Identifiability of Large Phylogenetic Mixture Models

John Rhodes and Seth Sullivant

University of Alaska–Fairbanks and NCSU

April 18, 2012

Main Result

Theorem (Rhodes-S 2011)

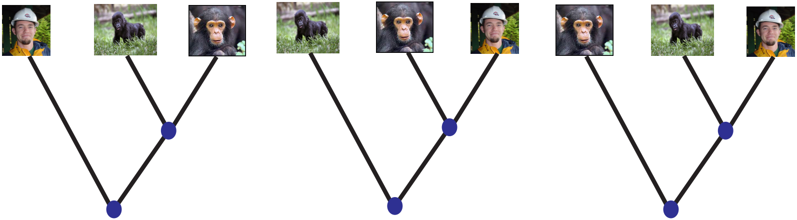
The tree and numerical parameters in a r-class, same tree phylogenetic mixture model on n-leaf trivalent trees are generically identifiable, if $r < 4^{\lceil n/4 \rceil}$.

- First result on numerical parameters.
- Exponential improvement over past results on this problem (Allman-Rhodes 2006)
- Large enough value of r for all practical uses
- Proofs depend on algebraic geometry
- New ideas: Large trees, tree and numerical parameters simultaneously

Phylogenetics

Problem

Given a collection of species, find the tree that explains their history.

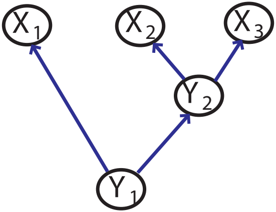


- Data consists of aligned DNA sequences from homologous genes

Human:	...ACCGTGCAACGTGAACGA...
Chimp:	...ACCTTGCAAGGTAAACGA...
Gorilla:	...ACCGTGCAACGTAAACTA...

Phylogenetic Models

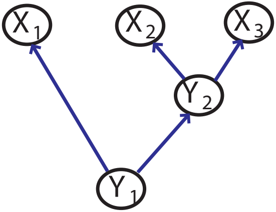
- Assuming site independence:
- Phylogenetic Model is a latent class graphical model
- Vertex $v \in T$ gives a random variable $X_v \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$
- All random variables corresponding to internal nodes are latent



$$P(x_1, x_2, x_3) = \sum_{y_1} \sum_{y_2} P(y_1) P(y_2|y_1) P(x_1|y_1) P(x_2|y_2) P(x_3|y_2)$$

Phylogenetic Models

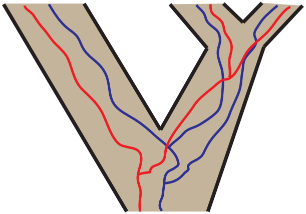
- Assuming site independence:
- Phylogenetic Model is a latent class graphical model
- Vertex $v \in T$ gives a random variable $X_v \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$
- All random variables corresponding to internal nodes are latent



$$p_{i_1 i_2 i_3} = \sum_{j_1} \sum_{j_2} \pi_{j_1} a_{j_2, j_1} b_{i_1, j_1} c_{i_2, j_2} d_{i_3, j_2}$$

Phylogenetic Mixture Models

- Basic phylogenetic model assume homogeneity across sites
- This assumption is not accurate within a single gene
  - Some sites more important: unlikely to change
- Tree structure may vary across genes



- Leads to mixture models for different classes of sites
- $\mathcal{M}(T, r)$ denotes a same tree mixture model with underlying tree T and r classes of sites

Identifiability: Numerical Parameters

Definition

A parametric statistical model is a function that associates a probability distribution to a parameter vector. The model is identifiable if the function is 1-to-1.

- Two types of parameters which we treat separately:
  - Numerical parameters (conditional distributions $f(x_v | x_{\mathrm{pa}(v)})$)
  - Tree parameter (combinatorial types of trees relating species)

Definition

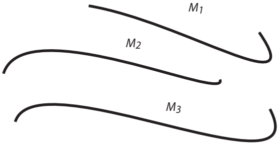
Fix a tree T. The numerical parameters of an r-class same tree phylogenetic mixture model are identifiable if the resulting polynomial map from numerical parameters to probability distributions is 1-to-1.

Identifiability: Tree Parameters

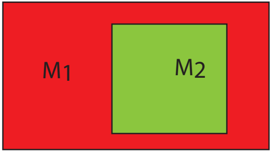
Definition

The tree parameters in an r class same tree phylogenetic mixture model are identifiable if for all n leaf trees $T_1 \neq T_2$,

$$\mathcal{M}(T_1, r) \cap \mathcal{M}(T_2, r) = \emptyset.$$



Identifiable



Not Identifiable

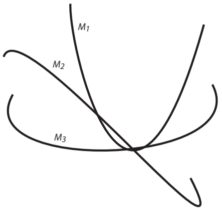
Generic Identifiability

- Identifiability is too strong a condition for mixture models
  - Numerical parameters not identifiable
  - Tree parameters not identifiable

Definition

- Numerical parameters are generically identifiable if there is a dense Zariski open subset of parameter space where identifiable.
- Tree parameters generically identifiable if for all T_1, T_2

$$\dim(\mathcal{M}(T_1, r) \cap \mathcal{M}(T_2, r)) < \min(\dim(\mathcal{M}(T_1, r)), \dim(\mathcal{M}(T_2, r))).$$



Identifiability Questions for Mixture Models

Question

For fixed number of trees r, are the tree parameters $T_1, \ldots, T_r$, and rate parameters of each tree (generically) identified in phylogenetic mixture models?

- $r = 1$ (Ordinary phylogenetic models)
  Most models are identifiable on $\geq 2, 3, 4$ leaves. (Rogers, Chang, Steel, Hendy, Penny, Székely, Allman, Rhodes, Housworth, ...)
- $k > 1$ $T_1 = T_2 = \cdots = T_r$ but no restriction on number of trees
  Not identifiable (Matsen-Steel, Stefankovic-Vigoda)
- $r > 1$, T_i arbitrary
  Not identifiable (Mossel-Vigoda)

Theorem (Rhodes-S 2010)

The tree and numerical parameters in a r-class, same tree phylogenetic mixture model on n-leaf trivalent trees are generically identifiable, if $r < 4^{\lceil n/4 \rceil}$.

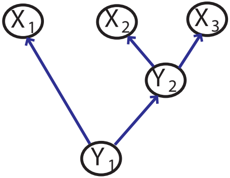
Proof Ideas.

- Phylogenetic invariants from flattenings
- Tensor rank (Kruskal's Theorem)
- Elementary tree combinatorics
- Solving tree and numerical parameter identifiability at the same time

□

Phylogenetics and Algebraic Geometry

- If we fix a tree T, get a rational map $\phi_T : \mathbb{R}^d \to \mathbb{R}^{4^n}$.



$$\phi_{i_1 i_2 i_3}(\pi, a, b, c, d) =$$

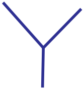
$$\sum_{j_1} \sum_{j_2} \pi_{j_1} a_{j_2, j_1} b_{i_1, j_1} c_{i_2, j_2} d_{i_3, j_2}$$

- $\Theta \subseteq \mathbb{R}^d$ as set of biologically meaningful parameters.
- $\mathcal{M}(T, 1) = \phi_T(\Theta)$ is the phylogenetic model.
- $\overline{\mathcal{M}(T, 1)}$ (Zariski closure) in the phylogenetic variety.
- r-class mixture $\overline{\mathcal{M}(T, r)}$ is the rth secant variety of $\overline{\mathcal{M}(T, 1)}$

Definition

The phylogenetic invariants of the model $\mathcal{M}(T, r)$ and the polynomials in the ideal:

$$I(T, r) = \mathcal{I}(\mathcal{M}(T, r)) \subseteq \mathbb{C}[p_{i_1 \cdots i_n} : i_j \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}].$$



$$p_{i_1 i_2 i_3} = \pi_A a_{i_1,A} b_{i_2,A} c_{i_3,A} + \pi_C a_{i_1,C} b_{i_2,C} c_{i_3,C} + \pi_G a_{i_1,G} b_{i_2,G} c_{i_3,G} + \pi_T a_{i_1,T} b_{i_2,T} c_{i_3,T}$$

$$V_T = \mathrm{Sec}^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$$

- Determining phylogenetic invariants is a hard problem.

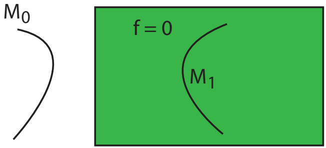
Proving Identifiability with Algebraic Geometry

Proposition

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

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

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

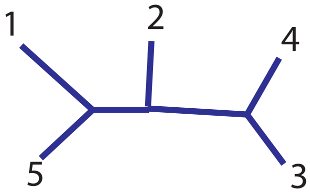


Splits and Tripartitions in a Tree

Definition

Let T be a tree with leave label set $\{1, 2, \ldots, n\}$.

- A partition $A_1|A_2|\cdots|A_t$ of the leaves is convex for T if $T|_{A_i} \cap T|_{A_j} = \emptyset$ for all $i \neq j$.
- Bipartitions $A_1|A_2$ of the leaves are called splits.
- A triparition $A|B|C$ is vertex induced if it obtained by removing a vertex in T.



- Convex: 15|234, 2|15|34
- Not Convex: 12|345
- Vertex Induced: 2|15|34
- Not Vertex Induced: 15|24|3

2-way Flattenings and Matrix Ranks

$$p_{ijkl} = \mathrm{P}(X_1 = i, X_2 = j, X_3 = k, X_4 = l)$$

$$\mathrm{Flat}_{12|34}(P) = \begin{pmatrix} p_{AAAA} & p_{AAAC} & p_{AAAG} & \cdots & p_{AATT} \\ p_{ACAA} & p_{ACAC} & p_{ACAG} & \cdots & p_{ACTT} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{TTAA} & p_{TTAC} & p_{TTAG} & \cdots & p_{TTTT} \end{pmatrix}$$

Proposition

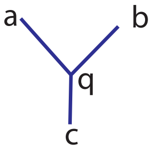
Let $P \in \mathcal{M}(T, r)$.

- *If $A|B$ is a convex split for T, then $\mathrm{rank}(\mathrm{Flat}_{A|B}(P)) \leq 4r$.*
- *If $C|D$ is not a nonconvex split for T, then generically $\mathrm{rank}(\mathrm{Flat}_{C|D}(P)) \geq \min(4r+1, 4^{\#A}, 4^{\#B})$.*

3-way Tensors and Kruskal's Theorem

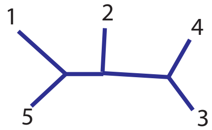
Theorem (Kruskal 1976)

Consider the generalized tree model $\mathcal{M}(a, b, c; q)$. This model is generically identifiable provided
$\min(a, q) + \min(b, q) + \min(c, q) \geq 2q + 2.$

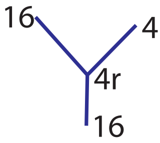


Proposition

Suppose $A|B|C$ is a vertex induced tripartition for T. Then $\mathcal{M}(T, r) \subseteq \mathcal{M}(4^{\#A}, 4^{\#B}, 4^{\#C}; 4r)$ and intersects the identifiable locus.



15|2|34



Putting It Together

Lemma

Every trivalent tree T with n leaves has a vertex induced tripartition $A|B|C$ with $\#A \geq \#B \geq \lceil n/4 \rceil$.

1. Use flattening rank invariants to find the tripartition from Lemma.
2. Use Kruskal's Theorem to recover numerical parameters in model $\mathcal{M}(T, r) \subseteq \mathcal{M}(4^{\#A}, 4^{\#B}, 4^{\#C}; 4r)$.
3. Use phylogenetic invariants to test for trees on each induced subtree on $T|_A$, $T|_B$, $T|_C$ and "untangle" slices.
4. Use results on identifiability of ordinary tree models to get numerical parameters for $T|_A$, $T|_B$, $T|_C$, and hence for T.

Further Results and the Future

- Same techniques yield results for different tree mixtures (joins) when all trees $T_1, \ldots, T_r$ have a common pair of deep splits

$$A|B \cup C \quad \text{and} \quad B|A \cup C.$$

- Generalizing to tree mixtures with no common structure requires studying new tensor decomposition.

Problem

Let $V^r_{12|34} * V^r_{13|24}$ be

$$\{P \in \mathbb{C}^r \otimes \mathbb{C}^r \otimes \mathbb{C}^r \otimes \mathbb{C}^r :$$

$$P = Q + R \text{ where } \mathrm{rank}\,(\mathrm{Flat}_{12|34}(Q)) \leq r, \mathrm{rank}\,(\mathrm{Flat}_{13|24}(R)) \leq r\}.$$

Determine phylogenetically relevant equations in $\mathcal{I}(V^r_{12|34} * V^r_{13|24})$.

Is it possible to drop "generic"?

Theorem (Allman-Rhodes-S 2012)

Let $T \neq T'$ be trivalent trees on n nodes. Then

$$\mathcal{M}(T', 1) \cap \mathcal{M}(T, 3) = \emptyset.$$

- Exploits the fact that we are not interested in general transition matrices in our underlying graphical model.
- All transition matrices of form $A = \exp(Qt)$ where Q is a "rate" matrix.
- This forces all variables to be positively correlated.
- Uses flattening invariants from convex splits.
- Might this "positive correlation" approach be useful for other graphical models?

Summary and Acknowledgments

- For practical purposes, same tree mixture models are identifiable
- Best available results require algebraic geometry
- Algebraic and tensor-based methods can likely be used for identifiability problems on other latent variable graphical models
- New algebraic results are needed for more general mixture models

- Acknowledgments
  - National Science Foundation
  - David and Lucille Packard Foundation

References

E. Allman, C. Matias, J. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, **37** no.6A (2009) 3099-3132.

E. Allman, J. Rhodes, S. Sullivant. When do phylogenetic mixture models mimic other phylogenetic models? `1202.2396`

F.A. Matsen and M. Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Systematic Biology*, 2007.

E. Mossel and E. Vigoda Phylogenetic MCMC Are Misleading on Mixtures of Trees. *Science* **309**, 2207–2209 (2005)

J. Rhodes, S. Sullivant. Identifiability of large phylogenetic mixture models. To appear *Bulletin of Mathematical Biology*, 2011. `1011.4134`

D. Stefankovic and E. Vigoda. Pitfalls of Heterogeneous Processes for Phylogenetic Reconstruction *Systematic Biology* **56**(1): 113-124, 2007.