DRAFT - Toric Geometry and Homogeneous Phylogenetic Models

Nicholas Eriksson
Department of Mathematics
University of California, Berkeley
eriksson@math.berkeley.edu

ABSTRACT

We consider the model on phylogenetic trees in which every node of the tree is an observed binary random variable and the transition probabilities are given by the same matrix on each edge. The ideal of algebraic invarients is a toric ideal, we show the results of many calculations of minimal generating sets for these ideals. Further, we prove that the normal fan of the toric variety has a finite number of types for several families of trees and give data for this normal fan for all trees.

1. PHYLOGENETIC TREES

Throughout, let T be a rooted tree with $n-1$ edges (and therefore n nodes). To each node v_i of T we associate a κ-ary random variable X_i. We write $\rho(v)$ for the parent of the node v.

The transition probabilities between $\rho(v)$ and v are given by a κ by κ matrix $A^{(v)}$. These transition matrices are generally picked from some specific families such as the Jukes- In this paper we consider the model where all $A^{(v)}$ are equal and all nodes are binary. We write $A^{(v)} = A = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix}$. Notice that by subdividing the edges of the tree we can approximate the more general model with this homogeneous model.

The probability of observing i at a node v is computed from the parent of v in the tree as follows

$$P(X_v = i) = a_{i0}P(X_{\rho(v)} = 0) + a_{i1}P(X_{\rho(v)} = 1).$$

We assume that the root of the tree has uniform distribution. We are interested in the algebraic relations satisfied by the joint distribution

$$p_{i_1 i_2 \dots i_n} := P(X_1 = i_1, \dots, X_n = i_n).$$

Writing the joint distribution in terms of the model parameters $a_{00}, a_{01}, a_{10}, a_{11}$ we have

$$p_{i_1 i_2 \dots i_n} := \prod_{j=2}^{n} a_{i_{\rho(j)} i_j} \qquad (1)$$

where the nodes are labeled $1 ton + 1$ starting with the root. That is to say, the probability of observing a certain labeling of the tree is the product of the a_{ij} corresponding to the transitions down all edges of the tree.

The 4 parameters a_{ij} trace out a variety of dimension 4 in $\mathbb{R}^{2^{n-1}}$. This is a toric variety - it is parameterized by monomials. The ideal of phylogenetic invariants of this model is the ideal of this variety. We denote this toric ideal by I_T.

To be explicit, the invariants vanish for a given distribution $(p_{i_1 \dots i_n})$ exactly when that distribution comes from our model. Thus the knowlegde of the generators of this ideal is potentially very useful for the fitting of data to a phylogenetic tree.

1.1 Vitterbi Sequences

Let A_T be the 4 by 2^n matrix where each column is the exponent vector of the a_{ij} in (1) for some $(i_1, \dots, i_n) \in \{0,1\}^n$. Let P_T be the convex hull of the columns of A_T. This is a polytope in $\mathbb{R}^4$, however since all the monomials in (1) have the same degree $n-1$, we see that this polytope is actually contained in $n-1$ times the unit simplex in $\mathbb{R}^4$. Thus, P_T is actually a 3 dimensional polytope. We call P_T the *Vitterbi* polytope.

This notation comes from a dual question. Given any observation $(i_1, \dots, i_n)$ of the tree, which matrices $A = (a_{ij})$ make $p_{i_1 \dots i_n}$ maximal among the coordinates of the distribution p?

To solve this problem, transform to logarithmic coordinates $b_{ij} = \log(a_{ij})$. Then the condition that $p_{i_1 \dots i_n} > p_{l_1 \dots l_n}$ for all $(l_1, \dots l_n) \in \{0,1\}^n$ is translated into the the linear system of inequalities

$$b_{x_1 x_2} + \cdots + b_{x_{\rho(n)} x_n} > b_{l_1 l_2} + \cdots + b_{l_{\rho(n)} l_n}$$

for all $(l_1, \dots l_n) \in \{0,1\}^n$

The set of solutions is a polyhedral cone. For most values of $i_1, \dots, i_n$, this cone will be empty. Those sequences $i_1, \dots, i_n$ for which the cone is maximal are called *Vitterbi* sequences.

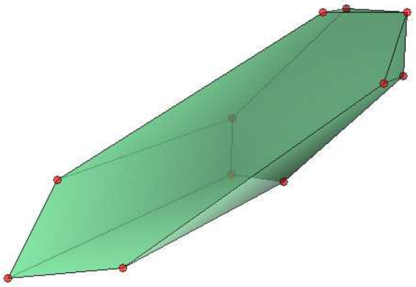


Figure 1: P_T for T a path with 6 edges

It is easy to see that the collection of these cones, as $(i_1, \dots, i_n)$ varies, is the normal fan of the polytope $P_T := Conv(A_T)$.

2. POLYTOPES

The polytopes P_T show remarkable finiteness properties as T varies. Since P_T is defined as the convex hull of 2^n vectors, it would seem that it could have arbitrarily bad structure. However, as it is contained in $n-1$ times the unit simplex, there are at most $O(n^3)$ integral points in P_T.

Example. Eric Kuo showed that if T is a path with n nodes, then P_T has only two combinatorial types for $n > 3$. The polytope for the path with 6 edges is shown in Figure 2. Think of this picture as a tetrahedron with the two vertices corresponding to all $0 \to 1$ and $1 \to 0$ transitions sliced off (since if a path has a $0 \to 1$ transition it must have a $1 \to x$ transition).

For all tested natural families of trees, this finiteness property seems to hold. In particular, if the tree is binary and completly balanced (that is, every leaf is at the same depth), a naive set of inequalities that hold for any binary tree actually characterize the polytope. Thus in some sense this tree is universal among binary trees.

Theorem: Let T be a balanced binary tree with more than 3 nodes. The associated polytope always has the same combinatorial type with 8 vertices and 6 facets, see Figure 2 and Figure 2.

However, arbitrary trees seem to have polytopes with many vertices. For example, Figure 2 and Figure 2 show a tree with 17 nodes that has a polytope with 34 vertices.

Table 2 shows data for all trees on at most 15 nodes. It appears that the maximum number of vertices grows approximately as $2n$. Notice that the tree with all leaves at depth 1 has P_T a tetrahedron, giving the unique minimum number, 4, of vertices for all trees.

Table 2 shows data for all binary tree on at most 21 nodes. It appears that the polytopes are much nicer, with the max-

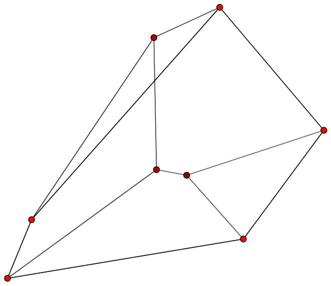


Figure 2: P_T for the complete binary tree with 15? nodes

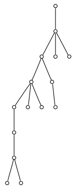


Figure 4: The tree associated to Figure 2

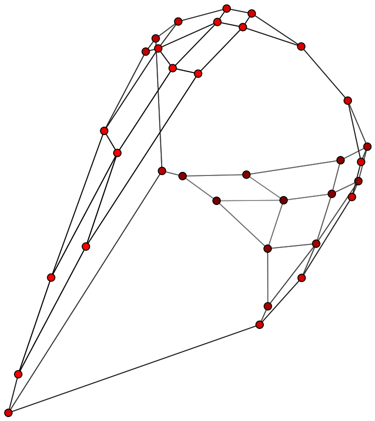


Figure 5: A polytope with 34 vertices

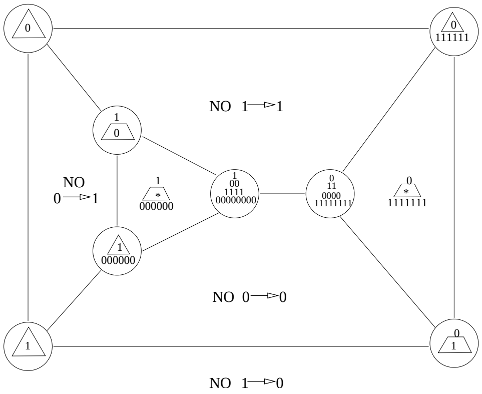


Figure 3: A Schlegel diagram for the polytope of the complete binary tree with facets labeled.

#nodes	#trees	min	max	ave
4	4	4	8	7
5	9	4	11	8
6	20	4	14	9.7
7	48	4	15	10.75
8	115	4	20	12.59
9	286	4	21	13.67
10	719	4	22	15.42
11	1842	4	25	16.60
12	4766	4	28	18.3
13	12486	4	31	19.5
14	32973	4	32	19.75
15	87811	4	34	22.6

Table 1: Minimum, maximum and average number of vertices of P_T over all trees with at most 15 nodes

imum number of vertices of P_T appearing to grow at most linearly. Thus we ask

Question: Is there a bound for the number of vertices of P_T as T ranges over all binary trees?

The obvious algorithm for computing P_T involves a 2^n loop followed by an elimination of duplicates and a convex hull computation. This algorithm can certainly be improved, but it is not know whether there is a polynomial time algorithm for constructing the polytope given a tree.

3. TORIC IDEALS

A toric ideal is an ideal parameterized by monomials or equivalently generated by binomials. A toric ideal has an (essentially) unique minimal generating set, called a Markov Basis. This can be calculated using 4ti2, by Raymond Hemmecke. Amazingly enough, 4ti2 was able to calculate Markov bases for ideals in 2048 variables.

The markov bases for binary trees are shown in Table 3.

Conjecture: The toric ideal corresponding to a binary tree is generated in degree 2.

It is conjectured that the quadratic generators correspond to moving subtrees betwee two trees in a compatible way. For an example, see Figure 3

#nodes	#trees	min	max	ave
3	1	4	4	4
5	1	7	7	7
7	2	8	10	9
9	3	8	13	11.33
11	6	10	14	11.66
13	11	11	13	11.91
15	23	8	16	14.35
17	46	12	17	13.82
19	98	10	20	14.65
21	207	8	19	14.8

Table 2: Minimum, maximum and average number of vertices of P_T over all binary trees with at most 21 nodes

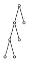
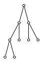
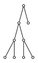
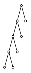
tree	degree	#MinGens	max deg
	4	4	2
	28	79	2
	92	441	2
	96	561	2
	210	2141	2
	220	2068	2
	210	2266	2

Table 3: Degree of I_T, munber of minimal generators, and maximum degree of the generators

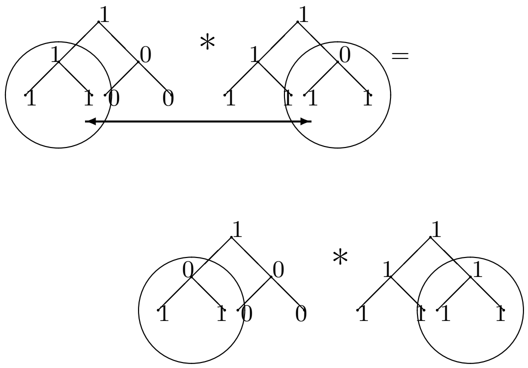


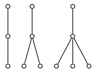
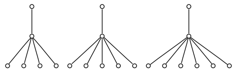
Figure 6: A quadratic generator

# nodes	degree	#MinGens	max deg	# of deg 3
3	6	6	3	2
4	19	32	3	4
5	36	102	3	6
6	61	259	3	8
7	90	540	3	10
8	127	1041	3	12
9	168	1842	3	14
10	217	3170	3	16

Table 4: Markov bases of paths

From examining the table of markov bases for paths (Table 3), we believe the following:

Conjecture: The toric ideal corresponding to a path is generated in degree 3.

Finally from examining Table 3 and the data we conjecture that the trees with highest degree generators are 
 . . .

4. ACKNOWLEGMENTS

The author appreciates many useful conversations with Seth Sullivant, Lior Pachter, and Bernd Sturmfels. The polytopal computations were done in `polymake`. The pictures were created with `javaview`. The program `4ti2` by Raymond Hemmecke [1] was used to compute the minimal generating set for the pylogenetic ideals.

nodes	trees	MinGens	Max deg
3	2	4-6	2-3
4	4	14-45	2-4
5	9	24-160	2-5
6	20	72-498	2-6
7	48	144-1342	2-7
8	115	282-3059	2-8

Table 5: Markov bases for all trees

5. REFERENCES

[1] R. Hemmecke and R. Hemmecke. 4ti2 version 1.1—computation of Hilbert bases, Graver bases, toric Gröbner bases, and more. Available at www.4ti2.de, Sept. 2003.