Dimensions of Group-Based Phylogenetic Mixtures

Hector Baños, et al. [full author details at the end of the article]

Received: 19 November 2017 / Accepted: 24 August 2018 / Published online: 17 September 2018
© Society for Mathematical Biology 2018

Abstract
Mixtures of group-based Markov models of evolution correspond to joins of toric varieties. In this paper, we establish a large number of cases for which these phylogenetic join varieties realize their expected dimension, meaning that they are nondefective. Nondefectiveness is not only interesting from a geometric point-of-view, but has been used to establish combinatorial identifiability for several classes of phylogenetic mixture models. Our focus is on group-based models where the equivalence classes of identified parameters are orbits of a subgroup of the automorphism group of the abelian group defining the model. In particular, we show that for these group-based models, the variety corresponding to the mixture of $r$ trees with $n$ leaves is nondefective when $n \geq 2r + 5$. We also give improved bounds for claw trees and give computational evidence that 2-tree and 3-tree mixtures are nondefective for small $n$.

1 Introduction

Tree-based Markov models of DNA sequence evolution are commonly used for model-based phylogenetic inference. These models are collections of probability distributions, and their geometric properties have been well studied within the field of phylogenetic algebraic geometry. A special class of tree-based Markov models is the group-based models, in which the transition matrices of the models are assumed to exhibit certain symmetries. An important observation of Evans and Speed (1993) and Hendy et al. (1994) is that the varieties associated with group-based phylogenetic tree models are not only algebraic varieties, but toric varieties. This allows us to apply tools of computational and combinatorial algebraic geometry to their investigation. The class of group-based models includes many commonly used models of sequence evolution, including the Jukes–Cantor (JC) and Cavender–Farris–Neyman (CFN) models (Daskalakis et al. 2011; Hendy et al. 1994; Jukes and Cantor 1969; Neyman 1971).
When flexibility beyond a single tree model is needed, mixtures of tree-based Markov models are used, since such models can account for variation in the mutation process across sites (Pagel and Meade 2004). The variety associated with a phylogenetic mixture model is the join (Casanellas 2012) of the varieties of the individual phylogenetic tree models in the mixture. In this paper, we study the varieties associated with mixtures of group-based phylogenetic models in the Fourier coordinates, which are the joins of toric varieties. The dimensions of the join varieties associated with 2- and 3-tree group-based phylogenetic mixture models have previously been studied for the standard group-based models in order to establish the identifiability of the tree parameters of these models (Allman et al. 2011; Long and Sullivant 2015). In this paper, we extend these dimension results to phylogenetic mixture models of an arbitrary number of trees as well as to a broader class of group-based phylogenetic models.

We consider the model varieties and their joins as living in projective space. The projective dimension of the join of \( r \) projective varieties is at most the sum of their dimensions plus \( r - 1 \). This upper bound is typically realized, as long as it does not exceed the dimension of the ambient space, and so is referred to as the expected dimension. If a join variety has the expected dimension, it is said to be nondefective and is otherwise defective. Proving the nondefectiveness of the join varieties associated with mixture models is a key tool in establishing identifiability results for phylogenetic mixtures. Specifically, in Allman et al. (2011), Long and Sullivant (2015), the strategy for proving identifiability relied on showing the join varieties associated with 2- and 3-tree mixtures for the CFN, JC, and Kimura 2-parameter (K2P) models have the expected dimension for trees with few leaves. In this paper, we extend these results in several directions. We prove the following theorem, which not only shows that nondefectiveness holds more generally for joins of group-based models, but also holds when there are more trees in the mixture, provided the trees have a sufficient number of leaves.

**Theorem 1.1** Let \( T_1, \ldots, T_r \) be phylogenetic \([n]\)-trees with \( n \geq 2r + 5 \), \( G \) be an abelian group, and \( B \subset \text{Aut}(G) \). Then, \( V_{T_1}^{(G, B)} \ast \cdots \ast V_{T_r}^{(G, B)} \) has the expected dimension.

In addition to its applications to phylogenetics, Theorem 1.1 is interesting geometrically as it adds to the growing body of knowledge on the structure of joins and secant varieties in algebraic geometry. Defectiveness of joins of toric varieties has been intensively studied in some very specific cases, such as secants of Veronese varieties (Alexander and Hirschowitz 1995) and of Segre–Veronese varieties (Abo and Brambilla 2012, 2013), but little is known about the general case or even cases outside of these examples.

Our primary tool for proving Theorem 1.1 is a tropical version of Terracini’s Lemma, a classical tool for computing the dimensions of joins. The tropical version, introduced by Draisma, gives lower bounds for dimensions of joins of toric varieties by rephrasing questions about the dimensions of toric varieties as questions about the convex geometry of lattice points (Draisma 2008).

This paper is organized as follows. In Sect. 2, we describe the class of toric varieties associated with group-based phylogenetic models. We also describe toric ideals
more generally and explain how Draisma’s Lemma can be used to establish nondefectiveness. In Sect. 3, we exploit the combinatorics of trees and use Draisma’s Lemma (Theorem 2.2) to prove a version of the main theorem (Theorem 3.1) that holds for a class of models called the general group-based models and for binary trees. In Sect. 4, we prove Theorem 1.1, which does not require that the trees in the mixture be binary and allows for group-based models with parameter identifications. Finally, in Sect. 5, we give some improved bounds for special cases, list computational results for trees with few leaves, and state some conjectures based on these computations.

2 Preliminaries

In this section, we describe the toric varieties associated with group-based phylogenetic models and the join varieties associated with their mixtures. Since the focus of this paper is the algebraic and combinatorial properties of the varieties, we will not discuss the details of the underlying group-based phylogenetic models nor the discrete Fourier transform that enables us to view the models as toric varieties. Instead, we will describe how to construct the Fourier parameterization of the toric variety associated with a group-based model given only the data of a phylogenetic tree \( T \), a finite abelian group \( G \), and a subgroup \( B \) of the automorphism group of \( G \). For a thorough discussion of group-based phylogenetic models and their connections to toric varieties, we refer the reader to Sturmfels and Sullivant (2005), Sullivant (2018). We conclude the section by laying the groundwork for the application of Draisma’s Lemma, which is the primary tool that we will use to establish our main results in the subsequent sections.

2.1 Toric Varieties of Group-Based Models

Let \( T \) be an \( n \)-leaf unrooted phylogenetic tree with leaves labeled by the set \( \{1, \ldots, n\} \). We will call such a tree an \( [n] \)-tree and let \( V(T) \), \( E(T) \), \( L(T) \) denote the vertex, edge, and leaf vertex sets of \( T \), respectively. Let \( m \) denote the number of edges of \( T \) and let \( G = \{g_1, \ldots, g_k\} \) be an abelian group. A consistent leaf \( G \)-labeling of \( T \) is a function \( \xi : L(T) \rightarrow G \) such that

\[
\sum_{v \in L(T)} \xi(v) = 0.
\]

The algebraic variety of the general \( G \)-based model on \( T \) in the Fourier coordinates, denoted \( V_G^T \), is the Zariski closure of the image of a parameterization map

\[
h_T : \mathbb{C}^{mk} \rightarrow \mathbb{C}^{k^n-1}.
\]

The coordinates \( q_\xi \) of the image space of \( h_T \) are indexed by all of the consistent leaf \( G \)-labelings of \( T \). The \( mk \) Fourier parameters are denoted \( a_e^g \) for all \( e \in E(T) \) and \( g \in G \). In order to describe the parameterization, we first choose an orientation for each edge \( e \in E(T) \) and define \( L(e) \subseteq L(T) \) to be the set of leaves on the arrow side
of $e$. With this orientation, each consistent leaf-labeling $\xi$ now induces a consistent edge $G$-labeling of $T$, $\xi' : \mathcal{E}(T) \to G$, defined by

$$e \mapsto \sum_{v \in L(e)} \xi(v).$$

The parameterization is then given by

$$q_\xi = \prod_{e \in \mathcal{E}(T)} a_{\xi'(e)}. $$

Notice that any leaf-labeling of $T$ can be written as a vector $\xi = (\xi_1, \ldots, \xi_n)$ where $\xi_i$ is the label of the leaf of $T$ labeled by $i$. Thus, the set of consistent leaf $G$-labelings is independent of the tree being considered and can be identified with the set of vectors in $G^n$ that satisfy $\sum_{v \in L(T)} \xi_i = 0$. Consequently, for fixed $G$, we may view the varieties of $G$-based models on $n$-leaf trees as all contained in the same ambient space.

**Remark 1** This description of $V^G_T$ does not depend on the chosen orientation of $T$. To see this, note that for any fixed consistent leaf-labeling $\xi$, reversing the orientation of an edge $e$ simply exchanges $a^e_g$ and $a^{-e}_g$ everywhere in the parameterization, and the variety remains unchanged. Moreover, if $g = -g$ for every element in the group, such as when the group is $\mathbb{Z}/2\mathbb{Z}$ or $(\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/2\mathbb{Z})$, then the map $h_T$ is itself invariant under changes of orientation of $T$, and so orientation can be ignored entirely.

If we do not place any restrictions on the parameters $a^e_g$, then the variety $V^G_T$ is that of the general $G$-based model on $T$. For example, the Kimura 3-parameter model is the general $G$-based model for the group $G = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. Often, in phylogenetic applications, we also consider group-based phylogenetic models in which certain entries of the transition matrices are chosen to be equal. After the Fourier transformation, this imposes restrictions on the Fourier parameters. As an example, the Jukes–Cantor model (JC) of nucleotide substitution is also a group-based model for the group $G = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. However, the restrictions on the entries of the transition matrices impose restrictions on the Fourier parameters, so that for each edge $e \in \mathcal{E}(T)$, $a^e_{(1,0)} = a^e_{(0,1)} = a^e_{(1,1)}$.

Arbitrarily identifying Fourier parameters, we can construct a toric variety, but such a variety will not necessarily correspond to a group-based phylogenetic model (see for example, Michałek 2011, Appendix A). However, we can ensure the resulting toric variety will be the transform of a group-based phylogenetic model by specifying the identifications according to an equivalence relation on $G$. We require the equivalence classes of this relation to be the orbits of a subgroup of $\text{Aut}(G)$ and insist that for all $e \in \mathcal{E}(T)$, $a^e_g = a^e_h$ if and only if $g$ and $h$ are in the same equivalence class. In particular, this means that the identity element is always in its own class. Therefore, the variety in Fourier coordinates for a group-based phylogenetic tree model is specified by an $n$-leaf directed tree $T$ and a pair $M = (G, B)$ where $G$ is a finite abelian group.
and $B$ is a subgroup of $\text{Aut}(G)$. As such, we will now use the notation $V^M_T = V^{(G,B)}_T$ for the variety of the $G$-based model on $T$ with parameters identified according to $B$, or simply $V^G_T$ when $B = \{1\}$. The number of parameters for these models will be $m(l + 1)$, where $(l + 1)$ is the number of orbits of $B$.

The pairs $(G, B)$ corresponding to the standard phylogenetic group-based models are given as follows:

- $\text{CFN} = (\mathbb{Z}/2\mathbb{Z}, \{1\})$,
- $\text{JC} = (\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}, S_3)$,
- $\text{K2P} = (\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}, S_2)$,
- $\text{K3P} = (\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}, \{1\})$,

where $\text{Aut}(\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z})$ is identified with the permutation group $S_3$. Observe also that the general group-based models are then precisely those models for which $B = \{1\}$. Our strategy for proving the main result, Theorem 1.1, will first be to prove some results for general group-based models in Sect. 3 and then to show in Sect. 4 that they still hold for models in which we identify certain parameters.

**Example 2.1** This example demonstrates the parameterization of one of the Fourier coordinates for the K3P model on the 4-leaf unrooted tree $T$ pictured below. For this model, $G = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ and $B = \{1\}$. As noted in Remark 1, because $G = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, the parameterization will remain unchanged if we reorient edges in $T$:

```
1  e1  e4  4
  e2
  e3  2
  3
```

The leaf-labeling $\xi = ((0, 1), (1, 0), (1, 1), (0, 0))$ is consistent since $(0, 1) + (1, 0) + (1, 1) + (0, 0) = (0, 0)$. Using $a^i_g$ for the Fourier parameter associated with the group element $g$ on the edge $e_i$,

$$q_{((0,1),(1,0),(0,0),(1,1))} = a^1_{(0,1)} a^2_{(1,0)} a^3_{(1,1)} a^4_{(0,0)} a^5_{(1,0)+ (1,1)} = a^1_{(0,1)} a^2_{(1,0)} a^3_{(1,1)} a^4_{(0,0)} a^5_{(1,0)}.$$

If instead, we consider the Jukes–Cantor model, where all of the non-identity parameters are assumed to be equal, this simply becomes

$$q_{((0,1),(1,0),(0,0),(1,1))} = a^1_{(1,0)} a^2_{(1,0)} a^3_{(1,0)} a^4_{(0,0)} a^5_{(1,0)}.$$

### 2.2 Joins of Toric Varieties and Mixture Models

Let $M = (G, B)$. Because $V^M_T$ is the image of a monomial map $h_T : \mathbb{C}^{m(l+1)} \rightarrow \mathbb{C}^{kn-1}$, it is a complex toric variety. Since the map is homogeneous, we can consider $V^M_T$ as a projective variety in $\mathbb{P}^{kn-1}$.$^{-1}$. The map $h_T$ can be described by a $m(l + 1) \times kn-1$ matrix $A$, the columns of which are the exponent vectors of the monomials.
parameterizing each Fourier coordinate. Let $A \subseteq \mathbb{R}^{m(l+1)}$ denote the set of column vectors of $A$ and let $P \subseteq \mathbb{R}^{m(l+1)}$ be the convex hull of $A$. The geometry of $P$ and $V^M_T$ are closely tied in many ways, and in fact $P$ has been studied in depth for several group-based models (Buczynska and Wisniewski 2007; Mauhar et al. 2017; Michałek 2011). We will be primarily interested in the relationship between the dimension of $V^M_T$ and the dimension of $P$. That is

$$\dim V^M_T = \dim \mathbb{R} P = \text{rank } A - 1,$$

where $\dim V^M_T$ denotes the projective dimension (Sturmfels 1996).

Fix $B$ and label the equivalence classes of $G$ induced by $B$ by the integers from $0, \ldots, l$, with $0$ labeling the equivalence class containing only the identity. Let $\delta : G \to \mathbb{R}^{l+1}$ be the map sending each group element to $e_i$ where $i$ is the label of its equivalence class and $e_0, \ldots, e_l$ are the $l+1$ standard basis vectors. The image of $\delta$ is the set of vertices of a standard $l$-simplex $\Delta_l := \text{conv}(e_0, \ldots, e_l)$, and its affine span is the hyperplane $K$ defined by $x_0 + \cdots + x_l = 1$ where $x_i$ denotes the $i$th coordinate.

Each of the column vectors of $A$ comes from a consistent edge-labeling of $T$, $(g_1, \ldots, g_m) \in G^m$, by the map

$$\delta^m : G^m \to (\mathbb{R}^{l+1})^m$$

$$(g_1, \ldots, g_m) \mapsto (\delta(g_1), \ldots, \delta(g_m)).$$

The image of $\delta^m$ is the set of lattice points corresponding to all possible $G$ edge-labelings of the graph (not just consistent ones), so $A \subseteq \text{Im } \delta^m$. The convex hull of $\text{Im } \delta^m$ is the polytope $\Delta^m_l$, e.g., for the CFN model, $\text{conv}(\text{Im } \delta^m) = \Delta^m_l$ is an $m$-dimensional cube. The polytope $\Delta^m_l$ has dimension $lm$, and its affine span is $K^m$. The polytope $P^M_T$ associated with $V^M_T$ is contained in $\Delta^m_l$, and thus,

$$\dim V^M_T \leq lm.$$

**Remark 2** The map $h_T$ is multi-homogeneous, which is reflected in the fact that $P^M_T$ is contained in the codimension-$m$ space $K^m \subseteq \mathbb{R}^{(l+1)m}$. It is sometimes convenient to consider the dehomogenized map by projecting away coordinate $x_0$ for each edge. The projected polytope $P^M_T \subseteq \mathbb{R}^{lm}$ differs only by a linear change of coordinates from the above, but now affinely spans the ambient space in the case that it realizes the upper bound and has dimension $lm$.

Given varieties $W_1, \ldots, W_r \subseteq \mathbb{P}^{N-1}$, we denote their join by $W_1 * \cdots * W_r$, which is defined as the Zariski closure of the set of linear spaces defined by one point from each variety. The variety $W_1 * \cdots * W_r$ can be considered as the closure of the image of the map

$$W_1 \times \cdots \times W_r \times \mathbb{P}^{r-1} \to \mathbb{P}^N-1,$$

$$(p_1, \ldots, p_r, [c_1 : \ldots : c_r]) \mapsto c_1 p_1 + \cdots + c_r p_r.$$
Just by counting parameters, the dimension of $W_1 \ast \ldots \ast W_r$ satisfies
\[ \dim(W_1 \ast \ldots \ast W_r) \leq \dim W_1 + \cdots + \dim W_r + (r - 1). \]

Another upper bound on $\dim(W_1 \ast \ldots \ast W_r)$ is $N - 1$, the dimension of the ambient space. The expected dimension of $W_1 \ast \ldots \ast W_r$ is
\[ \min\{\dim W_1 + \cdots + \dim W_r + (r - 1), N - 1\}. \]
If the dimension is less than the expected dimension, $W_1 \ast \ldots \ast W_r$ is said to be defective.

For $W \subseteq \mathbb{P}^{N-1}$, the join $W \ast W$ is called the secant variety of $W$ (or more specifically the second secant variety of $W$), also written $\sigma(W)$ or $\sigma_2(W)$. For any integer $r \geq 1$, the $r$th secant variety of $W$ is
\[ \sigma_r(W) := W \ast \cdots \ast W. \]
The expected dimension of $\sigma_r(W)$ is
\[ \min\{r(\dim W + 1) - 1, N - 1\}. \]

To prove the main theorem, we will rely heavily on Draisma’s Lemma, which allows us to determine lower bounds on the dimensions of joins and secants of toric varieties using tropical geometry. Let $W_1, \ldots, W_r$ be projective toric varieties in $\mathbb{P}^{N-1}$, and let $A_i$ be the $m_i \times N$ matrix associated with $W_i$ for $i = 1, \ldots, r$. Let $v = (v_1, \ldots, v_r)$ be a sequence of linear functionals, with each $v_i : \mathbb{R}^{m_i} \to \mathbb{R}$, considered as a row vector. Then, $v_i A_i$ is a row vector in $(\mathbb{R}^N)^*$. Let $W_i(v) \subseteq [N]$ denote the set of positions $j$ such that the $j$th entry of $v_i A_i$ is strictly less than the $j$th entry of $v_l A_l$ for all $l \neq i$. Thus, $W_1(v), \ldots, W_r(v)$ are disjoint, and for a generic choice of $v$ form a partition of $[N]$.

Now, let $D_i(v) \subseteq \mathbb{R}^{m_i}$ be the set of column vectors of $A_i$ in the positions given by $W_i(v)$. Let $\text{rank} D_i(v)$ denote the dimension of the span of $D_i(v)$. This is one more than $\dim \text{conv}(D_i(v))$, the dimension of the affine span of $D_i(v)$.

**Theorem 2.2** (Draisma’s Lemma, Corollary 2.3 of Draisma 2008) For any choice of $v \in \prod_{i=1}^r (\mathbb{R}^{m_i})^*$,
\[ \dim(W_1 \ast \ldots \ast W_r) \geq \text{rank} D_1(v) + \cdots + \text{rank} D_r(v) - 1. \]
It follows that if there exists $v$ such that $\text{rank} D_i(v) = m_i$ for all $i = 1, \ldots, r$, then $W_1 \ast \ldots \ast W_r$ has the expected dimension.

In the case of a secant variety, the picture is a bit simpler. Let the toric variety $W \subseteq \mathbb{P}^{N-1}$ have $m \times N$ matrix $A$ and let $\mathcal{A} \subseteq \mathbb{R}^m$ denote the set of column vectors of $A$. Let $v = (v_1, \ldots, v_r)$ be a sequence of linear functionals on $\mathbb{R}^m$. The sequence $v$ divides $\mathbb{R}^m$ into open convex regions $R_1(v), \ldots, R_r(v)$ defined by
\[ R_i(v) := \{ p \in \mathbb{R}^m \mid v_i(p) < v_l(p) \text{ for all } l \neq i \} \]

for \( i = 1, \ldots, r \).

**Corollary 2.3** For any choice of \( v \in \prod_{i=1}^r (\mathbb{R}^m)^* \),

\[ \dim \sigma_r(W) \geq \text{rank}(\mathcal{A} \cap R_1(v)) + \cdots + \text{rank}(\mathcal{A} \cap R_r(v)) - 1. \]

In the particular case of \( r = 2 \), the regions \( R_1(v) \) and \( R_2(v) \) can be described as the open half-spaces on either side of a hyperplane \( \mathcal{H} \). In this case, we denote the regions by \( \mathcal{H}^+ \) and \( \mathcal{H}^- \). For larger \( r \), it may also be useful to partition the space by \( r - 1 \) hyperplanes (although these are not the only sort of partitions allowed). Eventually, we will apply the following proposition to the polytopes \( P^M_T \) with vertices corresponding to consistent edge-labelings of \( T \).

**Proposition 2.4** Let \( \mathcal{H}_1, \ldots, \mathcal{H}_{r-1} \) be hyperplanes through a polytope \( P \), with no two intersecting in \( P \). Let \( P_1, \ldots, P_r \) be the connected components of \( P \setminus (\mathcal{H}_1 \cup \cdots \cup \mathcal{H}_{r-1}) \). Then, there is a sequence of functionals \( v = (v_1, \ldots, v_r) \) such that \( P_i = R_i(v) \cap P \).

**Proof** Because the hyperplanes do not intersect in \( P \), for each \( i \neq j \), \( \mathcal{H}_j \cap P \) is contained in \( \mathcal{H}_i^+ \) or \( \mathcal{H}_i^- \). Therefore by reindexing, and choosing plus and minus labels appropriately, we have

\[ (\mathcal{H}_1^+ \cap P) \supseteq (\mathcal{H}_2^+ \cap P) \supseteq \cdots \supseteq (\mathcal{H}_{r-1}^+ \cap P). \]

The connected components of \( P \setminus (\mathcal{H}_1 \cup \cdots \cup \mathcal{H}_{r-1}) \) are then \( P_1 = \mathcal{H}_1^- \cap P \), \( P_i = \mathcal{H}_{i-1}^+ \cap \mathcal{H}_i^- \cap P \) for \( 1 < i < r \), and \( P_r = \mathcal{H}_{r-1}^+ \cap P \). For each \( \mathcal{H}_i \), choose a functional \( \ell_i \) which vanishes on \( \mathcal{H}_i \) and is positive on \( \mathcal{H}_i^+ \). Then, let \( v_1 \) be the zero vector, and let \( v_i = -\ell_1 - \cdots - \ell_{i-1} \) for \( i = 2, \ldots, r \). For any \( p \in P_i \), \( \ell_j(p) > 0 \) for \( j < i \) and \( \ell_j(p) < 0 \) for \( j \geq i \). Therefore, \( v_i(p) \) is the unique minimum among \( v_1(p), \ldots, v_r(p) \), so \( p \in R_i(v) \). \( \square \)

### 3 Joins of General Group-Based Models on Binary Trees

In this section, we prove two intermediary theorems on our way to proving Theorem 1.1. Together, these two theorems establish the main result for all groups so long as the trees \( T_1, \ldots, T_r \) are binary and \( B \) is trivial. We require two theorems, since slightly different arguments are needed when \( G = (\mathbb{Z}/2\mathbb{Z}) \). This group-based model is particularly relevant to phylogenetic applications as it is exactly the CFN model. We handle that case first, and then the case \( |G| > 2 \). In the next section, we will generalize this result for arbitrary trees \( T_1, \ldots, T_r \) and for \( B \) an arbitrary subgroup of \( \text{Aut}(G) \).

#### 3.1 Joins for the CFN Model

We will prove the following theorem regarding binary trees using Draisma’s Lemma.
Theorem 3.1 Let $T_1, \ldots, T_r$ be binary phylogenetic $[n]$-trees with $n \geq 2r + 5$. Then, $V_{T_1}^{\mathbb{Z}/2\mathbb{Z}} \cdots V_{T_r}^{\mathbb{Z}/2\mathbb{Z}}$ has the expected projective dimension, $r(2n - 3) + r - 1$.

Since a $n$-leaf binary tree has $2n - 3$ edges, each toric variety $V_{T_i}^{\mathbb{Z}/2\mathbb{Z}}$ is parametrized by a monomial map with set of exponent vectors $A_i \subseteq \mathbb{R}^{2(2n-3)}$, which correspond to the consistent labelings of $T_i$. The affine span of $A_i$ has dimension $2n - 3$, and thus the projective dimension of $V_{T_i}^{\mathbb{Z}/2\mathbb{Z}}$ is $2n - 3$.

Per Draisma’s Lemma, we demonstrate the existence of a set of functionals $v = (v_1, \ldots, v_r)$ that partition the consistent labelings into sets $W_1(v), \ldots, W_r(v)$. Then, for all $1 \leq i \leq r$, $D_i(v)$ is a proper subset of $A_i$ constructed from $W_i(v)$ according to Draisma’s Lemma. The goal is to show that the dimension of the affine span of $D_i(v)$ is the same as the dimension of the affine span of $A_i$, namely $2n - 3$.

The general strategy will be to partition the consistent labelings based on the number of leaves not labeled by the identity. The subset of $A_i$ of vectors corresponding to labelings with exactly $c$ non-identity leaf labels has affine span of dimension at most $2n - 4$ because this imposes one linear constraint on the vectors. In the following lemma, we show that for even $2 \leq c \leq n - 5$, the affine span has exactly that dimension. (Note that the rank of the linear span is one larger than the affine dimension.) An example of the construction from this lemma is illustrated in Example 3.3.

As an exponent vector, each entry of an vector in $A_i \subseteq \mathbb{R}^{2(2n-3)}$ corresponds to a Fourier parameter $a^\epsilon_g$ with superscript $\epsilon \in \mathcal{E}(T)$ and subscript $g \in G = \mathbb{Z}/2\mathbb{Z}$. In order to recall this correspondence, we use the same superscripts and subscripts for the coordinates of $\mathbb{R}^{2(2n-3)}$, i.e., we will refer to the coordinates of $\mathbb{R}^{2(2n-3)}$ as $x^\epsilon_g$ where $\epsilon \in \mathcal{E}(T)$ and $g \in G$. We will refer to the unit vector corresponding to the coordinate $x^\epsilon_g$ as $e^\epsilon_g$. We also slightly abuse notation and interpret $\mathcal{L}(T)$ as the set of leaf edges of $T$ or as the set of leaf vertices of $T$ depending on the context. Similarly, we do not carefully distinguish between a leaf vertex and the leaf edge leading to that vertex.

Lemma 3.2 Let $G = \mathbb{Z}/2\mathbb{Z}$ and let $T$ be a binary phylogenetic $[n]$-tree. Let $A$ be the set of vectors describing the monomial map parameterizing $V_{T}^{\mathbb{Z}/2\mathbb{Z}}$. Let $S_c \subseteq \mathbb{R}^{2(2n-3)}$ be the hyperplane defined by

$$\sum_{\ell \in \mathcal{L}(T)} x^\ell_1 = c.$$

For even $2 \leq c \leq n - 5$, $\text{rank}(A \cap S_c) = 2n - 3$.

Proof Let $\pi : \mathbb{R}^{2(2n-3)} \to \mathbb{R}^{2n-3}$ be the projection that forgets coordinates $x_0^\epsilon$ for identity element $0 \in \mathbb{Z}/2\mathbb{Z}$ and each edge $\epsilon \in \mathcal{E}(T)$. Then, $\pi(A \cap S_c)$ is a set of $0/1$ vectors with $c$ leaf coordinates equal to $1$.

Let $E$ be an internal edge of $T$. Redraw $T$ as below where each $r_j$ is a rooted subtree of $T$ with root $\rho_j$. 

© Springer
For \(1 \leq j \leq 4\), choose \(L_j\) to be a subset of the leaves of \(r_j\) so that each \(|L_j|\) is odd and so that \(|L_1| + |L_2| + |L_3| + |L_4| = c\). This is always possible since \(c \leq n - 5\). Now, label all the leaves in \(L_1 \cup L_2 \cup L_3 \cup L_4\) by 1 to give a consistent leaf-labeling of \(T\). This induces a consistent edge-labeling of \(T\). Observe that in each rooted subtree \(r_j\) there is a unique leaf \(\lambda_j\) such that the path from \(\rho_j\) to \(\lambda_j\) involves only edges labeled by 1.

Let \(F_{kl} \in \pi(A \cap S_c)\) be the vector corresponding to the subforest of \(T\) induced by labeling all the leaves in

\[
\left( \bigcup_{1 \leq j \leq 4} (L_j \setminus \{\lambda_j\}) \right) \cup \{\lambda_k, \lambda_l\}
\]

by 1 and all other leaves by 0. Let \(F_E\) be the vector that corresponds to the subforest induced by labeling all of the leaves in

\[
\bigcup_{1 \leq j \leq 4} (L_j \setminus \{\lambda_j\})
\]

by 1 and all other leaves by 0. The key observation is that

\[
F_{kl} = F_E + \sum_{\epsilon \in p(\lambda_k, \lambda_l)} e^\epsilon,
\]

where \(p(\lambda_k, \lambda_l)\) is the path between \(\lambda_k\) and \(\lambda_l\). Therefore, we have

\[
F_{13} + F_{24} - F_{12} - F_{34} = 2e^E_1.
\]

Now suppose that \(\ell\) is a leaf edge of \(T\). Let \(\Lambda\) be a \(c\)-element subset of the leaves and label each leaf in this subset by 1 and all of the rest by 0. Since \(c\) is even, this is a consistent labeling. The vector corresponding to the subforest induced by this labeling is in \(\pi(A \cap S_c)\). Moreover, since we have already shown that \(\pi(A \cap S_c)\) contains \(e^\epsilon_1\) for any internal edge \(\epsilon\), the vector \(F_\Lambda = \sum_{\epsilon \in \Lambda} e^\epsilon\) must also be in \(\pi(A \cap S_c)\). For each leaf \(\epsilon \neq \ell\), let \(\Lambda_\epsilon\) be any \(c\)-element subset of the leaves that contains \(\epsilon\) but not \(\ell\). Then,

\[
F_{(\Lambda_\epsilon \setminus \{\epsilon\}) \cup \{\ell\}} - F_{\Lambda_\epsilon} = e^\ell_1 - e^\epsilon_1.
\]
and, we get,
\[ F_\Lambda + \sum_{\epsilon \in \Lambda} (e_1^\epsilon - e_1^\epsilon) = ce_1^\epsilon. \]

Since we can repeat this procedure for every leaf edge \( \ell \) and we have shown that \( e_1^E \in \langle A \cap S_e \rangle \) for every internal edge \( E \), we can conclude that \( \operatorname{rank} \langle A \cap S_e \rangle \geq \operatorname{rank} \langle \pi(A \cap S_e) \rangle \geq 2n - 3. \]

Example 3.3 The figure below represents a consistent leaf-labeling for which \( c = 16 \). Both the blue and red vertices are elements of the \( L_j \) (\( |L_1| = 7 \) and \( |L_2| = |L_3| = |L_4| = 3 \)) and the blue vertices are the \( \lambda_i \). All of the colored edges correspond to nonzero entries in the vector \( F_{12} \), and the red-colored edges to the nonzero entries in \( F_E \).

To use Draisma’s Lemma to prove Theorem 3.1, we need to construct sets \( D_i(v) \) with dimension \( 2n - 3 \) for each \( 1 \leq i \leq r \). In the following proof, we will show how to construct \( v \) so that \( A_i \cap S_{2i} \subseteq D_i(v) \). By Lemma 3.2, since the affine span of each \( A_i \cap S_{2i} \) has dimension \( 2n - 4 \), this ensures that \( \dim(D_i(v)) \geq 2n - 4 \). However, to prove the theorem, we will also need to ensure each \( D_i(v) \) contains a vector outside of the hyperplane \( S_{2i} \). For \( i = 1 \), this vector will be the vector corresponding to the trivial labeling, which we will call \( p_0 \). For each \( i > 1 \), we will need to “borrow” a vector \( p_i \) from an adjacent slice.

Proof of Theorem 3.1 Note that for \( n \geq 7 \),
\[ |A_i \cap S_{2i}| = \binom{n}{2i} \geq \binom{n}{2} > 2n - 3. \]

Therefore, the set \( A_i \cap S_{2i} \) must have some linear dependencies. Choose \( p_i \in A_i \cap S_{2i} \) such that \( \langle (A_i \cap S_{2i}) \setminus \{p_i\} \rangle \) still has dimension \( 2n - 3 \). Let \( L_i \) be the set of leaf edges labeled 1 in the labeling corresponding to the vector \( p_i \).

For each \( 1 \leq i \leq r \), let \( \pi_i : \mathbb{R}^{2(2n - 3)} \to \mathbb{R}^{2n} \) be the projection that forgets the coordinates of the non-leaf edges. Let \( \mathcal{H}_1, \ldots, \mathcal{H}_{r - 1} \) be the hyperplanes in \( \mathbb{R}^{2n} \) with \( \mathcal{H}_i \) defined by

\[ \mathcal{H}_i = \{ x \in \mathbb{R}^{2n} : x_{2i} = x_{2i+1} = \cdots = x_{2n-1} = 0 \}. \]
\[
\sum_{j \in [n]} x_j^i + \frac{2}{4i - 1} \sum_{j \in L_i} x_j^i = 2i + 1.
\]

This hyperplane is constructed so that \(\pi_i(q) \in \mathcal{H}_i^-\) for all \(q \in (\mathcal{A}_i \cap \mathcal{S}_{2i}) \setminus \{p_i\}\) but \(\pi_i(p_i) \in \mathcal{H}_i^+\) and \(\pi_{i+1}(q) \in \mathcal{H}_{i+1}^+\) for all \(q \in \mathcal{A}_{i+1} \cap \mathcal{S}_{2i+2}\).

By Proposition 2.4, there is a sequence of functionals \(v' = (v'_1, \ldots, v'_r)\) such that \(P_i = R_i(v') \cap P\) (where \(P, P_i\) and \(R_i(v')\) are defined in Proposition 2.4). Letting

\[v = (v'_1 \circ \pi_1, \ldots, v'_r \circ \pi_r),\]

we have \((\mathcal{A}_i \cap \mathcal{S}_{2i}) \cup \{p_{i-1}\} \setminus \{p_i\} \subseteq D_i(v)\) for \(1 \leq i \leq r\) and so the dimension of \(D_i(v)\) is \(2n - 3\). By Draisma’s Lemma (Theorem 2.2) and the comments after, \(V_{\mathcal{T}_i}^{Z/2Z} \ast \cdots \ast V_{\mathcal{T}_r}^{Z/2Z}\) has the expected projective dimension, \(r(2n - 3) + r - 1\).

3.2 Group-Based Models with \(|G| > 2\)

We continue to assume the trees \(\mathcal{T}_1, \ldots, \mathcal{T}_r\) are binary and that \(B\) is trivial, but consider group \(G\) with \(|G| > 2\).

**Theorem 3.4** Let \(\mathcal{T}_1, \ldots, \mathcal{T}_r\) be binary phylogenetic \([n]-trees\) with \(n \geq 2r + 4\), and let \(G\) be an abelian group with \(|G| > 2\). Then, \(V_{\mathcal{T}_1}^G \ast \cdots \ast V_{\mathcal{T}_r}^G\) has the expected projective dimension, \((|G| - 1)r(2n - 3) + r - 1\).

The proof follows the same structure as Theorem 3.1, but requires slightly different arguments. One way that the case \(|G| > 2\) is actually simpler is that for any integer \(2 \leq j \leq n\) there exists a consistent leaf-labeling that labels exactly \(j\) leaves by non-identity elements of \(G\) (this is only the case for even \(j\) when \(G = \mathbb{Z}/2\mathbb{Z}\)). Indeed, we have the following useful fact: Let \(G\) be a finite group with order \(|G| > 2\). For any \(g \in G\) and any \(N \geq 2\), \(g\) can be expressed as the sum of exactly \(N\) non-identity elements.

We will again partition the vectors corresponding to consistent leaf-labelings based on the number of leaves labeled by non-identity elements. But, we will not need to “borrow” the vectors \(p_i\) from adjacent slices as in the proof of Theorem 3.1.

While in this section we are focused on binary trees, we prove Lemma 3.5 for the more general, non-binary case.

**Lemma 3.5** Let \(\mathcal{T}_1, \ldots, \mathcal{T}_r\) be phylogenetic (not necessarily binary) \([n]-trees\) with \(m_1, \ldots, m_r\) edges, respectively. Let \(\mathcal{A}_i\) be the set of vectors describing the monomial map parameterizing \(V_{\mathcal{T}_i}^G\). Then, there exists \(v = (v_1, \ldots, v_r)\) in \(\prod_{i=1}^r (\mathbb{R}^{|G|m_i})^*\) such that \(D_1(v)\) contains the vectors of \(\mathcal{A}_1\) corresponding to consistent leaf-labelings of \(\mathcal{T}_1\) with 0, 2, or 3 non-identity labels and \(D_i(v)\) contains the vectors of \(\mathcal{A}_i\) corresponding to consistent leaf-labelings of \(\mathcal{T}_i\) with 2\(i\) or 2\(i + 1\) non-identity labels for \(2 \leq i \leq r\).

**Proof** Each matrix \(\mathcal{A}_i\) is in \(\mathbb{R}^{|G|m_i}\), where \(m_i\) is the number of edges of \(\mathcal{T}_i\). Let \(\pi_i : \mathbb{R}^{|G|m_i} \rightarrow \mathbb{R}^{|G|^n}\) be the projection which forgets the coordinates of the non-leaf edges.
Let $\mathcal{H}_1, \ldots, \mathcal{H}_{r-1}$ be the parallel hyperplanes in $\mathbb{R}^{|G|^n}$ with $\mathcal{H}_j$ defined by

$$
\sum_{i \in [n]} \sum_{g \in G \setminus \{0\}} x_g^i = \frac{3}{2} + 2j.
$$

These planes partition $\mathbb{R}^{|G|^n}$ into sets $R_1, \ldots, R_r$ where the projections of labelings with 0, 2, or 3 non-identity leaves are in $R_1$ and projections of labelings with $2i$ or $2i + 1$ non-identity leaves are in $R_i$ for $2 \leq i \leq r$. By Proposition 2.4, there is a sequence of functionals $v' = (v'_1, \ldots, v'_r)$ with $v'_i$ the minimum on $R_i$ and $v = (v'_1 \circ \pi_1, \ldots, v'_r \circ \pi_r)$ is the desired functional.

**Lemma 3.6** Let $T_1, \ldots, T_r$ be phylogenetic $[n]$-trees with $n \geq 2r + 4$. Let $1 \leq i \leq r$ and let $v$ and $D_i(v)$ be as in Lemma 3.5. If $T_i$ is binary, then

$$
\dim(D_i(v)) \geq (|G| - 1)(2n - 3) + 1.
$$

**Proof** It will be convenient to work with the dehomogenized vectors, so let $\pi : \mathbb{R}^{|G|(2n-3)} \to \mathbb{R}^{|G|-1)(2n-3)}$ be the projection that forgets coordinates $x_0^E$ for identity element $0 \in G$ for each edge $E \in \mathcal{E}(T_i)$. In order to prove the lemma, we will show that for any edge $E \in \mathcal{E}(T_i)$ and any non-identity $g \in G$, that $e_g^E$ is in the span of $\pi(D_i(v))$.

Choose any internal vertex $v$ of $T_i$ and redraw $T_i$ as below where each $r_j$ is a rooted subtree of $T_i$ with root $\rho_j$.

Choose the orientation of the edges as shown and so that all edges in $r_j$ are directed away from $\rho_j$. For $1 \leq j \leq 3$, choose $L_j$ to be a subset of the leaves of $r_j$ so that each $|L_j|$ is odd and so that $|L_1| + |L_2| + |L_3| = 2i + 1$. This is always possible since $i \leq r$ and $n \geq 2r + 4$. Now, we will distinguish a particular leaf in each $r_j$. There is a unique path from the root $\rho_j$ down $r_j$ such that every vertex on the path has an odd number of leaves in $L_j$ as descendants. Label the terminal vertex of this path in $r_j$ by $\lambda_j$. Construct a consistent leaf-labeling that labels each of the $\lambda_j$ by the identity and the other $2i - 2$ leaves by non-identity elements of $G$ so that in the induced consistent edge-labeling, every edge on the paths from $\rho_j$ to the $\lambda_j$ is labeled by the identity. By...
our construction, this is always possible using any single non-identity element and its inverse. Let $F$ be the vector that corresponds to this consistent leaf-labeling.

Now, fix non-identity $g \in G$. For any $h \in G$, let $f^j_h$ be the vector with a 1 in the entry corresponding to $e^E_h$ for each edge $E$ on the path from leaf $\lambda_j$ to $v$ and all other entries equal to zero. Then, for any $h \in G$, $(f^1_h + f^2_{-h} + F)$ and $(f^3_h + f^2_{-h} + F)$ are in $\langle \pi(D_1(v)) \rangle$, so their difference, $f^1_h - f^3_h$ is in the span. Likewise, $f^1_h - f^2_h$ is in the span. For any $2 \leq m \leq d - 1$, define the vector

$$
\begin{align*}
v_m^1 := & \left( f_{-mg}^1 + f_{g}^2 + f_{(m-1)g}^3 + F \right) + \left( f^1_g - f^2_g \right) + \left( f_{(m-1)g}^1 - f_{(m-1)g}^3 \right) \\
- & \left( f_{mg}^1 + f_{-mg}^2 + F \right) - \left( f_{mg}^1 - f_{-mg}^1 \right) \\
= & \left( -f_{mg}^1 + f^1_g + f_{(m-1)g}^3 \right).
\end{align*}
$$

Then, since each $v_m^1 \in \langle D_1(v) \rangle$, the vector

$$
\left( f^1_g + f^2_{(d-1)g} + F \right) + \left( f^1_{(d-1)g} - f^2_{(d-1)g} \right) + v_{d-1}^1 + \cdots + v_2^1 = df^1_g + F
$$

is in $\langle D_1(v) \rangle$. Let $a, b, c \in G$ be non-identity elements with $a + b + c = 0$. Then, $\langle D_1(v) \rangle$ also contains the vector

$$
\begin{align*}
(f^1_a + f^2_{-a} + F) + (f^2_b + f^3_{-b} + F) + (f^3_{-c} + f^3_c + F) \\
- (f^1_a + f^2_{b} + f^3_{c}) - (f^3_{-c} + f^2_{-a} + f^3_{-b} + F) \\
= F.
\end{align*}
$$

Consequently, we have $f^1_g \in \langle \pi(D_1(v)) \rangle$.

To show that $e^E_g$ is in $\langle \pi(A \cap H^-) \rangle$ we perform induction on the length of the path represented by $f^E_g$. For any leaf $E$, we can choose the vertex $v$ incident to this leaf edge so that $f^1_g = e^E_g$. If the path represented by $f^E_g$ has length $n$, all the edges $E' \neq E$ in the path have their path to a leaf shorter than $n$, so we assume that $e^E_{g'}$ is in the span. Subtracting these from $f^E_g$ leaves only $e^E_g$, so it is in the span as well.

This proves that $\langle \pi(D_1(v)) \rangle$ has full dimension $|G| - 1(2n - 3)$. It remains to show that $\langle D_1(v) \rangle$ has dimension one greater. In the above argument, $df^1_g$ is produced as an integer combination of vectors in $\pi(D_1(v))$. Note that the sum of the integer coefficients of this linear combination is 0, so there is a vector $w \in \langle D_1(v) \rangle$ with $\pi(w) = f^1_g$ satisfying $\sum_{g \in G} x^E_g = 0$ for all edges $E$; in particular, $w$ is the vector with a 1 in the entry corresponding to $e^E_g$, a $-1$ in the entry corresponding to $e^E_0$ for each edge $E$ on the path from $\lambda_1$ to $v$, and a zero in all other entries. Let $Z$ denote the subspace of $\mathbb{R}^{|G|(2n-3)}$ defined by the $2n - 3$ equations $\sum_{g \in G} x^E_g = 0$, so $Z$ has dimension $(|G| - 1)(2n - 3)$. Similarly, each vector $e^E_g \in \langle \pi(D_1(v)) \rangle$ can be produced as a linear combination of projections of vectors in $Z$, so $e^E_g - e^E_0 \in \langle D_1(v) \rangle$ for all edges $E$ and all $g \in G \setminus \{0\}$. Therefore $\langle D_1(v) \rangle$ contains $Z$. Note, however, that any vector in $D_1(v)$ is outside of $Z$, so

$\square$ Springer
\[ \dim \langle D_i(v) \rangle \geq (|G| - 1)(2n - 3) + 1. \]

Combining Lemma 3.5 and 3.6 gives a proof of Theorem 3.4.

4 Non-binary Trees and Non-general Group-Based Models

In this section, we show that many of the results of the previous sections generalize to non-binary trees and to group-based models where we identify the parameters of some group elements. First we tackle the case of non-binary trees.

4.1 Non-binary Trees

To prove the result for non-binary trees, we use the fact that any [n]-tree with no degree-two vertices can be resolved into a binary [n]-tree. We then apply our construction for binary trees to a resolution of each tree and adapt this construction to obtain an analogous result for the unresolved tree.

Lemma 4.1 Let \( T_1, \ldots, T_r \) be phylogenetic [n]-trees with \( n \geq 2r + 4 \) and let \( m_i \) be the number of edges of \( T_i \) for \( 1 \leq i \leq r \). There exist functionals \( v = (v_1, \ldots, v_r) \) on the parameter spaces of \( V_{T_1}^G, \ldots, V_{T_r}^G \) such that \( \dim \text{conv}(D_i(v)) = (|G| - 1)m_i \) for \( 1 \leq i \leq r \).

Proof Let \( S_i \) be a binary phylogenetic [n]-tree obtained by resolving \( T_i \). That is, \( T_i \) is obtained after contracting \( 2n - 3 - m_i \) internal edges of \( S_i \).

Again we will work with dehomogenized vectors. For \( 1 \leq i \leq r \), let \( \pi'_i : \mathbb{R}^{|G|(2n-3)} \to \mathbb{R}^{|G|(-1)(2n-3)} \) be the projection that forgets coordinates \( x_0^G \) for identity element \( 0 \in G \) from the parameter space for \( V_{S_i}^G \). Let \( \pi_i : \mathbb{R}^{|G|m_i} \to \mathbb{R}^{|G|(-1)m_i} \) be the equivalent projection for \( V_{T_i}^G \).

Let \( p_i : \mathbb{R}^{|G|(-1)(2n-3)} \to \mathbb{R}^{|G|(-1)m_i} \) be the projection from the reduced parameter space of \( V_{S_i}^G \) to the reduced parameter space of \( V_{T_i}^G \) that forgets the coordinates of the parameters for the contracted edges. The kernel of \( p_i \) has dimension \((|G| - 1)(2n - 3 - m_i)\). If \( A'_i \) is the set of exponent vectors associated with the toric variety \( V_{S_i}^G \) and \( A_i \) the set of vectors corresponding to \( V_{T_i}^G \), then \( \pi_i(A_i) = p_i(\pi'_i(A'_i)) \).

Choose functionals \( v' = (v'_1, \ldots, v'_r) \in \prod_{i=1}^r (\mathbb{R}^{|G|(2n-3)})^* \) as in the proof of Theorem 3.1 or Theorem 3.4 (depending on whether \( G = \mathbb{Z}/2\mathbb{Z} \)). As shown above, \( \dim \text{conv}(D_i(v')) = (|G| - 1)(2n - 3) \) since \( \pi'_i \) is independent of the dimension of the subspaces of \( A'_i \). Let \( v = (v_1, \ldots, v_r) \) be in \( \prod_{i=1}^r (\mathbb{R}^{|G|m_i})^* \) where \( v_i \) is \( v'_i \) after projecting away the \( |G|(2n - 3 - m_i) \) entries corresponding to the contracted edges. Each \( v'_i \) depends only on the coordinates of the leaf edges of \( S_i \), so each consistent labeling has the same evaluation by \( v' \) and \( v \). The minimum value is achieved at the same index \( i \) so

\[ \pi_i(D_i(v)) = p_i \left( \pi'_i(D_i(v')) \right). \]
Therefore, there is a functional $v$. Proof: Lemma 4.1 proves the case that $|v|$. In which case $\dim \text{conv}(\mathcal{A}_i)$ has dimension $(|G| - 1)m_i$.

However, $\dim \text{conv}(D_i(v))$ cannot exceed $(|G| - 1)m_i$ because $D_i(v) \subseteq \mathcal{A}_i$ and the affine span of $\mathcal{A}_i$ has dimension $(|G| - 1)m_i$.

The above result allows us to extend Theorem 3.4 to non-binary trees, and henceforth we can work more generally with arbitrary trees with no degree-2 vertices.

### 4.2 Non-general Group-Based Models

With the results from the previous sections, we now have all of the pieces necessary to prove Theorem 1.1 for arbitrary trees and for the special case of general group-based models. Thus, the final step is to show that the same results hold for the non-general group-based models in which we allow Fourier parameters to be identified according to a non-trivial subgroup $B$ of $\text{Aut}(G)$. As discussed previously, many of the most commonly used models in phylogenetics, including the JC and K2P models, are non-general group-based models. Recall that all group elements in the same orbit of $B$ are assigned the same parameter.

**Lemma 4.2** Let $T_1, \ldots, T_r$ be phylogenetic $[n]$-trees with $n \geq 2r + 4$ and let $m_i$ be the number of edges of $T_i$ for $1 \leq i \leq r$. Let $G$ be a finite abelian group with $B$ a subgroup of $\text{Aut}(G)$. Then, there exist functionals $v = (v_1, \ldots, v_r)$ on the parameter spaces of $V_{T_1}^{(G,B)}, \ldots, V_{T_r}^{(G,B)}$ such that $\dim \text{conv}(D_i(v)) = lm_i$ for $1 \leq i \leq r$.

**Proof** Lemma 4.1 proves the case that $B$ is trivial. Assume that $B$ is non-trivial in which case $|G| > 2$.

For $1 \leq i \leq r$, let $p_i : \mathbb{R}^{|G|m_i} \rightarrow \mathbb{R}^{(|G| - 1)m_i}$ be the projection from the parameter space of $V_{T_i}^{(G,1)}$ to the parameter space of $V_{T_i}^{(G,B)}$ by summing the coordinates of the parameters that are identified in $B$ for each edge. The kernel of $p_i$ has dimension $(|G| - 1)m_i$. If $\mathcal{A}_i'$ is set of exponent vectors associated to toric variety $V_{T_i}^{(G,1)}$ and $\mathcal{A}_i$ the vectors to $V_{T_i}^{(G,B)}$ then $\mathcal{A}_i = p_i(\mathcal{A}_i')$.

Choose functionals $v' = (v'_1, \ldots, v'_r)$ with $v'_i$ acting on $\mathbb{R}^{|G|m_i}$ as in the proof of Lemma 4.1. As was shown above, $D_i(v')$ has affine dimension $(|G| - 1)m_i$. Because we are in the case $|G| > 2$, each $v'_i$ depends only on the total number of non-identity leaf labels, which does not change when parameters are identified according to $B$. Therefore, there is a functional $v_i$ on $\mathbb{R}^{(|G| - 1)m_i}$ such that $v'_i = v_i \circ p_i$. Let $v = (v_1, \ldots, v_r)$. Each consistent labeling has the same evaluation by $v'$ and $v$ so the minimum value is achieved at the same index $i$. Then,

$$D_i(v) = p_i(D_i(v')).$$ 

The dimension of $D_i(v)$ has the bound

$$\dim \text{conv}(D_i(v)) \geq \dim \text{conv}(D_i(v')) - \dim \ker(p_i) = lm_i.$$
Since \( \text{dim}_i \) is the dimension of the affine span of \( \mathcal{A}_i \), it is also an upper bound on \( \text{dim conv}(D_i(v)) \).

**Proof of Theorem 1.1** Applying Draisma’s Lemma to Lemma 4.2 shows that for \( T_1, \ldots, T_r \) phylogenetic \([n]\)-trees (not necessarily binary) with \( n \geq 2r + 5 \), \( G \) an abelian group, and \( B \) a subgroup of \( \text{Aut}(G) \), the join variety \( V^{(G,B)}_{T_1} \ast \cdots \ast V^{(G,B)}_{T_r} \) has projective dimension \( M + r - 1 \) where \( M \) is the sum of the number of edges among \( T_1, \ldots, T_r \) and \( l + 1 \) is the number of orbits of \( B \) in \( G \). This is the expected dimension, so we have completed the proof of Theorem 1.1 in full generality.

\[ \square \]

### 5 Improved Bounds for Special Cases

Our proof of Theorem 1.1 holds when \( r \), the number of phylogenetic tree models in the mixture, is not too large compared to the \( n \), the number of leaves of the trees, according to the bound \( n \geq 2r + 5 \). It should be noted though that the bound \( n \geq 2r + 5 \) merely reflects the limitations in our proof techniques. In our experiments, we have not come across any defective mixtures of phylogenetic tree models, and we have no reason to believe that these models have defective join dimensions for larger \( r \), so we state the following conjecture.

**Conjecture 5.1** Let \( T_1, \ldots, T_r \) be phylogenetic \([n]\)-trees with \( n \geq 3 \), and let \( G \) be an abelian group and \( B \subset \text{Aut}(G) \). Then \( V^{(G,B)}_{T_1} \ast \cdots \ast V^{(G,B)}_{T_r} \) has the expected dimension.

#### 5.1 Claw Trees

In some special cases, we can improve the bound. For instance, Theorem 3.4 states that when the group \( G \) has order at least 3, then joins have the expected dimension for \( n \geq 2r + 4 \). When each tree in the mixture is the \( n \)-leaf claw tree, we can improve this bound.

**Proposition 5.2** Let \( T \) be the \([n]\)-leaf claw tree and let \( G \) be an abelian group and \( B \subset \text{Aut}(G) \). Then, for \( n \geq 2r + 1 \), the \( r \)th secant variety \( \sigma_r(V^{(G,B)}_T) \) has the expected dimension.

**Proof** The proof outline follows that of Theorems 3.1 and 3.4, but with simplifications that allow for the improved bound \( n \geq 2r + 1 \).

For \( G = \mathbb{Z}/2\mathbb{Z} \), as in Lemma 3.2 let \( S_c \) the hyperplane defined by

\[
\sum_{E \in \mathcal{E}(T)} x_E^c = c.
\]

We show that for even \( 2 \leq c \leq n - 1 \), \( \text{rank}(\mathcal{A} \cap S_c) = n \). For any pair of edges \( E_1, E_2 \), let \( L \) be any collection of \( c - 1 \) edges not containing \( E_1 \) or \( E_2 \) and \( F = \sum_{j \in L} e_j^1 \). Then, in \( \mathcal{A} \cap S_c \) is

\[
(e_1^{E_1} + F) - (e_1^{E_2} + F) = e_1^{E_1} - e_1^{E_2}.
\]
Fixing edge $E$, now let $L$ be any set of $c - 1$ edges not containing $E$ and again $F = \sum_{j \in L} e^j_1$. The vector

$$ (e^E_1 + F) + \sum_{j \in L} (e^E_1 - e^j_1) = ce^E_1 $$

is in $\langle A \cap S_c \rangle$ for every edge $E$ so rank $\langle A \cap S_c \rangle = n$.

The remainder of the proof exactly follows the proof of Theorem 3.1 replacing dimension $2n - 3$ (the number of edges of a binary $[n]$-tree) with $n$ (the number of edges of the claw $[n]$-tree).

For $|G| > 2$, as in Lemma 3.5 choose functionals $v = (v_1, \ldots, v_r)$ that divide the vectors corresponding to consistent leaf-labelings so that $D_i(v)$ contains the leaf-labelings with $2i$ or $2i + 1$ non-identity edges for $1 \leq i \leq r$.

Let $\pi : \mathbb{R}^{\lvert G \rvert \times n} \to \mathbb{R}^{(|G|-1)n}$ be the dehomogenization map that forgets the coordinates of $x^E_0$ for each edge $E$. Working in the dehomogenized coordinates, fix $i$ and any edge $E_1$. Let $L$ be any collection of $2i - 2$ edges not containing $E_1$ and $F = \sum_{j \in L} e^j_1$. Choose $E_2$ and $E_3$ to be additional edges not in $L$. Follow the argument in the proof of Lemma 3.6 but replacing each $f^j_h$ with $e^E_j$. This shows that $e^E_1 \in \langle \pi(D_i(v)) \rangle$ for any non-identity $g \in G$, and consequently that

$$ \dim \langle D_i(v) \rangle = (|G| - 1)n + 1. $$

Applying Draisma’s Lemma, this proves the result for $B$ trivial. The argument in the proof of Lemma 4.2 can be applied here for the case that $B$ is non-trivial.

5.2 Trees with Few Leaves

For a specific value of $r$ and a specific model $M = (G, B)$, there are a finite number of collections of $[n]$-trees $T_1, \ldots, T_r$ with $n < 2r + 5$, so one can check whether all joins have the expected dimension by explicit computation. We perform some of these computations in the computer algebra system Macaulay2 (Grayson and Stillman 2002) using the package PhylogeneticTrees (Baños et al. 2016).

One can efficiently compute the dimensions of joins of parametrized varieties using the principle of Terracini’s Lemma. The dimension of a join is equal to the dimension of the tangent space at a generic point on the join variety. Choosing random parameter values, we obtain the tangent space dimension from the rank of the Jacobian of the parameterization map. To further improve efficiency, we compute the rank over a finite field $\mathbb{F}_p$ for a large prime $p$.

Note that this algorithm is probabilistic. With small probability the random parameter values may be non-generic, leading to a drop in the dimension of the tangent space. Additionally, for some parameter values there may be more linear dependencies in the Jacobian over $\mathbb{F}_p$ than over characteristic zero. Both situations produce a lower value than the true dimension, so this algorithm only certifies a lower bound. However, if the algorithm returns a value equal to the expected dimension, it is a proof that the join is not defective.
Here, we state the results that we are able to obtain by combining Theorem 1.1 and computational results obtained for small \( n \).

**Proposition 5.3** Let \( T \) be a phylogenetic \([n]\)-tree with \( n \geq 3 \). Then, the second secant variety \( V^M_T \ast V^M_T \) has the expected dimension for \( M \) equal to CFN, JC, K2P or K3P.

This statement was previously proved for models JC and K2P in Allman et al. (2011). The case of \( n \geq 9 \) in Proposition 5.3 is proved by Theorem 1.1, and \( n \geq 8 \) for \( M \neq \text{CFN} \) by Theorem 3.4. For each model \( M \), we check the dimension of \( V^M_T \ast V^M_T \) for all \([n]\)-trees \( T \) with \( n \leq 7 \) and \( V^Z_T \ast V^Z_T \) for \([8]\)-trees \( T \). The dimension of the secant variety is invariant under relabeling the leaves of \( T \), so we need only check one tree for each equivalence class. The equivalence classes correspond to the set of unlabeled trees with \( n \) leaves.

**Proposition 5.4** Let \( T \) be a phylogenetic \([n]\)-tree with \( n \geq 3 \). Then, the third secant variety \( V^M_T \ast V^M_T \ast V^M_T \) has the expected dimension for \( M \) equal to CFN, and for \( M \) equal to JC, K2P or K3P with \( n \neq 9 \).

For Proposition 5.4, the case of \( n \geq 11 \) is proved by Theorem 1.1, and \( n \geq 10 \) for \( M \neq \text{CFN} \) by Theorem 3.4. For \( M = \text{CFN} \), we check the dimension for all \([n]\)-trees \( T \) with \( n \leq 10 \). For \( M \) equal to JC, K2P or K3P, we were able to check the dimension for all \([n]\)-trees \( T \) with \( n \leq 8 \), but were not able to complete the computation on \([9]\)-trees.

**Proposition 5.5** Let \( T_1, T_2 \) be phylogenetic \([n]\)-trees with \( n \geq 3 \). Then, \( V^M_{T_1} \ast V^M_{T_2} \) has the expected dimension for \( M \) equal to CFN, and for \( M \) equal to JC, K2P or K3P with \( n \neq 7 \).

For Proposition 5.5, the case of \( n \geq 9 \) is proved by Theorem 1.1, and \( n \geq 8 \) for \( M \neq \text{CFN} \) by Theorem 3.4. For \( M = \text{CFN} \), we check the dimension of \( V^M_{T_1} \ast V^M_{T_2} \) for all pairs \((T_1, T_2)\) of \([n]\)-trees with \( n \leq 8 \). Again, the dimension of the join is invariant under permutations of \([n]\), but a relabeling applies to both trees in the pair \((T_1, T_2)\). Therefore, we can let \( T_1 \) vary over the set of unlabeled trees, but \( T_2 \) must then be chosen from the full set of labeled \([n]\)-trees. For \( M \) equal to JC, K2P or K3P, we were able to check the dimension for all pairs of \([n]\)-trees with \( n \leq 6 \), but were not able to complete the computation on pairs of \([7]\)-trees.

**Remark 3** Note that while we resolve most of the cases for binary trees with few leaves in this section for 3-tree secant varieties and 2-tree joins with respect to the JC, K2P, and K3P models, the \( n = 8 \) case for 3-tree secant varieties and the \( n = 7 \) case for 2-tree joins remains an open computational question.

In conclusion, we have shown that, for \( n \geq 2r + 5 \), the join varieties associated with a large class of group-based models, including the CFN, JC, K2P, and K3P models have the expected dimension. In order to provide a complete answer to Conjecture 5.1, we expect different proof techniques would need to be used to handle the \( n < 2r + 5 \) case. However, we showed how this bound could be improved for the case of claw trees in the proof of Proposition 5.2.
Acknowledgements  This work began at the 2016 AMS Mathematics Research Community on “Algebraic Statistics,” which was supported by the National Science Foundation under Grant number DMS-1321794. RD was supported by NSF DMS-1401591. EG was supported by NSF DMS-1620109. RW was supported by a NSF GRF under Grant number PFG-031543, NSF RTG Grant 0943832, and a Ford Foundation Dissertation Fellowship. HB was supported in part by a research assistantship, funded by the National Institutes of Health Grant R01 GM117590. PEH was partially supported by NSF Grant DMS-1620202.

References

Abo H, Brambilla MC (2012) New examples of defective secant varieties of Segre–Veronese varieties. Collectanea Mathematica 63(3):287–297
Abo H, Brambilla MC (2013) On the dimensions of secant varieties of Segre-Veronese varieties. Annali di Matematica Pura ed Applicata 192(1):1–32
Alexander J, Hirschowitz A (1995) Polynomial interpolation in several variables. J Algebraic Geom 4(2):201–222
Allman ES, Petrović S, Rhodes JA, Sullivant S (2011) Identifiability of 2-tree mixtures for group-based models. IEEE/ACM Trans Comput Biol Bioinf 8(3):710–722
Baños H, Bushek N, Davidson R, Gross E, Harris PE, Krone R, Long C, Stewart A, Walker R (2016) Phylogenetic trees. arXiv preprint arXiv:1611.05805 (2016)
Buczynska W, Wismiowski JA (2007) On the geometry of binary symmetric models of phylogenetic trees. J Eur Mathe Soc 9(3):609–635
Casanellas M (2012) Algebraic tools for evolutionary biology. Math Soc 12–17
Daskalakis C, Mossel E, Roch S (2011) Evolutionary trees and the ising model on the Bethe lattice: a proof of steel’s conjecture. Probab Theory Rel Fields 149(1):149–189
Draisma J (2008) A tropical approach to secant dimensions. J Pure Appl Algebra 212(2):349–363
Evans SN, Speed TP (1993) Invariants of some probability models used in phylogenetic inference. Ann Stat 21(1):355–377
Grayson DR, Stillman ME (2002) Macaulay 2, a software system for research in algebraic geometry
Hendy MD, Penny D, Steel MA (1994) A discrete fourier analysis for evolutionary trees. Proc Natl Acad Sci 91(8):3339–3343
Jukes TH, Cantor CR (1969) Evolution of protein molecules. Mamm Protein Metab 3(21):132
Long C, Sullivant S (2015) Identifiability of 3-class Jukes–Cantor mixtures. Adv Appl Math 64:89–110, 3
Mauhar M, Rusinko J, Vernon Z (2017) H-representation of the kimura-3 polytope for the m-claw tree. SIAM J Discret Math 31(2):783–795
Michałek M (2011) Geometry of phylogenetic group-based models. J Algebra 339(1):339–356
Neyman J (1971) Molecular studies of evolution: a source of novel statistical problems. Stat Decis Theory Rel Top 1:1–27
Pagel M, Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53(4):571–581
Sturmfels B (1996) Gröbner bases and Convex Polytopes, vol 8. American Mathematical Society, Providence
Sturmfels B, Sullivant S (2005) Toric ideals of phylogenetic invariants. J Comput Biol 12(2):204–228
Sullivant S (2018) Algebraic statistics. Graduate studies in mathematics. American Mathematical Society, Providence

Affiliations

Hector Baños1 · Nathaniel Bushek2 · Ruth Davidson3 · Elizabeth Gross4 · Pamela E. Harris5 · Robert Krone6 · Colby Long7 · Allen Stewart8 · Robert Walker9

Hector Baños
hdbanoservantes@alaska.edu
Nathaniel Bushek
bush0320@d.umn.edu

Ruth Davidson
redavid2@illinois.edu

Pamela E. Harris
pamela.e.harris@williams.edu

Robert Krone
rckrone@ucdavis.edu

Colby Long
long.1579@mbi.osu.edu

Allen Stewart
stewaral@seattleu.edu

Robert Walker
robmarsw@umich.edu

1 Department of Mathematics and Statistics, University of Alaska Fairbanks, Fairbanks, AK 99775, US
2 Department of Mathematics and Statistics, University of Minnesota Duluth, Duluth, MN 55812, USA
3 Department of Mathematics, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
4 Department of Mathematics, University of Hawai’i at Mānoa, Honolulu, HI 96822, USA
5 Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267, USA
6 Department of Mathematics, University of California Davis, Davis, CA 95616, USA
7 Mathematical Biosciences Institute, The Ohio State University, Columbus, OH 43210, USA
8 Department of Mathematics, Seattle University, Seattle, WA 98122, USA
9 Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, USA