Identifiability of Phylogenetic Mixture Models

Elizabeth Allman, Sonja Petrović, John Rhodes, and Seth Sullivant

U. of Alaska– Fairbanks, U. of Illinois– Chicago, and NCSU

Harmony of Gröbner bases and the Modern Industrial Society Conference
Osaka, Japan

28 June 2010

The Main Results: Two-tree Mixtures

Theorem

The tree parameters of the phylogenetic mixture model $\mathcal{M}_{T_1} * \mathcal{M}_{T_2}$ are generically identifiable under the Jukes-Cantor and Kimura 2-parameter models if T_1, T_2 are trivalent with $n \geq 4$ leaves.

Theorem*

The continuous parameters of the phylogenetic mixture model $\mathcal{M}_{T_1} * \mathcal{M}_{T_2}$ are generically identifiable under the Jukes-Cantor and Kimura 2-parameter models if T_1, T_2 are trivalent with $n \geq 5$ leaves.

Outline

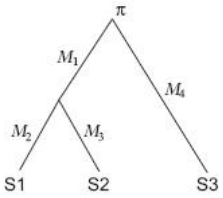
1. Phylogenetic Mixture Models
2. Group-based Phylogenetic Models
3. The Identifiability Problem
4. Proof of Tree Identifiability
    1. Quartets
    2. Sextets
5. Proof* of Parameter Identifiability
6. Some Mathematical Surprises

Phylogenetic Models

Let T be a trivalent tree with n leaves. Leaves are labeled by $[n] := \{1, 2, 3, \ldots, n\}$.

Associated to each edge of tree e is a Markov (structured) transition matrix M_e.

Once we specify T, and the M_e, get a probability distribution of characters at the leaves of the tree.



$$Prob(i, j, k) = \sum_{l=1}^{4} \sum_{m=1}^{4} r_l M_1(l, m) M_2(m, i) M_3(m, j) M_4(l, k)$$

Think of phylogenetic model as a map

$$\phi_T : \Theta \subseteq \mathbb{R}^k \to \Delta_{4^n}$$

Given by polynomials:
$\mathcal{M}_T := \mathrm{im}\phi_T = \phi_T(\Theta)$, is the phylogenetic model.

Phylogenetic Mixture Models

Suppose there are k classes of sites in the genome.
Each class $j \in [k]$ evolved according to tree T_j on n leaves.
Assuming that the classes are hidden, we observe a probability distribution of the form:

$$\phi_{T_1,\dots,T_k}(\pi, \{M_e\}) = \pi_1 \cdot \phi_{T_1}(\{M_e^1\}) + \pi_2 \cdot \phi_{T_2}(\{M_e^2\}) + \cdots + \pi_k \cdot \phi_{T_k}(\{M_e^k\})$$

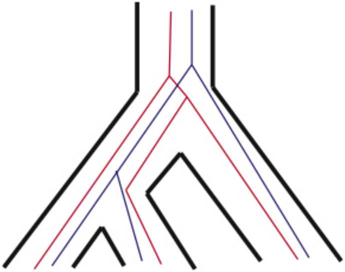
where π_j is the relative proportion of sites of class j.

Definition

Let $T_1, \dots, T_k$ be trees with n leaves. The phylogenetic mixture model

$$\mathcal{M}_{T_1} * \mathcal{M}_{T_2} * \cdots * \mathcal{M}_{T_k} = \left\{ \sum_{j=1}^{k} \pi_j p^j : \pi_j \geq 0, \sum \pi_j = 1, p^j \in \mathcal{M}_{T_j} \right\}.$$

Why Mixture Models?



- Differing gene tree topologies
- Could explain evolution with recombination

Group-based Models

For remainder we focus on group-based models. Phylogenetic models with structured transition matrices.

$$\begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix} \quad \begin{pmatrix} \alpha & \beta & \beta & \beta \\ \beta & \alpha & \beta & \beta \\ \beta & \beta & \alpha & \beta \\ \beta & \beta & \beta & \alpha \end{pmatrix} \quad \begin{pmatrix} \alpha & \beta & \gamma & \gamma \\ \beta & \alpha & \gamma & \gamma \\ \gamma & \gamma & \alpha & \beta \\ \gamma & \gamma & \beta & \alpha \end{pmatrix} \quad \begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \alpha & \delta & \gamma \\ \gamma & \delta & \alpha & \beta \\ \delta & \gamma & \beta & \alpha \end{pmatrix}$$

Cavender-Farris-Neyman (CFN), Jukes-Cantor (JC), Kimura 2-Parameter (K2P), Kimura 3-Parameter (K3P)

Transition structure is governed by a finite Abelian group G, such that

$$M_e(g,h) = f_e(g-h).$$

Theorem (Evans-Speed 1993, Hendy-Penny 1993)

Group-based models can be diagonalized by means of the discrete Fourier transform over G (Hadamard conjugation). In the Fourier coordinates, group-based models are toric varieties.

Fourier Coordinates

For each split $A|B$ in tree introduce a set of Fourier parameters

$$\{a_g^{A|B} : g \in G\}.$$

Theorem (Evans-Speed 1993, Hendy-Penny 1993)

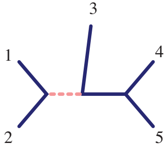
In the Fourier coordinates, a group-based phylogenetic model is given parametrically by:

$$q_{g_1,\ldots,g_n} = \begin{cases} \prod_{A|B \in \Sigma(T)} a^{A|B}_{\sum_{a \in A} g_a} & \text{if } g_1 + \cdots + g_n = 0 \\ 0 & \text{if } g_1 + \cdots + g_n \neq 0 \end{cases}$$

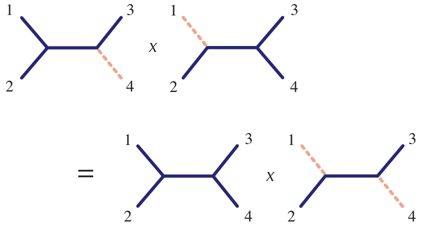
In the JC, K2P, K3P, we take $G = \mathbb{Z}_2 \times \mathbb{Z}_2 = \{A, C, G, T\}$.
In the K2P model, we have $a_G^{A|B} = a_T^{A|B}$ for all $A|B$
In the JC model, we have $a_C^{A|B} = a_G^{A|B} = a_T^{A|B}$ for all $A|B$.

Doodles for Group-Based Models

$$q_{CCTGC} = a^1_C a^2_C a^3_T a^4_G a^5_C a^{12|345}_A a^{123|45}_T$$



$$q_{CGTA} q_{ACTG} = q_{CGCG} q_{ATTA}$$



The Identifiability Problem

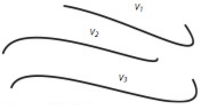
Definition

The tree parameters $T_1, \dots, T_k$ in a k-class phylogenetic mixture model are identifiable if for all

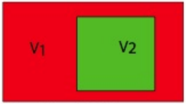
$$p \in \mathcal{M}_{T_1} * \cdots * \mathcal{M}_{T_k}$$

there does not exist another set of k trees $T'_1, \dots, T'_k$ such that

$$p \in \mathcal{M}_{T'_1} * \cdots * \mathcal{M}_{T'_k}.$$



Identifiable



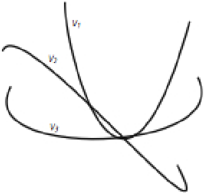
Not Identifiable

Generic Identifiability

Definition

The tree parameters in a k-class phylogenetic mixture model are generically identifiable if for all nonequal multisets $T_1, \ldots, T_k$, and $T'_1, \ldots, T'_k$,

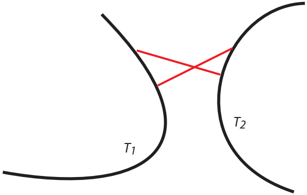
$$\dim(\mathcal{M}_{T_1} * \cdots * \mathcal{M}_{T_k} \cap \mathcal{M}_{T'_1} * \cdots * \mathcal{M}_{T'_k}) < \dim(\mathcal{M}_{T_1} * \cdots * \mathcal{M}_{T_k}).$$



Generic Identifiability of Continuous Parameters

Definition

Fix trees $T_1, \ldots, T_k$ on n leaves. The continuous parameters of phylogenetic mixture model are generically identifiable if $\phi_{T_1,\ldots,T_k}$ is one-to-one (off of a set of measure zero (up to label swapping)).



Past Work on Identifiability of Tree Mixtures

- Identifiability Results:
  - Allman and Rhodes (2006) $T_1 = \ldots = T_k$, $k < n$.
  - Stefankovic and Vigoda (2007) $T_1 = \ldots = T_k$, JC, K2P
  - Matsen, Mossel, and Steel (2008)
- Non-Identifiability Results:
  - Matsen and Steel (2007)
  - Stefankovic and Vigoda (2007)
  - Mossel and Vigoda (2005)

Algebraic Methods for Proving Identifiability

Proposition

Let $\mathcal{M}_0$ and $\mathcal{M}_1$ be two algebraic models. If there exist polynomials f_0 and f_1 such that

$$f_i(p) = 0 \text{ for all } p \in \mathcal{M}_i, \text{ and } f_i(p) \neq 0 \text{ for some } p \in \mathcal{M}_{1-i}, \text{ then}$$

$$\dim(\mathcal{M}_0 \cap \mathcal{M}_1) < \min(\dim \mathcal{M}_0, \dim \mathcal{M}_1).$$

Proposition

Let $\mathcal{M}_0$ and $\mathcal{M}_1$ be two algebraic models. If there is a polynomial f_0 such that

$$f_0(p) = 0 \text{ for all } p \in \mathcal{M}_0, \text{ and } f_0(p) \neq 0 \text{ for some } p \in \mathcal{M}_1, \text{ and}$$

$$\dim \mathcal{M}_1 \leq \dim \mathcal{M}_0 \text{ then}$$

$$\dim(\mathcal{M}_0 \cap \mathcal{M}_1) < \min(\dim \mathcal{M}_0, \dim \mathcal{M}_1).$$

Proof of Tree Parameter Identifiability

Theorem

The tree parameters of the phylogenetic mixture model $\mathcal{M}_{T_1} * \mathcal{M}_{T_2}$ are generically identifiable under the Jukes-Cantor and Kimura 2-parameter models if T_1, T_2 are trivalent with $n \geq 4$ leaves.

Strategy: Prove theorem for quartets $n = 4$ (using linear invariants), then lift to arbitrary sized trees:

Proposition

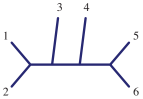
Let T_1, T_2, T_3, T_4 be n leaf trivalent trees. Suppose that there is a four element set $Q \subseteq [n]$ such that $\{T_1|_Q, T_2|_Q\} \neq \{T_3|_Q, T_4|_Q\}$. Then

$$\dim(\mathcal{M}_{T_1} * \mathcal{M}_{T_2} \cap \mathcal{M}_{T_3} * \mathcal{M}_{T_4}) < \dim(\mathcal{M}_{T_1} * \mathcal{M}_{T_2}).$$

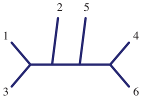
From Six to Infinity

Proposition

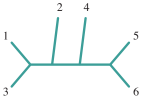
There are no quartet-matched pairs of trees with 5 leaves. The only pair of quartet-matched pairs of trees on 6 leaves are:



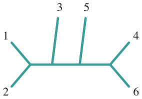
T_1



T_2



T_3



T_4

Proposition

There are linear invariants that distinguish T_1, T_2 from T_3, T_4.

Theorem (Matsen, Mossel, Steel 2007)

If two-tree mixtures are identifiable for trivalent trees with $n = 6$ trees, they are identifiable for all trees with $n \geq 6$ leaves.

Identifiability of Continuous Parameters

Theorem*

The continuous parameters of the phylogenetic mixture model $\mathcal{M}_{T_1} * \mathcal{M}_{T_2}$ are generically identifiable under the Jukes-Cantor and Kimura 2-parameter models if T_1, T_2 are trivalent with $n \geq 5$ leaves.

$$\begin{pmatrix} \alpha & \beta & \beta & \beta \\ \beta & \alpha & \beta & \beta \\ \beta & \beta & \alpha & \beta \\ \beta & \beta & \beta & \alpha \end{pmatrix} \qquad \begin{pmatrix} \alpha & \beta & \gamma & \gamma \\ \beta & \alpha & \gamma & \gamma \\ \gamma & \gamma & \alpha & \beta \\ \gamma & \gamma & \beta & \alpha \end{pmatrix}$$

Definition

Theorem* means that the result holds with high probability.

Proposition

Let $\phi : \mathbb{C}^d \to \mathbb{C}^m$ be a rational map. Then there is a Zariski open set $\Theta \subseteq \mathbb{C}^d$, such that $\#\phi^{-1}(\phi(\theta))$ is constant over Θ.

1. So to prove* the Theorem* for a particular size tree, generate random rational parameter choices θ and then symbolically solve the simultaneous polynomial system

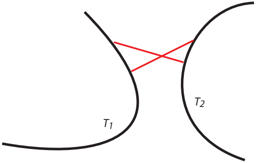
$$\phi(t) = \phi(\theta)$$

   and hope for one solution.
2. We check this using software SINGULAR, for JC and K2P on 4 and 5 leaf trees.
3. Recovering parameters uniquely on quartets $\implies$ recover edge lengths $\implies$ recover parameters on arbitrary sized trees.

Why $n = 5$ in Theorem*?

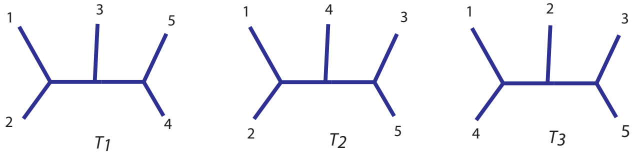
Proposition*

For T a four leaf tree under the Jukes-Cantor model, the continuous parameters in $\mathcal{M}_T * \mathcal{M}_T$ are not generically identifiable. The map $\phi_{T,T}$ is generically 6-to-1 (up to label swapping).



For biologically relevant parameters, we observed between 1 and 4 biologically relevant preimages.

Another Mathematical Surprise



Theorem

For the Jukes-Cantor model

$$\overline{\mathcal{M}_{T_2}} \subseteq \overline{\mathcal{M}_{T_1} * \mathcal{M}_{T_3}}.$$

Can the closure be dropped; i.e. does it happen for biologically meaningful parameter values?

Future Directions

- Develop methods to remove the * from a Theorem*
- Deal with the other group-based models (CFN, K3P)

  (K3P: current joint work with M. Casanellas - computational)
- Beyond group-based models, GTR, GMM
- Beyond 2-tree mixtures to *k*-tree mixtures

  (Recent work: M. Casanellas, J. Fernández-Sánchez, A. Kedzierska: some non-identifiability results)