \min\{n, \kappa\}\). Note that as all partitions of \([n]\) are of size at most \(n\), \(\Sigma_k\) equals the set \(\Sigma_m\) of partitions of size at most \(m\).

If \(m = 1\), then \(\Sigma_k\) contains a single element and there is nothing to prove. Assume that \(m \geq 2\) and consider a linear combination as in Eqn. (5).

Note that the coordinate \(\tilde{x}_\chi\) of \(q_\sigma\) is zero if \(\sigma\) does not refine \(\sigma(\chi)\). Let \(\tilde{x}_\chi\) be a coordinate such that \(\sigma(\chi)\) has the maximum size \(m\). Then \(\tilde{x}_\chi\) is different from zero only for \(q_{\sigma(\chi)}\) (because the other points \(q_\sigma\) correspond to partitions that do not refine \(\sigma(\chi)\)). Thus, \(\lambda_{\sigma(\chi)} = 0\) and hence in (5) we have \(\lambda_\sigma = 0\) for all \(\sigma\) of size \(m\). Thus, we are left with a linear combination such as
\[
\sum_{\sigma \in \Sigma_{m-1}} \lambda_\sigma q_\sigma = 0, \quad \sum_{\sigma \in \Sigma_{m-1}} \lambda_\sigma = 0.
\]
The result follows by the induction hypothesis.

We compute the dimension of \(E^{\pi_U}\) for the uniform distribution \(\pi_U\):
Lemma 18. If \( \pi_U \) is the uniform distribution, then the set of linear model invariants is spanned by the set of polynomials \( x_\chi - x_\chi' \) for \( \sigma(\chi) = \sigma(\chi') \). In particular, the set of vectors \( E^\pi_U \) where the model invariants vanish has dimension equal to \( |\Sigma_n| \).

**Proof.** For the uniform distribution, each polynomial \( x_\chi - x_\chi' \) for \( \sigma(\chi) = \sigma(\chi') \) is clearly a model invariant (see Remark 22). Thus the set of vectors \( E^\pi_U \) where these polynomials vanish has dimension less than or equal to \(|\Sigma_n|\). The set of points considered in Lemma 17 for \( \pi_U \) is contained in \( E^\pi_U \cap H \), and hence (as \( H \) is an equation linearly independent with the previous polynomials), the dimension of \( E^\pi_U \) is \(|\Sigma_n|\). It follows that the inclusion \( E^\pi_U \subseteq \{ x \in \mathbb{R}^N | x_\chi = x_\chi' \text{ if } \sigma(\chi) = \sigma(\chi') \} \) is actually an equality and the set of model invariants is spanned by the polynomials \( x_\chi - x_\chi' \) for \( \sigma(\chi) = \sigma(\chi') \). \( \square \)

Now we bound the dimension of \( E^\pi \) from above by the dimension of \( E^\pi_U \):

Lemma 19. For a generic distribution \( \pi \), the dimension of \( E^\pi \) (respectively \( E^\pi_T \)) is less than or equal to the dimension of \( E^{\pi_0} \) (respectively \( E^{\pi_0}_T \)) for a particular distribution \( \pi_0 \).

**Proof.** We think first of the coordinates of \( \pi \) as parameters, so that we consider model invariants as linear polynomials in the variables \( x_\chi \) with coefficients in the field of rational functions \( \mathbb{R}(\pi) \) (i.e. the field of fractions of the ring of polynomials \( \mathbb{R}[\pi_1, \ldots, \pi_n] \)). The set of all model invariants is a \( \mathbb{R}(\pi_1, \ldots, \pi_n) \)-vector space. Consider a basis \( l_1, \ldots, l_t \) of this space and let \( E \) be its orthogonal subspace, \( E = \{ x \in \mathbb{R}^N | l_i(x) = 0, i = 1, \ldots, t \} \) so that \( \dim E = N - t \). When we substitute \( \pi \) by a particular value \( \pi_0 \), \( l_1, \ldots, l_t \) may not be linearly independent any more, and the corresponding space \( E^{\pi_0} \) may have dimension \( \geq \dim E \). But for a generic \( \pi \), the dimension of the corresponding space coincides with dimension of \( E \) (because \( \pi \) moves in an irreducible space). Therefore, for a generic \( \pi \) we have \( \dim E^\pi = \dim E \leq \dim E^{\pi_0} \). The same argument can be applied for \( E^\pi_T \) and \( E^{\pi_0}_T \). \( \square \)

Proof of Theorem 16. By Lemma 19, the dimension of \( E^\pi \) is less than or equal to \( \dim E^{\pi_U} \) for the uniform distribution \( \pi_U \) and the dimension of this vector space is \(|\Sigma_n| \) (by Lemma 18). Thus, \( \dim E^\pi \leq |\Sigma_n| - 1 \). Now, Lemma 17 and the inclusion
\[
\langle q_\sigma \mid \sigma \in \Sigma_n \rangle \subseteq D^\pi = E^\pi \cap H
\]
gives the claim. \( \square \)

Remark 20. Fix \( \kappa \geq n \). Then the dimension of the space of mixtures \( D^\pi \) for generic \( \pi \) is \( B_n - 1 \) and coincides with the dimension of the space of mixtures for \( n \) states for generic \( \pi \); it also coincides with the dimension of the space of mixtures on the fully symmetric model (uniform \( \pi \)) with \( \kappa \) states. This is an immediate consequence of Theorem 16 once one notes that if \( \kappa \geq n \), the dimension of \( D^\pi \) does not depend on \( \pi \) nor \( \kappa \) (if \( \pi \) is a generic distribution). The dimension of the space of mixtures for the Jukes-Cantor model has already been computed in [3] and in [7].

Remark 21. Although we do not give the explicit linear model invariants that generate \( L^\pi \) for all distributions \( \pi \) (not even for generic \( \pi \)'s), these can be obtained from Theorem 16 because we explicitly give a set of linearly independent points in the space of mixtures in Lemma 17.

Then, for a particular \( \pi \), one can compute a basis of the set of equations that define this set of points. It may happen that, if \( \pi \) is not generic, this basis contains equations that do not come actually from model invariants. Hence, in order to obtain a basis of \( L^\pi \), one just needs select the equations in the basis that correspond to model invariants.

Example 22. We give here the coordinates of the points that span the spaces of mixtures on trees with \( n = 4 \) and \( \kappa = 4 \) or \( \kappa = 3 \).

For \( \kappa = 4 \) we have \(|\Sigma_4| = B_4 = 15\) and \( D^\pi = \langle q_\sigma \mid \sigma \in \Sigma_4 \rangle \). We start with 12 partitions \( \sigma \) that correspond to forests in the star tree \( T^* \). We call \( q^*_\sigma \) the point corresponding to the trivial
subforest of $T_*$ (formed by singletons). We call $q_{ij}$ the points corresponding to the full subforest of $T_*$ formed by the tree $T[i, j]$ and singletons (this gives six points, $q_{ij}$, $i < j$). Then we consider the forests formed by a subtre of three leaves $i, j, k$ and a singleton, which gives four points $q_{123}$, $q_{124}$, $q_{134}$, $q_{234}$. Finally, we denote by $q_{1234}$ the point corresponding to the forest $F = \{T_\ast\}$. To simplify notation we write the normalized coordinates $\tilde{x}_{1, \ldots, 4}$ instead of $x_{1, \ldots, 4}$. Let the space of states $S$ be $\{x, y, z, w\}$. In order to prove that the 15 points we provide are affine linearly independent, it is enough to look at the following 15 coordinates of these points:

$$\tilde{x}_{12}, \tilde{x}_{13}, \tilde{x}_{14}, \tilde{x}_{23}, \tilde{x}_{24}, \tilde{x}_{34}, \tilde{x}_{123}, \tilde{x}_{124}, \tilde{x}_{134}, \tilde{x}_{234}, \tilde{x}_{1234}, \tilde{x}_{1243}, \tilde{x}_{1342}, \tilde{x}_{2341}. $$

In Table 1 we write the coordinates of the first 12 points considered above.

If we consider the previous points plus the point $q_{1234}$ that corresponds to the forest $\{T[1, 2], T[3, 4]\}$ on the tree $T_{1234}$, then we obtain a set of linearly independent points that span $\mathcal{D}_{1234}^\pi$. In Table 2 we show the coordinates of this new point.

Now we consider the points corresponding to the forests compatible for the remaining quartets, $q_{1324}, q_{1423}$ (their coordinates are shown in Table 3). The previous points together with these two points span the space of mixtures $\mathcal{D}^\pi$. Consider now the case $\kappa = 3$. Then, according to Theorem 10, $\mathcal{D}^\pi$ has dimension 13 for generic $\pi$. Indeed, if we consider the 15 points above, then they are no longer linearly independent when the last column of the table is removed. The last 14 points suffice to span $\mathcal{D}^\pi$ in this case.

### Table 1. Linearly independent points for $\mathcal{D}_{12}$.

| $q_{ij}$ | $x_{1}$ | $x_{2}$ | $x_{3}$ | $x_{4}$ | $\tilde{x}_{12}$ | $\tilde{x}_{13}$ | $\tilde{x}_{14}$ | $\tilde{x}_{23}$ | $\tilde{x}_{24}$ | $\tilde{x}_{34}$ | $\tilde{x}_{123}$ | $\tilde{x}_{124}$ | $\tilde{x}_{134}$ | $\tilde{x}_{234}$ | $\tilde{x}_{1234}$ |
|-----------|--------|--------|--------|--------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| $q_{12}$  | 1      | 1      | 1      | 1      | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              |
| $q_{13}$  | $\frac{1}{\pi_1}$ | $\frac{1}{\pi_2}$ | 0      | 0      | $\frac{1}{\pi_3}$ | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              |
| $q_{14}$  | $\frac{1}{\pi_1}$ | $\frac{1}{\pi_2}$ | 0      | 0      | $\frac{1}{\pi_3}$ | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              |
| $q_{23}$  | $\frac{1}{\pi_1}$ | $\frac{1}{\pi_2}$ | 0      | 0      | $\frac{1}{\pi_3}$ | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              |
| $q_{24}$  | $\frac{1}{\pi_1}$ | $\frac{1}{\pi_2}$ | 0      | 0      | $\frac{1}{\pi_3}$ | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              |
| $q_{34}$  | $\frac{1}{\pi_1}$ | $\frac{1}{\pi_2}$ | 0      | 0      | $\frac{1}{\pi_3}$ | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              | 0              |

### Table 2. The new point added for tree 12|34.

| $q_{1234}$ | $\tilde{x}_{12}$ | $\tilde{x}_{13}$ | $\tilde{x}_{14}$ | $\tilde{x}_{23}$ | $\tilde{x}_{24}$ | $\tilde{x}_{34}$ | $\tilde{x}_{123}$ | $\tilde{x}_{124}$ | $\tilde{x}_{134}$ | $\tilde{x}_{234}$ | $\tilde{x}_{1234}$ |
|------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $\frac{1}{\pi_2}$ | 0               | 0               | 0               | 0               | 0               | 0               | 0               | 0               | 0               | 0               | 0               |

### Table 3. The two points added when considering the quartets 13|24 and 14|23.

| $q_{1324}$ | $\tilde{x}_{13}$ | $\tilde{x}_{14}$ | $\tilde{x}_{23}$ | $\tilde{x}_{24}$ | $\tilde{x}_{34}$ | $\tilde{x}_{132}$ | $\tilde{x}_{134}$ | $\tilde{x}_{234}$ | $\tilde{x}_{1324}$ |
|------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $\frac{1}{\pi_2}$ | 0               | 0               | 0               | 0               | 0               | 0               | 0               | 0               | 0               |

| $q_{1423}$ | $\tilde{x}_{14}$ | $\tilde{x}_{24}$ | $\tilde{x}_{34}$ | $\tilde{x}_{142}$ | $\tilde{x}_{143}$ | $\tilde{x}_{243}$ | $\tilde{x}_{1423}$ |
|------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $\frac{1}{\pi_2}$ | 0               | 0               | 0               | 0               | 0               | 0               | 0               |
In this section we compute the dimension of the space of phylogenetic mixtures on a tree and we explain whether Lake-type invariants of Proposition 8 suffice to describe the space of phylogenetic invariants. Recall that $E_T^*$ is the space of vectors where the linear phylogenetic invariants vanish. We know by Lemma 11(b) that a homogeneous linear polynomial vanishes on all distributions $\mathbb{P}_{T,\Theta}$ if and only if it vanishes on all distributions of type $\mathbb{P}_{T,\Theta_P}$ for $F$ subforest of $T$. Therefore we have $$E_T^* = \langle \tilde{q}_F | F \in \mathcal{F}_T \rangle.$$ 

Example 23. Let $n = 3$, let $T$ be the tripod tree and assume that $\kappa \geq 3$. We prove here that the vectors $\tilde{q}_F$, for $F \in \mathcal{F}_T$ are linearly independent. These vectors are: $\tilde{q}_\bullet$ corresponding to the trivial subforest, $\tilde{q}_{123}$, $\tilde{q}_{132}$, $\tilde{q}_{231}$, corresponding to forests with one singleton, and $\tilde{q}_{23}$ corresponding to the tree itself. We choose three states $x, y, z$ and we provide in Table 4 the submatrix corresponding to the coordinates $x_{xxy}, x_{xyy}, x_{xyy}, x_{xxy}, x_{xyy}, x_{xyy}$. It is clear that this submatrix has nonvanishing determinant if $\pi$ is positive.

Let $T$ be a tree on $[n]$, $n \geq 4$, and assume that leaves $n$ and $n - 1$ form a cherry $c$. Let $u$ be the interior node of this cherry, and let $e$ be the edge adjacent to $u$ and not to $n, n - 1$. Let $T'$ be the subtree $T - \{e_n, e_{n-1}\}$. We denote by $\mathcal{F}_c$ the set of full subforests of $T$ that contain a tree with the cherry $c = \{e_n, e_{n-1}\}$. For any leaf $l$ we let $\mathcal{F}_l$ be the set of full subforests of $T$ that contain $l$ as a singleton. Then $\mathcal{F}_T$ is the disjoint union of $\mathcal{F}_c$ and $\mathcal{F}_{n-1} \cup \mathcal{F}_n$.

Lemma 24. We have isomorphisms of vector spaces:

$$\langle \tilde{q}_F | F \in \mathcal{F}_c \rangle \cong \langle \tilde{q}_G | G \in \mathcal{F}_T \rangle$$

$$\langle \tilde{q}_F | F \in \mathcal{F}_G \rangle \cong \langle \tilde{q}_G | G \in \mathcal{F}_T \rangle.$$ 

Proof. Let $V_n$ be the vector space $\langle \tilde{q}_F | F \in \mathcal{F}_n \rangle$. For any state $s \in S$ we denote by $f^s$ the projection map from $\mathbb{R}^n$ to the subspace $R_s$ corresponding to coordinates $x_{\chi_1, \ldots, \chi_{n-1}}$, so that we can view $\mathbb{R}^n$ as the direct sum $R_{s_1} \oplus \cdots \oplus R_{s_\kappa}$. For a vector $v \in \mathbb{R}^n$ we denote by $(f^1(v), \ldots, f^\kappa(v))$ the decomposition of $v$ according to this direct sum. Note that if $F \in \mathcal{F}_n$, then $\mathbb{P}_T(\chi_1, \ldots, \chi_n | \Theta_F) = \pi_{\chi_1} \mathbb{P}_T(\chi_1, \ldots, \chi_{n-1} | \Theta_{F | T'})$. In particular, we have $f^s(\tilde{q}_F) = \pi_s \tilde{q}_\mathcal{F}_c$, for any $s \in S$ and $\tilde{q}_\mathcal{F}_c = (\pi_{s_1} \tilde{q}_{\mathcal{F}_c 1}, \ldots, \pi_{s_\kappa} \tilde{q}_{\mathcal{F}_c \kappa})$.

We prove here that (for any $s \in S$) the linear map $f^s$ is an isomorphism between $V_n$ and the target vector space. First of all, the linear map $f^s_{|V_n}$ is injective. Indeed, if $f^s_{|V_n}(v) = 0$ for a certain $v = \sum_{F \in \mathcal{F}_n} \lambda_F \tilde{q}_F$, then $0 = \sum_{F \in \mathcal{F}_n} \lambda_F f^s(\tilde{q}_F) = \sum_{F \in \mathcal{F}_n} \lambda_F \pi_s \tilde{q}_\mathcal{F}_c$, and hence $\sum_{F \in \mathcal{F}_n} \lambda_F \tilde{q}_\mathcal{F}_c = 0$. This implies that $v = (0, \ldots, 0)$ in $R_{s_1} \oplus \cdots \oplus R_{s_\kappa}$ and so $f^s_{|V_n}$ is an injective linear map.

We prove that the image of $f^s_{|V_n}$ is $\langle \tilde{q}_G | G \in \mathcal{F}_T \rangle$. From the above, one can easily see that $\text{Im} f^s_{|V_n}$ is contained in $\langle \tilde{q}_G | G \in \mathcal{F}_T \rangle$. Now for any $G \in \mathcal{F}_T$, we consider $\tilde{G}$ the full subforest of $T$ defined by: the singleton $\{n\}$, $t$ for any $t \in G$ not containing $e$ nor $u$, $t \cup e_{n-1}$ if there is

| $x_{xxy}$ | $x_{xyy}$ | $x_{xyy}$ | $x_{xxy}$ | $x_{xyy}$ |
|-----------|-----------|-----------|-----------|-----------|
| $\tilde{q}_\bullet$ | $\pi_x$ | $\pi_x$ | $\pi_x$ | $\pi_x$ |
| $\tilde{q}_{123}$ | $\pi_x^3$ | $\pi_x^2$ | $\pi_x^2$ | $\pi_x$ |
| $\tilde{q}_{132}$ | $\pi_x^2$ | $\pi_x$ | $\pi_x$ | $\pi_x$ |
| $\tilde{q}_{231}$ | $\pi_x$ | $\pi_x$ | $\pi_x$ | $\pi_x$ |

Table 4. Table of example 23.
t ∈ G containing e, and the singleton \( \{n-1\} \) if G contains the singleton \( \{u\} \). In this way we have \( \bar{G}_{|T'} = G \) and \( \bar{q}_G = \frac{1}{\pi_G} f_{|V_0} p_{|V_0} \bar{q}_G \in \text{Im} f_{|V_0}^\ast \), so the other inclusion is proved.

As far as the second isomorphism is concerned, we consider the subspace \( L \subset \mathbb{R}^n \) given by coordinates of type \( \chi \chi_{1\cdots n-3}^s \) for any \( \chi_1, \ldots, \chi_{n-2}, s \) in S. We have \( \mathbb{R}^n = L \oplus L^\perp \) and if \( f \) denotes the projection to \( L \), then any vector \( v \) can be decomposed as \( (f(v), v-f(v)) \). If \( F \in \mathcal{F} \), then \( \mathbb{P}^T(\chi \chi_{1\cdots n-1}(\Theta F)) \) is zero if \( \chi_{n-1} \neq \chi_n \) and is equal to \( \mathbb{P}^T(\chi \chi_{1\cdots n}(\Theta F')) \) if \( \chi_{n-1} = \chi_n \). Hence, if \( F \in \mathcal{F} \) we have \( \bar{q}_F = (f(\bar{q}),0) = (\bar{q}_{F'|T'},0) \). Now we prove that \( f_{|V_0} \) is a injective. Moreover the image of this map is included in the subspace \( \langle \bar{q}_G | G \in \mathcal{F}_T' \rangle \). For any \( G \in \mathcal{F}_T' \) we consider the full subforest \( G \) of \( T \) defined by: the trees in \( G \) that do not contain \( e \), \( t \cup c \) if \( t \) contains \( e \), and the cherry \( c \) if \( G \) contains the singleton \( \{u\} \). Therefore we have \( \bar{G}_{|T'} = G \) and \( \bar{q}_G = f_{|V_0} \bar{q}_G \in \text{Im} f_{|V_0}^\ast \).

**Theorem 25.** Let \( T \) a phylogenetic tree on \( n \) leaves, \( n \geq 3 \), evolving under the EI/RC model for any distribution \( \pi \) on \( \kappa \geq 3 \) states. Then, the dimension of the affine space of phylogenetic mixtures on \( T \) is \( |\mathcal{F}_T| - 1 \) (i.e. \( \dim D^*_T = |\mathcal{F}_T| - 1 \)). In the case where \( T \) is binary, \( \dim D^*_T \) is the Fibonacci number \( F_{2n-1} \).

**Proof.** We proceed by induction on \( n \). The statement of the theorem is equivalent to \( \dim E^*_T = |\mathcal{F}_T| \).

The case \( n = 3 \) is handled by Example 23. For \( n \geq 4 \), suppose that \( T \) is a binary tree. We may assume that the statement is true for trees with strictly less than \( n \) leaves. We suppose that \( n \) and \( n-1 \) form a cherry and adopt the notation fixed above. Then we have that

\[
E^*_T = \langle \bar{q}_F | F \in \mathcal{F}_T \rangle = \langle \bar{q}_F | F \in F_{n-1} \cup F_n \rangle + \langle \bar{q}_F | F \in \mathcal{F}_T \rangle.
\]

Note that \( \langle \bar{q}_F | F \in \mathcal{F}_{n-1} \cup \mathcal{F}_n \rangle \) equals \( \langle \bar{q}_F | F \in \mathcal{F}_{n-1} \rangle + \langle \bar{q}_F | F \in \mathcal{F}_n \rangle \). We know that \( \langle \bar{q}_F | F \in \mathcal{F}_{n-1} \rangle \) and \( \langle \bar{q}_F | F \in \mathcal{F}_n \rangle \) have dimension \( |\mathcal{F}_{T'}| \) by Lemma 24 and the induction hypothesis. These subspaces intersect in \( \langle \bar{q}_F | F \in F_{n-1} \cap \mathcal{F}_n \rangle \). By Lemma 24 (applied twice) and the induction hypothesis, this linear space has dimension \( |\mathcal{F}_{T''}| \) where \( T'' \) is a tree on \( n-2 \) leaves. Therefore, using Grassmann’s formula (\( \dim(U + W) = \dim U + \dim W - \dim(U \cap W) \) for subspaces \( U, W \) of a vector space) we have that \( \dim(\langle \bar{q}_F | F \in \mathcal{F}_{n-1} \rangle + \langle \bar{q}_F | F \in \mathcal{F}_n \rangle) = |\mathcal{F}_{T'}| + |\mathcal{F}_{T''}| - |\mathcal{F}_{T''}|. \)

As all of these trees are binary, this dimension equals the Fibonacci number \( F_{2n-2} = F_{2n-3} + F_{2n-3} = F_{2n-5} \).

On the other hand, by Lemma 24 and the induction hypothesis, \( \langle \bar{q}_F | F \in \mathcal{F}_T \rangle \) has dimension \( |\mathcal{F}_{T'}| = F_{2n-3} \). Let us prove now that \( \langle \bar{q}_F | F \in \mathcal{F}_T \rangle \) and \( \langle \bar{q}_F | F \in \mathcal{F}_{n-1} \cup \mathcal{F}_n \rangle \) only intersect in the zero vector. Let \( v \) be a vector in the intersection,

\[
v = \sum_{F \in \mathcal{F}_{n-1} \cup \mathcal{F}_n} \lambda_F \bar{q}_F = \sum_{G \in \mathcal{F}_T} \mu_G \bar{q}_G.
\]

Looking at the right-hand side we see that all the coordinates of \( v \) of type \( \chi \chi_{1\cdots n-3}^s \) for \( s \neq s' \) are zero. Let us fix \( \chi_1, \ldots, \chi_{n-2}, s \in S \) and we shall prove that the coordinate \( \chi \chi_{1\cdots n-3}^s \) of \( v \), \( \chi \chi_{1\cdots n-3}^s(\chi) \), is 0. Let us split the sum \( \sum_{F \in \mathcal{F}_{n-1} \cup \mathcal{F}_n} \lambda_F \bar{q}_F + \sum_{H \in \mathcal{F}_T} \lambda_H \bar{q}_H \). We denote by \( \mathcal{F}_T' \) the restriction of a forest \( F \) to \( T' \). Note that

\[
x_{\chi_1 \cdots n-2}^s(\chi) = \pi_s x_{\chi_1 \cdots n-2}^s \left( \sum_{F \in \mathcal{F}_{n-1}} \lambda_F \bar{q}_F \right) + \pi_s x_{\chi_1 \cdots n-2}^s \left( \sum_{H \in \mathcal{F}_T} \lambda_H \bar{q}_H \right).
\]
For each \( \alpha \in s \) we denote by \( a(\alpha) \) the value of the coordinate \( x_{\chi_1 \ldots \chi_{n-2} \alpha} \) of \( \sum_{F \in \mathcal{F}_{n-1}} \lambda_F \vec{q}_F \) and by \( b(\alpha) \) the value of this coordinate at \( \sum_{H \in \mathcal{F}} \lambda_H \vec{w} \). We want to prove that \( a(s) + b(s) = 0 \). Consider \( s' \) and \( s'' \) states in \( S \) different from \( s \) (this is possible because \( \kappa \geq 3 \)). As \( 0 = x_{\chi_1 \ldots \chi_{n-2}s''}(v) = \pi_{s''}a(s) + \pi_{s''}b(s') \), \( 0 = x_{\chi_1 \ldots \chi_{n-2}s''}(v) = \pi_{s''}a(s') + \pi_{s''}b(s) \), \( 0 = x_{\chi_1 \ldots \chi_{n-2}s''}(v) = \pi_{s''}a(s') + \pi_{s''}b(s') \), and \( 0 = x_{\chi_1 \ldots \chi_{n-2}s''}(v) = \pi_{s''}a(s') + \pi_{s''}b(s'') \), we have

\[
a(s) + b(s) = -\frac{\pi_{s'}}{\pi_{s'}}(b(s') + a(s')) = \frac{\pi_{s'}}{\pi_{s'}}(b(s'') + a(s'')).
\]

But now we use the analogous relations between \( a(s), a(s'''), b(s), b(s'') \):

\[
0 = x_{\chi_1 \ldots \chi_{n-2}s'''}(v) = \pi_{s'''}a(s) + \pi_{s'''}b(s''),
\]

in order to obtain that \( a(s) + b(s) = -\frac{\pi_{s'''}}{\pi_{s'''}}(b(s'') + a(s'')) \). Therefore, \( a(s) + b(s) = -a(s) - b(s) \) and this quantity vanishes.

Summing up, applying Grassmann’s formula again, we have \( \langle \vec{q}_F | F \in \mathcal{F}_{n-1} \cup \mathcal{F}_n \rangle \cap \langle \vec{q}_F | F \in \mathcal{F}_c \rangle = 0 \) and

\[
dim E_T^\pi = \dim(\langle \vec{q}_F | F \in \mathcal{F}_{n-1} \rangle + \langle \vec{q}_F | F \in \mathcal{F}_n \rangle) + \dim(\langle \vec{q}_F | F \in \mathcal{F}_c \rangle).
\]

We have already seen that the first term is equal to \( F_{2n-2} \). The second term is equal to \( F_{2n-3} \) by Lemma 24 and the induction hypothesis. Therefore \( \dim E_T^\pi = F_{2n-1} = |\mathcal{F}_T| \).

Let us assume now that \( T \) is not binary. We already know that \( E_T^\pi = \langle \vec{q}_F | F \in \mathcal{F}_T \rangle \) and we only need to check that the vectors \( \vec{q}_F \), \( F \in \mathcal{F}_T \), are linearly independent. As the forests in \( T \) are also subforests of any binary tree that refines \( T \), these vectors are linearly independent by the binary tree case proved above.

We prove now that Lake–type invariants suffice to define the space of linear topology invariants of a tree. We first need a combinatorial lemma.

**Lemma 26.** For any phylogenetic tree on \([n]\) and any partition \( \sigma \) that is incompatible with \( T \) there exist two blocks \( B, B' \) of \( \sigma \) and leaves \( x \in B, x' \in B' \) and an interior vertex \( v \) of \( T \) in the path connecting \( x \) and \( x' \) for which the following holds:

For each leaf \( l \) of \( T \) in the same connected component of \( T - v \) as \( x, l \in B \) or \( \{l\} \in \sigma \).

For each leaf \( l \) of \( T \) in the same connected component of \( T - x' \) as \( x, l \in B' \) or \( \{l\} \in \sigma \).

**Proof.** First suppose that \( \sigma \) has no singleton blocks. Let us say that an edge \( e = \{u, v\} \) of \( T \) is terminating if:

(i) all the leaves of \( T \) that are in the subtree \( t_e \) of \( T - v \) containing \( u \) are contained in a single block of \( \sigma \) (say, \( B_1 \)), and

(ii) at least two of the other subtrees of \( T - v \) contain elements of \([n]\) not in \( B_1 \).

For each such terminating edge \( e \) delete the pendant subtree \( t_e \) from \( T \) and label \( u \) by \( B_1 \). Let \( T' \) be the resulting tree. This tree \( T' \) has at least four leaves (since \( \sigma \) is incompatible with \( T \)) and so \( T' \) has a cherry (two leaves that are adjacent to a shared vertex \( v \)). This vertex \( v \) and the label sets of the incident leaves (\( B \) and \( B' \)) then satisfies the property claimed in the lemma. The extension to allow \( \sigma \) to have singleton blocks is now straightforward – we can simply delete them first, repeat the argument above, and add them in afterwards.

**Corollary 27.** If \( \pi_U \) is the uniform distribution and \( \kappa \geq n \), then the Lake–type invariants of Proposition 8 and model invariants generate the space of linear phylogenetic invariants for \( T \).

**Proof.** By Lemma 18 the space of model invariants \( L \) is spanned by the polynomials \( x_{\chi} - x_{\chi'} \) for \( \sigma(\chi) = \sigma(\chi') \) and has dimension \( \kappa^n - |\Sigma_n| \) (because \( \kappa \geq n \)). We also have that \( \dim L_T = \kappa^n - \dim E_T^{\pi_U} = \kappa^n - |\mathcal{F}_T| = \kappa^n - (|\Sigma_n| - |\Inc(T)|) \) and \( \dim L = \kappa^n - \dim E_T^{\pi_U} = \kappa^n - |\Sigma_n| \). Hence,
we have \( \dim L_T/L = \dim L_T - \dim L = |\text{Inc}(T)| \). So we need to prove that Lake’s invariants give a set of \( |\text{Inc}(T)| \) linearly independent vectors in \( L_T/L \).

Note that in \( L_T/L \) we can work with polynomials on indeterminates \( x_\sigma, \sigma \in \Sigma_n \).

Let us prove that, if \( \sigma \) is an incompatible partition on \( T \), then \( x_\sigma \) is a linear combination of \( x_{\sigma'} \) for compatible partitions \( \sigma' \) of size greater than \( |\sigma| \). To this end, we proceed by induction on \( m = n - |\sigma| \).

If \( m = 0 \) or 1, then \( \sigma \) is convex on \( T \) and there is nothing to prove. Let \( m \geq 2 \) and assume that we have proved the statement for partitions of size less than \( m \). Let \( \sigma = \{ B_1, \ldots, B_s \} \) and we call \( s_1, \ldots, s_r \) the states associated to \( \sigma \). Assume first that \( \sigma \) has no singletons. Then, according to Lemma 23 we can find two blocks of \( \sigma \), say \( B_1, B_2 \), an interior vertex \( v \) for which all leaves in one of the subtrees \( T'_1 \) of \( T - v \) are in \( B \), and all leaves in one of the other subtrees \( T'_2 \) of \( T - v \) are in \( B_2 \). We write \( l'_1 \) for the set of leaves in \( T'_1 \) so that if \( B \) is the disjoint union of \( l'_1 \) and some other set \( l \). Let \( E \) be the event that leaves from \( B_i \) are in state \( s_i \) for \( i \geq 3 \), leaves in \( l_1 \) are in state \( s_1 \) and leaves in \( l_2 \) are in state \( s_2 \). As the fully symmetric model satisfies the partial separability property (PS) and as \( |\sigma| \leq n - 2 \leq \kappa - 2 \), we can consider two new states \( s'_1, s'_2 \) for the set of leaves in \( l \) and \( \kappa \) for the set of leaves in \( l' \). Thus we obtain the following linear invariant (written in terms of partitions because the states do not matter, as soon as they are different):

\[
x_{\sigma} = x_{l'_1 | l'_2 | l_1 | B_2 | \ldots | B_r} - x_{l'_1 | l_2 | B_1 | \ldots | B_r} - x_{B_1 | l_1 | l_2 | B_3 | \ldots | B_r}.
\]

Note that all partitions involved in this expression, except for \( \sigma \), have size larger than \( |\sigma| \) and we can apply the induction hypothesis to them.

If \( \sigma \) has singletons, we remove these singletons in \( T \) and \( \sigma \) obtaining a tree \( T_0 \) and a partition \( \sigma_0 \) without singletons on \( T_0 \). We apply the previous argument to \( \sigma_0 \) and \( T_0 \) to obtain a linear invariant. Then we apply the Extension Lemma 14(a) recursively to add singletons and we end up also with a linear polynomial that involves \( \sigma \) and partitions of larger size. Hence, we can apply the induction hypothesis again.

The linear invariants obtained in this way for each incompatible partition \( \sigma \) are of Lake–type and form a set of linearly independent vectors in \( L_T/L \) because they involve partitions of larger size.

**Remark 28.** Case \( \kappa = 2 \). For \( \kappa = 2 \), Theorem 25 and Corollary 27 do not apply. In this case it is already known (see [13]) that there are no topology invariants for the uniform distribution \( \pi_{t_1} \) and hence \( D_{T}^\pi = D_{T_1}^\pi \) for any tree \( T \) (see [13]). Using Lemma 19 one can actually prove that this also holds for any generic distribution \( \pi \) and this space has dimension \( |\Sigma_n| = 2^{n-1} - 1 \).

**Remark 29.** Case \( \kappa = 3 \). For \( \kappa = 3 \) and \( n = 4 \), we cannot apply Corollary 27 either. But in this case we can provide another topology invariant. We describe it in the following lemma for \( n = 4 \) but can be easily generalized for the uniform distribution to any tree by using a similar argument as in Proposition 8. Moreover, it is not difficult to see that for \( \kappa \geq 4 \) it can be derived from the Lake-type invariants.

**Lemma 30.** For the tree 12|34 and any positive distribution \( \pi \) on a set \( S \) of \( \kappa \geq 3 \) states, the polynomial

\[
\tilde{x}_{xyz} + \tilde{x}_{xyz} + \tilde{x}_{xyzz} - \tilde{x}_{xyzz} - \tilde{x}_{xyz} + \tilde{x}_{xyz}\,
\]

for any three different states \( x, y, z \in S \), is a topology invariant if \( T \) evolves under the EI/RC model.

**Proof.** According to Lemma 10 we need to prove that [12] vanishes when we evaluate it at the points \( q_F \), \( F \in F_T \). If \( F \) is a forest such that \( \sigma(F) \) does not refine any of the partitions \{1, 3\}, \{2, 4\}, then the coordinates that appear in [12] are all zero. If \( \sigma(F) \) is a refinement of \{1, 3\}, \{2, 4\}, then \( \sigma(F) \) is either \{1, 3\}, \{2, 4\}, or \{2, 4\}, \{1\}, \{3\} or the trivial forest. In both cases [12] evaluated at \( q_F \) vanishes. As the evaluation of any coordinate \( \tilde{x} \) at the
point associated to the trivial forest is one; it also vanishes in this case. The remaining cases follow form the symmetry of leaves 3 and 4 in (6).

\[\begin{align*}
\text{Remark 31. Case } \kappa = 4. & \text{ For } n = 5 \text{ not all linear topology invariants are of Lake–type.} \text{ In [7] a complete list of 17 linear invariants that generate the space of linear topology invariants is given. For example, for the fully symmetric model on the set of states } \{x, y, z, w\} \text{ (i.e Jukes-Cantor model),}
\end{align*}\]

\[x_{xyzwy} + x_{xyzuw} - x_{xyyzw} - x_{xyzzw}\]

is a topology linear invariant that cannot be described by Proposition 8.

7. Explicit linear invariants for quartet trees

In this section we assume that \(\kappa \geq 4\) and we shall deal with quartet trees and the star tree on four leaves. Note that in the previous section we gave an explicit description of linear phylogenetic invariants only when the distribution was uniform. For a generic distribution \(\pi\) we managed to compute the dimension of the set of the space of linear phylogenetic invariants, but we did not give a set of generators. We do it in this section for the case \(n = 4\) and \(\kappa \geq 4\).

\[\text{Proposition 32. Let } T \text{ be an } X\text{-tree and fix and edge } e_0 \in E(T). \text{ Let } T_0 \text{ be the tree obtained from } T \text{ by contracting edge } e_0, \text{ and let } T_1, T_2 \text{ be the two trees obtained when removing edge } e_0, \text{ such that } L(T) = L_1 \cup L_2 \text{ if } L_i = L(T_i). \text{ If } \sum a_i x_i \text{ is a linear phylogenetic invariant for } T_0 \text{ such that for any pair of nontrivial full subforests } F_1, F_2, F_1, F_2 \in F_T, \text{ we have}
\]

\[\sum a_i \mathbb{P}_{T_1}(\chi_{|L_1}|(\Theta_{F_1})) \mathbb{P}_{T_2}(\chi_{|L_2}|(\Theta_{F_2})) = 0,
\]

then \(\sum a_i x_i\) is also a linear phylogenetic invariant for \(T\).

In particular, if \(\sum a_i x_i \neq 0\), \(\sum x_i\) is a linear phylogenetic invariant for \(T_0\) such that, for any \(\chi\) with \(a_i \neq 0\), \(\chi_{|L_i}\) consists of different states for either \(i = 1\) or \(i = 2\), then \(\sum a_i x_i\) is a linear phylogenetic invariant for \(T\).

\[\text{Proof. According to lemma 10 we need to check that } \sum a_i \mathbb{P}_{T}(\chi|\Theta_F) = 0 \text{ for any full subforest of } T, \text{ so let } F \text{ be a full subforest of } T.
\]

If \(e_0\) is an edge for some tree in \(F\), then for any character \(\chi\) we have \(\mathbb{P}_{T}(\chi|\Theta_F) = \mathbb{P}_{T_0}(\chi|\Theta_{F_0})\) where \(F_0\) is the subforest of \(T_0\) obtained by contracting \(e_0\) in \(F\). As \(\sum a_i x_i\) is a linear invariant for \(T_0\) by hypothesis, we are done.

Now we assume that the trees of \(F\) do not contain \(e_0\). Then any tree in \(F\) is either a subtree of \(T_1\) or \(T_2\) and \(F\) can be decomposed into two full subforests \(F_1, F_2, F_1, F_2 \in F_T\). If \(F_2\) is the trivial subforest then \(\mathbb{P}_{T}(\chi|\Theta_F) = \mathbb{P}_{T_1}(\chi|\Theta_{F_1})\) where \(F'\) is formed by \(F_1\) and the singletons of leaves in \(T_2\), so this case is also covered by the hypothesis that \(\sum a_i x_i\) is a linear invariant on \(T_0\). The case of \(F_1\) being the trivial forest follows analogously. In particular, we obtain the last assertion of the proposition because if \(\chi_{|L_i}\) consists of different states, then the only forests \(F = F_1 \cup F_2\) for which \(\mathbb{P}_{T}(\chi|\Theta_F) = 0\) have trivial \(F_1\).

Now we assume that \(F_1, F_2\) are nontrivial subforests. In this case, for any character \(\chi\) we have \(\mathbb{P}_{T}(\chi|\Theta_F) = \mathbb{P}_{T_1}(\chi_{|L_1}|(\Theta_{F_1})) \mathbb{P}_{T_2}(\chi_{|L_2}|(\Theta_{F_2})).\) Therefore the hypothesis (1) guarantees that \(\sum a_i \mathbb{P}_{T}(\chi|\Theta_F) = 0\) for any forest \(F = F_1 \cup F_2\) of this type. \(\square\)

\[\text{Remark 33. This proposition is specially useful for quartet trees } T = AB|CD \text{ because in this case, once we have a phylogenetic invariant for the star tree on four leaves, we just need to check condition (1) when } F_1 \text{ and } F_2 \text{ are } T_{A,B} \text{ and } T_{C,D} \text{ respectively.}\]
Proposition 34. Let $x, y, z, w$ be four different states and define
\[
\beta_{x,y} = \pi_x^2 \tilde{x}_{xxy} + \pi_x \pi_y [\tilde{x}_{xyy} + \tilde{x}_{xzy}] + \pi_x \pi_y [\tilde{x}_{zzw} + \tilde{x}_{zwz} + \tilde{x}_{xww}] + \pi_y^2 \tilde{x}_{yzyw},
\]
\[
\delta_{x,y} = \pi_x^2 [\pi_x \tilde{x}_{xzz} + \pi_y \tilde{x}_{xxy} + \pi_z \tilde{x}_{xxx} + \pi_w \tilde{x}_{xxw}] + \pi_x \pi_y [\pi_x \tilde{x}_{xyy} + \pi_y \tilde{x}_{xyx} + \pi_z \tilde{x}_{xyz} + \pi_w \tilde{x}_{xyw}] + \pi_x \pi_y [\pi_x \tilde{x}_{zzw} + \pi_y \tilde{x}_{yzy} + \pi_z \tilde{x}_{zwy} + \pi_w \tilde{x}_{zww}] + \pi_y^2 [\pi_x \tilde{x}_{xzw} + \pi_y \tilde{x}_{xyw} + \pi_z \tilde{x}_{xwy} + \pi_w \tilde{x}_{xww}],
\]

Then follow are linear model invariants for quartet trees evolving under the EI/RC model:
\[
(7) \quad \pi_y \tilde{x}_{xxy} + \pi_z \tilde{x}_{xzy} = \pi_y \tilde{x}_{xyz} + \pi_z \tilde{x}_{zyz},
\]
\[
(8) \quad \pi_x \tilde{x}_{xyy} + \pi_w \tilde{x}_{xwy} = \pi_w \tilde{x}_{xyw} + \pi_x \tilde{x}_{xyy},
\]
\[
(9) \quad \beta_{x,y} = \beta_{y,x},
\]
\[
(10) \quad \delta_{x,y} = \delta_{y,x}.
\]

One obtains analogous linear model invariants by considering any permutation of the set of leaves.

Proof. From the extension Lemma 14(b) it follows that (7) and (8) are model invariants. Indeed, if we consider the star tree $T_0$ on two leaves, then it is easy to check that $\pi_y \tilde{x}_{xxy} + \pi_z \tilde{x}_{xzy} - \pi_z \tilde{x}_{xyz} - \pi_y \tilde{x}_{zyz}$ is a linear phylogenetic invariant. By identifying $T_2$ with the star tree $T_{3,4}$ on leaves 3, 4 we can apply Lemma 14(b) with $\mu = xx$ to obtain (7) for the quartet tree $T = 1234$ (because $T$ can be obtained by attaching the tripod tree $T_{1,2}$ to the edge leading to leaf 3 of $T_2$). In particular, (7) vanishes for the star tree $T_0$ on four leaves. Similarly, in order to see that (8) is a phylogenetic invariant for the star tree $T_0$, we use the phylogenetic invariant
\[
\pi_x \tilde{x}_{xxy} + \pi_w \tilde{x}_{xwy} - \pi_w \tilde{x}_{xyw} - \pi_x \tilde{x}_{xww}
\]
for the tree $T_2 = T_{1,2}$ and apply Lemma 14(b) with $\mu = yz$. By Lemma 14(c) we see that (8) is a phylogenetic invariant for the quartet tree 1234.

As we have already proved that the expressions are phylogenetic invariants for the star tree $T_0$, which is obtained by contracting the interior edge of any quartet tree, we use Proposition 32 in order to prove that (7) and (8) are also phylogenetic invariants for the other two quartet trees. Actually, if $T = 13|24$ or 14|23, then we can use directly the last assertion of Proposition 32.

We check that (9) is a phylogenetic invariant for the star tree $T_0$ using Lemma 10. Let $F$ be a full subforest of $T_0$. If $F$ is the trivial subforest, then $\beta_{x,y}$ evaluated at $\Theta_F$ is $\pi_x^2 + 6\pi_x \pi_y + \pi_y^2$. As this expression is symmetric for $x$ and $y$, (9) vanishes in this case.

If $F$ is not trivial, then $F$ is formed by a subtree $T'$ of $T_0$ of $m$ leaves and singleton trees, $m \geq 2$.

We start with $m = 2$. Using the equalities of lemma 11 we can see that $\beta_{x,y}$ is symmetric under the permutation of leaves 1, 2, and 3. Thus we only need to consider that either $T'$ is formed by the edges $\{e_1, e_2\}$ or the edges $\{e_3, e_4\}$. In the first case, when we evaluate $\beta_{x,y}$ at $\Theta_F$, only the first two summands are nonzero and the evaluation equals $\frac{\pi_y^2}{\pi_x} + \frac{\pi_y^2}{\pi_x}$. As this is symmetric in $x$ and $y$, (7) also vanishes in this case. If $T'$ consists of the edges $\{e_3, e_4\}$, then only $\beta_{xxyy}$ and $\delta_{xww}$ in the evaluation of $\beta_{x,y}$ on $\Theta_F$ are nonzero and the evaluation equals $\frac{\pi_y^2}{\pi_x} + \frac{\pi_y^2}{\pi_x}$. Again, this is symmetric in $x$ and $y$, (9) vanishes.

Now we consider $m = 3$. Let us assume first that $T'$ is formed by edges $e_1, e_2$, and $e_3$. In this case, the evaluation of $\beta_{x,y}$ at $\Theta_F$ equals $\frac{\pi_y^2}{\pi_x} = 1$ and therefore (7) vanishes. If $T'$ contains $e_4$, then all terms in the evaluation of $\beta_{x,y}$ at $\Theta_F$ are zero and (9) trivially vanishes.
If \( m = 4 \), then (9) vanishes trivially because all its terms are 0.

Therefore we have proven that (9) is a linear phylogenetic invariant for the star tree \( T_0 \).

In order to prove that it is a phylogenetic invariant for any quartet tree \( T = AB/CD \) we use Proposition 32. We need to consider the full subforest \( F = F_1 \cup F_2 \) where \( F_1 = T_{AB}, F_2 = T_{CD} \) and check that condition (\*) of Proposition 32 is satisfied. The expression \( \beta_{xy} \) evaluated at \( \Theta_F \) gives \( \pi_x \pi_y \tilde{p}_{xy} \) if \( T = 12|34 \), \( \pi_x \pi_y \tilde{p}_{xy} \) if \( T = 13|24 \), and \( \pi_x \pi_y \tilde{p}_{xy} \) if \( T = 14|23 \). In any case this value is symmetric on \( x \) and \( y \) because it actually equals 1, so (9) vanishes.

(10) arises by marginalization over the 4th leaf. Indeed, marginalizing over the four leaf at any quartet tree we obtain the tripod tree \( T_3 \) for which

\[
\pi_x^2 \tilde{x}_{xxx} + \pi_x \pi_y \tilde{x}_{xyy} + \tilde{x}_{xyx} + \tilde{x}_{yxx} + \tilde{x}_{yxz} + \tilde{x}_{yzy} + \pi_y^2 \tilde{x}_{xyz} - \left( \pi_y^2 \tilde{x}_{yxy} + \pi_x \pi_y \tilde{x}_{xyy} + \tilde{x}_{yxx} + \tilde{x}_{yxz} + \pi_x^2 \tilde{x}_{yzy} \right)
\]

is a phylogenetic invariant. As \( \bar{p}(\chi(1)\chi(2)\chi(3)) = \sum_{\chi_1} \bar{p}(\chi(1)\chi(2)\chi(3) \chi_4) \), we have \( \bar{p}_{x_1} \chi_2 \chi_3 = \sum_{\chi_4} \pi_{x_4} \bar{p}_{x_1} \chi_2 \chi_3 \chi_4 \) and we obtain the phylogenetic invariant \( \delta_{x,y} = \delta_{y,x} \) using Lemma 13.

When we make a permutation of the set of leaves, the resulting polynomials are phylogenetic invariants because we have just proven that the original ones are linear model invariants.

\[\square\]

**Remark 35.** Although one could substitute (9) by another phylogenetic invariant obtained from marginalization from a phylogenetic invariant relating \( \tilde{x}_{xy} \) and \( \tilde{x}_{yzz} \) on the tripod, this expression would have less symmetries than (9) and therefore we decided to use (9) instead.

**Theorem 36.** For any distribution \( \pi \), the space of linear model invariants \( L^n \) for \( n = 4 \) and \( \kappa \geq 4 \) is generated by the phylogenetic invariants of Proposition 32 together with \( x_\chi = x_{\chi'} \) for any \( \chi' \equiv \chi \) and has dimension \( B_4 = 15 \).

Note that this theorem is stated for any \( \pi \) without the restriction of being “generic”, so we cannot apply directly Theorem 16. For the fully symmetric model we have already seen in Remark 12 that \( x_\chi = x_{\chi'} \) are linear phylogenetic invariants if \( \sigma(\chi) = \sigma(\chi') \). In this case this set of invariants defines the vector space as the phylogenetic invariants in Theorem 36.

**Proof.** Here it is easier to work with the coordinates \( \tilde{x}_\chi \) (obtained form \( x_\chi \) by a straightforward change of coordinates). We let \( F^n \) be the space of vectors where all the linear polynomials in the statement vanish. Then we shall prove that for the vectors in \( F^n \), any coordinate \( \tilde{x}_\chi \) can be expressed as a linear combination of the following 15 coordinates:

\[
\tilde{x}_{xxx}, \tilde{x}_{xyy}, \tilde{x}_{xyz}, \tilde{x}_{xzy}, \tilde{x}_{xzx}, \tilde{x}_{yxx}, \tilde{x}_{yxy}, \tilde{x}_{yxz}, \tilde{x}_{yzy}, \tilde{x}_{xyz}, \tilde{x}_{xy}, \tilde{x}_{yz}, \tilde{x}_{yzw}, \tilde{x}_{xyw}
\]

This will prove that \( F^n \) is a vector space of dimension 15 or lower. By Lemma 11 we know that \( \dim D^\pi \geq \mid x_\chi \mid - 1 \), which is \( B_4 - 1 = 14 \) for \( n = 4 \). As we have the inclusion \( D^\pi = E^n \cap H \subseteq F^n \cap H \) this will finish the proof.

First note that by Lemma 11 we have \( \tilde{x}_{xyy'y'} = \tilde{x}_{xyxy}, \tilde{x}_{xyxy'y'} = \tilde{x}_{xyxy}, \tilde{x}_{x'y'z''w'} = \tilde{x}_{xyzw} \) for any \( y' \neq x, x', z' \neq y, y', x, x', w' \neq x, y, z, x', y', z'' \).

Using the equation (7)=0 one can put \( \tilde{x}_{x'y'z''w'} \) as a linear combination of \( \tilde{x}_{xyxyz} \) and \( \tilde{x}_{xyxyw} \).

The equation (7)=0 allows us to put \( \tilde{x}_{xyy'y'} \) as a linear combination of \( \tilde{x}_{xyxy} \) if \( y' \neq y \). In order to write \( \tilde{x}_{yxx} \) (or similarly \( \tilde{x}_{yxy} \)) in terms of the allowed coordinates we need to do two steps. We use Eqn. (7) three times to put first \( \tilde{x}_{yxx} \) in terms of \( \tilde{x}_{yxyz} \) first, then \( \tilde{x}_{yxyz} \) in terms of \( \tilde{x}_{yxxx} \) and finally \( \tilde{x}_{yxxx} \) in terms of \( \tilde{x}_{xyyy} \). Changing the role of leaves 1,2 by 3,4 we also obtain \( \tilde{x}_{x'y'z''y} \) as a linear combination of \( \tilde{x}_{xyxy} \) if \( x' \neq x \). In the same way, we can use the equation (7)=0 to put \( \tilde{x}_{x'y'z''y'} \) as a linear combination of \( \tilde{x}_{xyxy} \) and other coordinates which we now know that
are linear combinations of the allowed coordinates. Finally, we use the equation \( (10) = 0 \) to put 
\[ \tilde{x}_{x'} = (1 - a')x \] 
for \( x' \neq x \) as a linear combination of \( \tilde{x}_{x_{xxx}} \) and other allowed coordinates.

By considering these relations above and all permutations of the leaves, we end up with every coordinate written as a linear combination of the allowed list of 15 coordinates. □

We now consider the two linear topology invariants that we obtained in Example 5 in terms of the \( \tilde{x}' \)s above, the corresponding equations for the quartet tree 12|34 these are

\[ H_1 : \tilde{x}_{xyxy} + \tilde{x}_{xyyx} = \tilde{x}_{xyyw} + \tilde{x}_{xyxw} \]
\[ H_2 : \tilde{x}_{zyyx} + \tilde{x}_{zywy} = \tilde{x}_{zyyw} + \tilde{x}_{zyxw} \]

Equations \( H_1 \) and \( H_2 \) are linearly independent and drop the dimension by two. In total, we have that \( D^\pi \) is contained in an affine space \( E^\pi \cap H \cap H_1 \cap H_2 \) of dimension 12. As the dimension of \( D^\pi \) is 12 and for the star tree \( \dim D^\pi_{|T_*} = 11 \) we have:

**Corollary 37.** For \( n = 4 \) and any distribution \( \pi \) one has

\[ D^\pi = E^\pi \cap H \]
\[ D^\pi_{|T_*} = E^\pi \cap H \cap H_1 \cap H_2 \cap H_3 \]

where \( H_3 : \tilde{x}_{xyxy} + \tilde{x}_{zywy} = \tilde{x}_{zyyw} + \tilde{x}_{zyxw} \) and \( T_* \) denotes the star tree on four leaves.

8. **The infinite-state random cluster model \( RC_\infty \)**

Recall that in the random cluster model, each edge of \( T \) is cut with some probability \( \theta_e \) to obtain a resulting partition \( \sigma \) of the leaf set \( X \). Each block is then assigned a state independently according to the distribution \( \pi \). However, we could just consider the partition \( \sigma \) itself as the output of this process (rather than assigning states, which has the effect of combining some blocks together when they receive the same state). We call this the **infinite state RC model** \( RC_\infty \) since it has a natural interpretation as the limiting distribution on partitions induced by the \( EI/RC \) model as the number of states \( k \) in \( S \) tends to infinity when states have at least roughly similar probabilities.

More precisely, under the \( RC \) model, the probability that two blocks of \( \sigma \) are assigned a same state in the equal state model is at most \( n |S| \to \infty \), by Boole’s inequality (note that there are at most \( n \) blocks in \( \sigma \)). Suppose that \( \pi_{a,b} \in [a/k, b/k] \) for some fixed \( a, b \) then as \( k = |S| \to \infty \) all blocks of \( \sigma \) receive distinct states with probability converging to 1 (this restriction on \( \pi \) can be weakened a little further). The \( RC_\infty \) model is sometimes referred to as the ‘Kimura’s infinite alleles’ model in phylogenetics, and it was studied mathematically in [13].

8.1. **Linear invariants for \( RC_\infty \).** The linear phylogenetic invariants for the infinite-state random cluster model are particularly easy to describe.

Let \( \text{Inc}(T) \) be the set of partitions of \( |n| \) that are not convex on \( T \) (i.e. they are ‘incompatible’ with \( T \)). Thus \( |\text{Inc}(T)| = B_n - |\text{co}(T)| \) where the Bell number \( B_n \) is the total number of partitions of the set \( |n| \).

Let \( p_{\Theta} = P_T(\sigma|\Theta) \) be the probability of generating partition \( \sigma \) on \( T \) under the \( RC_\infty \) model with edge cut probabilities \( \Theta = (\theta_e) \).

**Proposition 38.** Under the \( RC_\infty \) model:

(i) \( p_T(\sigma|\Theta) = 0 \) for all \( \Theta \) if and only if \( \sigma \in \text{Inc}(T) \).

(ii) \( \{\tau_{\sigma} : \sigma \in \text{Inc}(T)\} \) forms a basis for the vector space \( L_T \) of linear phylogenetic invariants for \( T \). Consequently, this space has dimension \( |\text{Inc}(T)| = B_n - |\text{co}(T)| \).

(iii) The space of all phylogenetic mixtures on \( T \) has dimension \( |\text{co}(T)| - 1 \).
(iv) The space of all phylogenetic mixtures on all trees under the RC\(_\infty\) model has dimension \(B_n - 1\).

Proof. (i) Suppose that \(\sigma \in \text{Inc}(T)\). Then there exists two blocks \(B, B'\) of \(\sigma\) and leaves \(x, y \in B\) and \(x', y' \in B'\) for which the paths \(P(T; x, y)\) and \(P(T; x', y')\) share at least one vertex. Now since \(x, y \in B\) and \(x', y' \in B'\) the only way to generate \(\sigma\) under \(\text{RC}_\infty\) is if none of the edges in the two paths \(P(T; x, y)\) and \(P(T; x', y')\) is cut. Since these paths intersect on a vertex this implies that \(x\) and \(x'\) must be the same block, i.e. that \(B = B'\). Thus \(\sigma\) cannot be generated with positive probability under the \(\text{RC}_\infty\) model. Conversely, suppose that \(\sigma\) is convex on \(T\). Then set \(\theta_x = 0\) for all edges in \(\{T[B] : B \in \sigma\}\) and set \(\theta_{e} = 1\) for all other edges. Then \(p_{\sigma} = 1\).

(ii) If \(\sum_{\lambda} x_{\sigma}\) is a linear phylogenetic invariant, then for any \(\sigma\) convex on \(T\) we can choose a set of parameters \(\Theta\) such that \(p_{\sigma} = 1\) (see above). This implies that \(\lambda_{\sigma} = 0\) for any \(\sigma \in \text{co}(T)\). This and (i) show that the span is the space of all linear phylogenetic invariants, and linear independence follows immediately from the observation that each polynomial involves a variable not present in any other polynomial in this set. Note that all these polynomials are topology invariants.

(iii) The space of phylogenetic mixtures \(D_T\) on \(T\) is equal to \(E_T \cap H\) where \(E_T\) is the space of vectors on which the linear phylogenetic invariants vanish and \(H\) is the hyperplane defined by the trivial equation \(\sum_{x} x_{\sigma} = 1\) (the sum is over all partitions of \([n]\)). By (ii), \(E_T\) has dimension \(B_n - \text{Inc}(T) = |\text{co}(T)|\) and we are done.

(iv) Note that in the basis \(\{x_{\sigma} : \sigma \in \text{Inc}(T)\}\) of (ii) there are no model invariants. Therefore, the set \(D\) of phylogenetic mixtures on all trees coincides with the trivial hyperplane \(H\) and has dimension \(B_n - 1\). \(\square\)

The construction of (quadratic) phylogenetic invariants for \(\text{RC}_\infty\) is also quite easy. Let \(x \sim y\) denote the event that \(x\) and \(y\) are in the same block of the partition generated by a phylogeny under the \(\text{RC}_\infty\) model, and let \(p(x, y)\) denote the probability of that event. Note that \(p(x, y)\) is a sum of \(p_{\sigma}\) values over all \(\sigma\) for which \(x\) and \(y\) are in different blocks. Then \(p(x, y) = \prod_{e \in P(T; x, y)} (1 - \theta_{e})\), where \(P(T; x, y)\) is the path in \(T\) between \(x\) and \(y\). It follows (from the four point condition) that if the quartet tree obtained by restricting \(T\) to \(x, y, w, z\) is either \(xy/wz\) or the star tree, then

\[p(x, w)p(y, z) - p(x, z)p(y, w) = 0.\]

9. Future work

It would be interesting to generalize Lake-type invariants in such a way that they generate the space of linear topology invariants for \(\kappa < n\) (cf. Corollary 27). On the other hand, it also would be useful to give explicit linear model invariants for any number of leaves, as was done in Section 4 for \(n = 3, 4\). These model invariants could be used for model selection as it was done in [10] for the uniform distribution. Extending the work of Section 4 to other models is also of interest because this would increase the range of models that can be considered in certain model selection software such as SPIn (http://genome.crg.es/cgi-bin/phylo_mod_sel/AlgModelSelection.pl).

Acknowledgments

Part of this research was performed while MC was visiting the Biomathematics Research Center of the University of Canterbury. MC would like to thank the Biomathematics Research Center (and specially its director) for the invitation, the support provided, and the great working atmosphere. MC is partially supported by MTM2015-69135-P and Generalitat de Catalunya 2014 SGR-634.
References

[1] E. S. Allman, J. A. Rhodes, and S. Sullivant, *When do phylogenetic mixture models mimic other phylogenetic models?*, Syst. Biol., 61 (2012), pp. 1049–1059.

[2] M. Casanellas and J. Fernández-Sánchez, *Relevant phylogenetic invariants of evolutionary models*, J. Math. Pures Appl., 96 (2011), pp. 207–229.

[3] M. Casanellas, J. Fernández-Sánchez, and A. M. Kedzierska, *The space of phylogenetic mixtures for equivariant models*, Alg. Mol. Biol., 7 (2012), 7:33.

[4] J. T. Chang, *Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency*, Math. Biosci., 137 (1996), pp. 51–73.

[5] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Sunderland, MA, 2004.

[6] J. Fernández-Sánchez and M. Casanellas, *Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages*, Syst. Biol., (in press), 2015. doi: 10.1093/sysbio/syv086.

[7] Y. X. Fu, *Linear Invariants under Jukes’ and Cantor’s one-parameter model*, J. Theor. Biol. (1995) 173 (1995), pp. 339–352.

[8] Y. X. Fu and W. H. Li, *Necessary and sufficient conditions for the existence of certain quadratic invariants under a phylogenetic tree*, Math. Biosci. 105 (1991), pp. 229–238.

[9] M. D. Hendy and D. Penny, *Complete families of linear invariants for some stochastic models of sequence evolution, with and without the molecular clock assumption*, J. Comput. Biol., 3 (1996), pp. 19–31.

[10] A. Kedzierska, M. Deton, R. Guigó, and M. Casanellas, *SPIn: model selection for phylogenetic mixtures via linear invariants*, Molecular Biology and Evolution 29 (2012), pp. 929–937.

[11] J. G. Kemeny and J. L. Snell, *Finite Markov chains*, Springer-Verlag, New York, 1976.

[12] J. Lake, *A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony*, Molec. Biol. Evol., 4 (1987), pp. 167–191.

[13] F. A. Matsen, E. Mossel, and M. Steel, *Mixed-up trees: The structure of phylogenetic mixtures*, Bull. Math. Biol., 70 (2008), pp. 1115–1139.

[14] E. Mossel and M. Steel, *A phase transition for a random cluster model on phylogenetic trees*, Math. Biosci. 187 (2004), pp. 189–203.

[15] C. Semple and M. Steel, *Phylogenetics*, Oxford University Press, 2003.

[16] M. A. Steel, L. A. Székely, and M. D. Hendy, *Reconstructing trees when sequence sites evolve at variable rates*, J. Comput. Biol., 1 (1994), pp. 153–163.

[17] M.A. Steel and Y. X. Fu, *Classifying and counting linear phylogenetic invariants for the Jukes-Cantor model*, J. Comput. Biol., 2 (1995), pp. 39–47.

[18] M. Steel, *Can we avoid ‘SIN’ in the House of ‘No Common Mechanism’?*, Syst. Biol., 60 (2011), pp. 96–109.

[19] D. Štefaković and E. Vigoda, *Phylogeny of mixture models: robustness of maximum likelihood and non-identifiable distributions*, J. Comput. Biol., 14 (2007), pp. 156–189.

MC: Department of Mathematics – BGSMath, Universitat Politècnica de Catalunya, Barcelona, Spain

MS: Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand