Phylogenetic Invariants: Recent Progress and New Directions

Elizabeth S. Allman*
Department of Mathematics and Statistics
University of Southern Maine

Institut Henri Poincaré
Mathematics of Evolution and Phylogeny
June 18, 2005

* and other authors

All work joint with

John A. Rhodes
Department of Mathematics
Bates College

Introduction

In 1987,

Cavender and Felsenstein (JC)

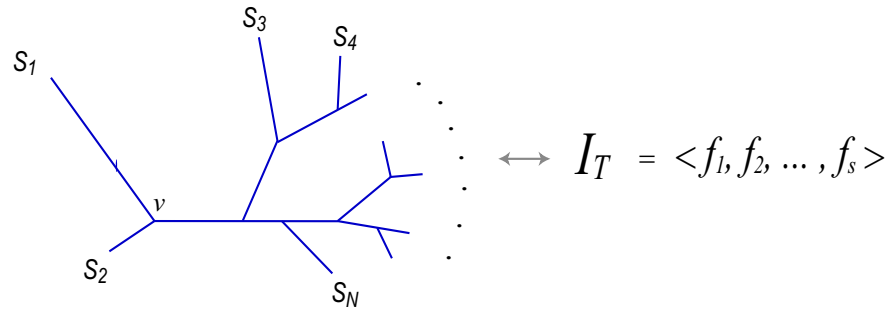
Lake ('K2P')

proposed a new method of inference of trees using

phylogenetic invariants

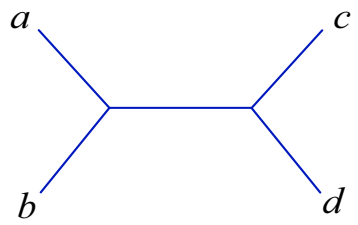
Phylogenetic invariants are polynomials $f_1, f_2, f_3, \ldots$

- associated to a model of sequence mutation M (JC, K2P, GM, ...) and topological tree T relating n taxa
- with the property that if T and M are the correct tree and model describing the evolutionary process, then $f_1, \cdots, f_s$ vanish when evaluated at (perfect) sequence data.



Phylogenetic invariants are polynomials in variables $p_{i_1 i_2 \ldots i_n}$ that represent expected pattern frequencies,

4-taxon case:



$$P(i,j,k,l) = p_{ijkl},$$

is the expected frequency of observing

$$i \text{ at } a, \quad j \text{ at } b, \quad k \text{ at } c, \quad l \text{ at } d.$$

P is the $4 \times 4 \times 4 \times 4$ joint distribution tensor describing probabilities of bases at leaves

(tensor = multi-dimensional array = table)

Viewing P as a tensor of variables:

A phylogenetic invariant $f(P)$ is a polynomial in 4^n variables.

Replacing P with a joint distribution tensor P_0 arising from any specific parameters for T, M:

$$f(P_0) = 0$$

Idea was to evaluate invariants at observed pattern frequencies in aligned sequences (data):

a: ATTAGGTACATGATTAG

b: ATTCGGTACATGATTAG

c: ATTCGCTACATGATCCG

d: ATTTGCTACATGTTCCG

$$\widehat{p}_{AAAA} = 3/17,\ \widehat{p}_{ACCT} = 1/17, ...$$

If T, M, are the correct tree and mutation model relating the sequences, then $\widehat{P} = (\widehat{p}_{ijkl})$ and $\widehat{P} \approx P_0$.

Since $f(P_0) = 0$, then $f(\widehat{P}) \approx 0$

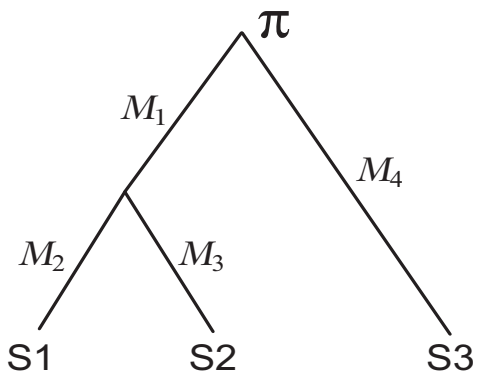
Near vanishing of f on $\widehat{P}$

$\updownarrow$

support for T, M as correct tree, model

From a more algebraic/geometric viewpoint...

Ex: π = root distribution vector, $\{M_e\}$ Markov transition matrices on edges



Given T, $\boldsymbol{\pi}$,$\{M_e\}$, compute joint distribution of bases at leaves:
E.g., $GAA \rightsquigarrow 311$,

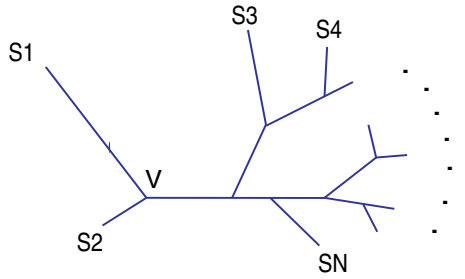
$$p_{311} = \sum_{i=1}^{4}\sum_{j=1}^{4} \pi_i M_1(i,j) M_2(j,3) M_3(j,1) M_4(i,1)$$

$P = (p_{ijk})$ is a $4 \times 4 \times 4$ tensor.

each p_{ijk} is polynomial in unknown parameters.

More generally....

Fix an n-taxon tree T, κ states at each node, GM model of mutation



$\kappa = 4$ (DNA), $\kappa = 2$ (R/Y), or $\kappa = 20$ (proteins)

$$\text{parameters} = \begin{cases} \text{tree} \\ \text{root distribution vector} \\ \text{Markov matrix on each edge} \end{cases}$$

(more restrictive models given by additional assumptions)

Stochastic Parameter Space: $S \subset [0,1]^N,\ \ s \in S,\ s = (\boldsymbol{\pi}, \{M_e\})$

$$N = (\kappa - 1) + (2n - 3)\kappa(\kappa - 1)$$

Joint distribution space: $[0,1]^{\kappa^n}$

Polynomial parameterization map:

$$\phi_T : S \longrightarrow [0,1]^{\kappa^n}$$

$$\phi_T(s) = P$$

P, an n-dimensional $\kappa \times \cdots \times \kappa$ tensor giving the *joint distribution of pattern frequencies* at the leaves of the n-taxon tree T.

Since ϕ_T is polynomial, extend to a polynomial map

$$\phi_T : \mathbb{C}^N \longrightarrow \mathbb{C}^{\kappa^n}$$

We can use algebraic geometry to understand the image, the *phylogenetic variety*,

$$V_T = \overline{\phi_T(\mathbb{C}^N)}.$$

Since $\kappa^n >> N$, points in the image of ϕ_T ('pattern frequencies') will satisfy polynomial relations.

These equations are the *phylogenetic invariants* for (T, GM),

Finding invariants $\leftrightsquigarrow$ finding an implicit description of V_T, as a zero set of polynomials

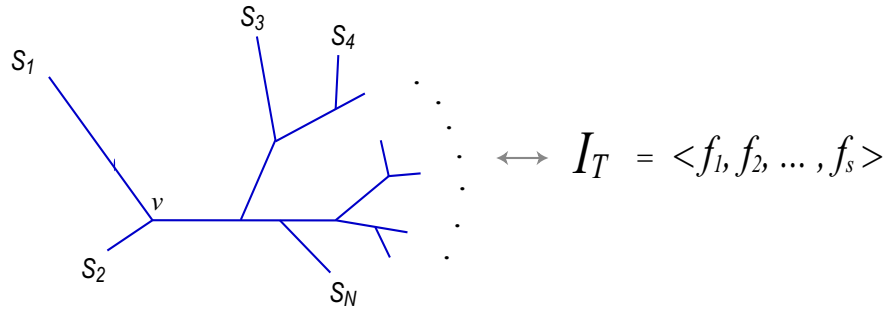
$$\phi_T : \mathbb{C}^N \longrightarrow V_T \subseteq \mathbb{C}^{\kappa^n}$$

i.e. finding the kernel of the associated ring map

$$\Phi_T : \mathbb{C}[p_{0\ldots0}, \cdots, p_{\kappa\ldots\kappa}] \longrightarrow \mathbb{C}[s_1, \cdots, s_N]$$

$$\ker \Phi_T = I_T \equiv \textit{ phylogenetic ideal,}$$

the ideal of polynomials in $p_{0\ldots0}, \ldots, p_{\kappa\ldots\kappa}$ vanishing for all choices of (stochastic or complex) parameters.



Notions from algebraic geometry:

S set-theoretically defines V: A variety V is the zero set of a collection S of polynomials, $V = Z(S)$. (S need not be an ideal.)

I ideal-theoretically defines V: A variety V corresponds to an ideal I, $I = Id(V)$, where I contains *all polynomials* vanishing on V. (I is a radical ideal.)

Generally,

$$S \subsetneq Id(Z(S))$$

Ex. In $\mathbb{R}^2$,

$$S = \{(y-x)^2\} \subsetneq Id(Z((y-x)^2)) = (y-x).$$

Most earlier work on phylogenetic invariants focused on set-theoretic notion, or even weaker idea of finding *some* polynomials vanishing on V.

Ideal theoretic results are generally much more difficult to obtain.

(Revisited) Idea is to evaluate invariants at pattern frequencies in aligned sequences (data):

a: ATTAGGTACATGATTAG

b: ATTCGGTACATGATTAG

c: ATTCGCTACATGATCCG

d: ATTTGCTACATGTTCCG

$$\widehat{p}_{AAAA} = 3/17,\ \widehat{p}_{ACCT} = 1/17, ...$$

If T, GM, are the correct tree and mutation model relating the sequences, then $\widehat{P} \approx P_0 = \phi(s) \in V_T$, for some parameters s.

For $f \in I_T$, $f(P_0) = 0$, so $f(\widehat{P}) \approx 0$

'Plan' of Implementation: Given data $\widehat{P}$,

- Fix M.
- For each T,
  - Find some/most/all invariants f for V_T (set- vs. ideal theoretic)
  - Test if $f(\widehat{P}) \approx 0$.
- Return tree for which $\widehat{P}$ " $\in$ " V_T (as best possible)

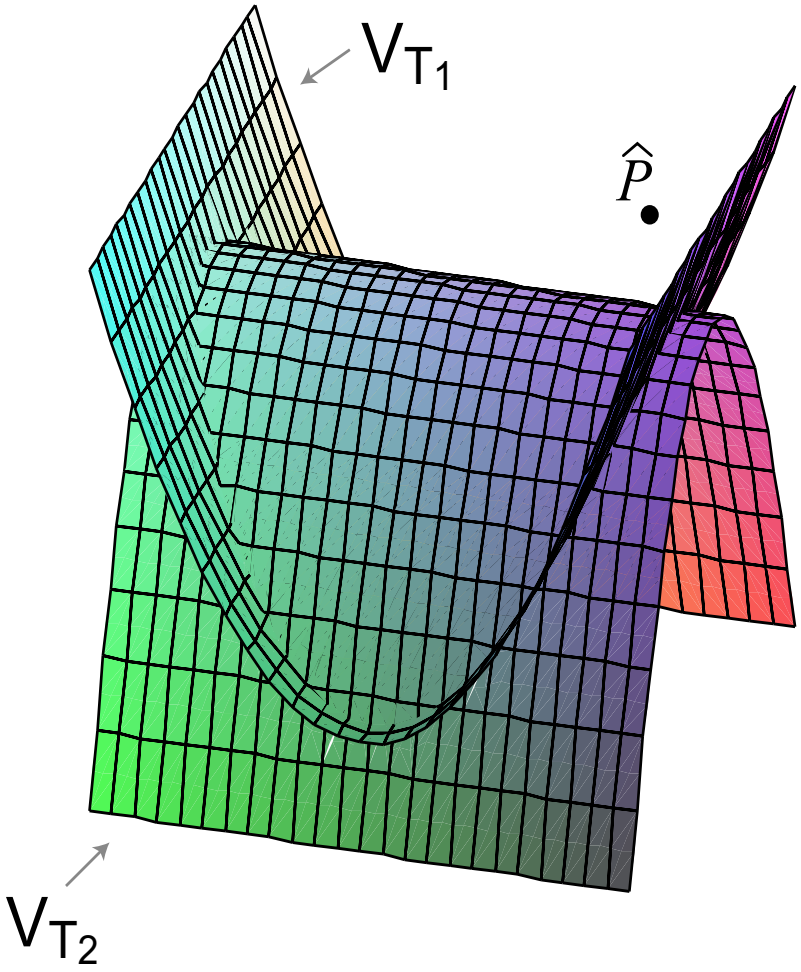
This method is statistically consistent.

More generators of I_T known $\rightsquigarrow$ improved tree inference.

Issues:

Invariants will not be identically zero, only close to zero

- statistical issues (finite length sequences, imperfect model)
- algebraic issues (evaluation at points off V_T,
  precise form affects "near" vanishing)


V_{T_1}
$\hat{P}$
V_{T_2}

Finding invariants:

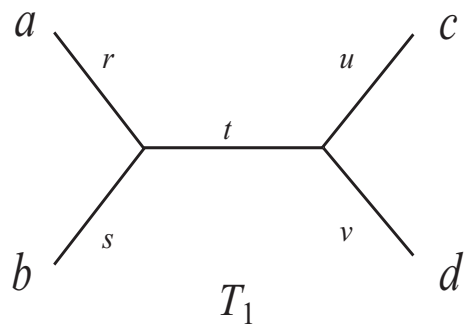
Only one invariant is easy to see – stochastic invariant

$$1 - \sum_{ijkl} p_{ijkl}$$

Other invariants are typically higher degree and reflect the topology of the tree T and choice of mutation model.

Cavender and Felsenstein

- Symmetric 2-state model



The stochastic invariant: $p_{0000} + \cdots + p_{1111} - 1.$

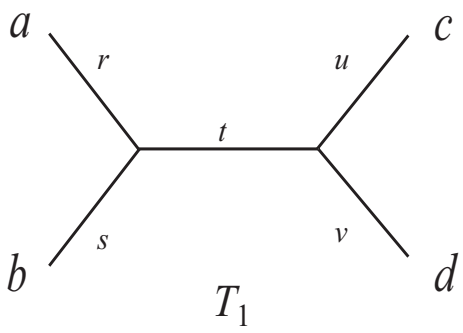
The symmetry invariants: $p_{0000} - p_{1111},$ etc.

An informative invariant:

$$f_1 = (p_{0100} + p_{1011} - p_{0111} - p_{1000})(p_{0010} + p_{1101} - p_{0001} - p_{1110}) - (p_{0110} + p_{1001} - p_{0101} - p_{1010})(p_{0000} + p_{1111} - p_{0011} - p_{1100})$$

$$P \in V_{T_1} \longleftrightarrow f_1(P) = 0$$

T_1 (and JC) correct.



Origin of C–F informative invariant is

The 4-point condition for tree metrics

$$d_{ab} + d_{cd} \leq d_{ac} + d_{bd} = d_{ad} + d_{bc}$$

using log-det distance

$$f_1(P) = 0 \longleftrightarrow e^{d_{ad}} e^{d_{bc}} - e^{d_{ac}} e^{d_{bd}} = 0$$

Work on determining invariants

Introduction of Invariants:

- Cavender and Felsenstein
- Lake

Hadamard Conjugation/Fourier transform methods: (Group-based models: Kimura 3ST model and submodels)

- Hendy, Hendy and Penny, ...
- Evans and Speed/Evans and Zhou
- Steel, Székely, Erdös, Waddell

Numerical approach (K2P 4-taxon, limited rate variation, ...)

Ferretti and Sankoff

Computational Algebra approach using Gröbner bases

Hagedorn

Special case: Linear invariants

Many authors...

Linear invariants exist only for certain models.

Their study was motivated by insensitivity to rate variation.

Simulation study: How well do invariants work for inference?

Huelsenbeck

Lake's linear invariants inefficient in practice.

Very long sequences needed for good performance.

More recent progress...

Group-based models (Sturmfels – Sullivant, 2004):

complete description of the full phylogenetic ideal for a group-based model, arbitrary tree

Builds on previous work,

Hadamard transform = change of variables

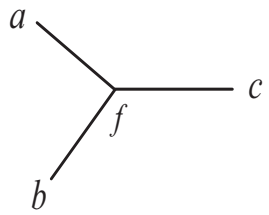
This gives a simpler form to parameterization map ϕ_T

polynomials $\rightsquigarrow$ monomials

Varieties parameterized by monomials are called toric varieties, and are a well-understood class in algebraic geometry.

General Markov model (AR, 2003):

Construction of many invariants for arbitrary trees... but most important is 3-taxon



- Look at 'slices' and marginalizations of $\kappa \times \kappa \times \kappa$ tensor P.
- Express these in terms of parameters through matrix algebra
- Construct matrix products from these that must be simultaneously diagonalizable
- Invariants arise from commutation of simultaneously diagonalizable matrices

Invariants from commutation relations

- have degree $\kappa + 1$ (lowest possible),
- have many terms (≈ 180 for $\kappa = 4$)
- do not generate full ideal, (up to explicit saturation, radical)
- for $\kappa = 4$ gives all degree 5 invariants (1728-d space)

Note: Expressed in matrix form, invariants may be better evaluated through numerical linear algebra (?)

An example ($\kappa = 4$):

$$
\begin{aligned}
f \;=\; & -p_{121}p_{133}p_{002}p_{212}p_{322} + p_{121}p_{133}p_{002}p_{222}p_{312} + p_{121}p_{133}p_{202}p_{012}p_{322} \\
& - p_{121}p_{133}p_{202}p_{022}p_{312} - p_{121}p_{133}p_{302}p_{012}p_{222} + p_{121}p_{133}p_{302}p_{022}p_{212} \\
& + p_{321}p_{103}p_{012}p_{122}p_{232} - p_{321}p_{103}p_{012}p_{132}p_{222} - p_{321}p_{103}p_{112}p_{022}p_{232} \\
& + p_{321}p_{103}p_{112}p_{032}p_{222} + p_{321}p_{103}p_{212}p_{022}p_{132} - p_{321}p_{103}p_{212}p_{032}p_{122} \\
& - p_{321}p_{113}p_{002}p_{122}p_{232} + p_{321}p_{113}p_{002}p_{132}p_{222} + p_{321}p_{113}p_{102}p_{022}p_{232} \\
& - p_{321}p_{113}p_{102}p_{032}p_{222} - p_{321}p_{113}p_{202}p_{022}p_{132} + p_{321}p_{113}p_{202}p_{032}p_{122} \\
& + p_{321}p_{123}p_{002}p_{112}p_{232} - p_{321}p_{123}p_{002}p_{132}p_{212} - p_{321}p_{123}p_{102}p_{012}p_{232} \\
& + p_{321}p_{123}p_{102}p_{032}p_{212} + p_{321}p_{123}p_{202}p_{012}p_{132} - p_{321}p_{123}p_{202}p_{032}p_{112} \\
& - p_{321}p_{133}p_{002}p_{112}p_{222} + p_{321}p_{133}p_{002}p_{122}p_{212} + p_{321}p_{133}p_{102}p_{012}p_{222} \\
& - p_{321}p_{133}p_{102}p_{022}p_{212} - p_{321}p_{133}p_{202}p_{012}p_{122} + p_{321}p_{133}p_{202}p_{022}p_{112} \\
& - p_{323}p_{101}p_{212}p_{022}p_{132} + p_{323}p_{101}p_{212}p_{032}p_{122} + p_{323}p_{111}p_{002}p_{122}p_{232} \\
& - p_{323}p_{111}p_{002}p_{132}p_{222} - p_{323}p_{111}p_{102}p_{022}p_{232} + p_{323}p_{111}p_{102}p_{032}p_{222} \\
& + p_{323}p_{111}p_{202}p_{022}p_{132} - p_{323}p_{111}p_{202}p_{032}p_{122} - p_{323}p_{121}p_{002}p_{112}p_{232} \\
& + p_{323}p_{121}p_{002}p_{132}p_{212} + p_{323}p_{121}p_{102}p_{012}p_{232} - p_{323}p_{121}p_{102}p_{032}p_{212} \\
& - p_{323}p_{121}p_{202}p_{012}p_{132} + p_{323}p_{121}p_{202}p_{032}p_{112} + p_{323}p_{131}p_{002}p_{112}p_{222} \\
& - p_{323}p_{131}p_{002}p_{122}p_{212} - p_{323}p_{131}p_{102}p_{012}p_{222} + p_{323}p_{131}p_{102}p_{022}p_{212} \\
& + p_{323}p_{131}p_{202}p_{012}p_{122} - p_{323}p_{131}p_{202}p_{022}p_{112} - p_{223}p_{111}p_{302}p_{022}p_{132} \\
& + p_{223}p_{111}p_{302}p_{032}p_{122} - p_{121}p_{103}p_{012}p_{232}p_{322} - p_{221}p_{103}p_{012}p_{122}p_{332} \\
& + p_{221}p_{103}p_{012}p_{132}p_{322} + p_{221}p_{103}p_{112}p_{022}p_{332} - p_{221}p_{103}p_{112}p_{032}p_{322} \\
& - p_{221}p_{103}p_{312}p_{022}p_{132} + p_{221}p_{103}p_{312}p_{032}p_{122} + p_{221}p_{113}p_{002}p_{122}p_{332} \\
& - p_{221}p_{113}p_{002}p_{132}p_{322} - p_{221}p_{113}p_{102}p_{022}p_{332} + p_{221}p_{113}p_{102}p_{032}p_{322} \\
& + p_{221}p_{113}p_{302}p_{022}p_{132} - p_{221}p_{113}p_{302}p_{032}p_{122} - p_{221}p_{123}p_{002}p_{112}p_{332} \\
& + p_{221}p_{123}p_{002}p_{132}p_{312} + p_{221}p_{123}p_{102}p_{012}p_{332} - p_{221}p_{123}p_{102}p_{032}p_{312} \\
& - p_{221}p_{123}p_{302}p_{012}p_{132} + p_{221}p_{123}p_{302}p_{032}p_{112} + p_{221}p_{133}p_{002}p_{112}p_{322} \\
& - p_{221}p_{133}p_{002}p_{122}p_{312} - p_{221}p_{133}p_{102}p_{012}p_{322} + p_{221}p_{133}p_{102}p_{022}p_{312} \\
& + p_{221}p_{133}p_{302}p_{012}p_{122} - p_{221}p_{133}p_{302}p_{022}p_{112} - p_{223}p_{101}p_{012}p_{132}p_{322} \\
& - p_{223}p_{101}p_{112}p_{022}p_{332} + p_{121}p_{103}p_{212}p_{032}p_{322} + p_{121}p_{103}p_{312}p_{022}p_{232} \\
& - p_{123}p_{101}p_{012}p_{222}p_{332} + p_{123}p_{101}p_{012}p_{232}p_{322} + p_{123}p_{101}p_{212}p_{022}p_{332} \\
& - p_{123}p_{101}p_{212}p_{032}p_{322} - p_{123}p_{101}p_{312}p_{022}p_{232} + p_{123}p_{101}p_{312}p_{032}p_{222} \\
& + p_{123}p_{111}p_{002}p_{222}p_{332} - p_{123}p_{111}p_{002}p_{232}p_{322} - p_{123}p_{111}p_{202}p_{022}p_{332}
\end{aligned}
$$

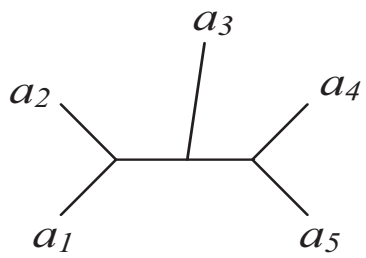
$$
\begin{aligned}
&+p_{123}p_{111}p_{202}p_{032}p_{322} + p_{123}p_{111}p_{302}p_{022}p_{232} - p_{123}p_{111}p_{302}p_{032}p_{222}\\
&+p_{123}p_{131}p_{002}p_{212}p_{322} - p_{123}p_{131}p_{002}p_{222}p_{312} - p_{123}p_{131}p_{202}p_{012}p_{322}\\
&+p_{123}p_{131}p_{202}p_{022}p_{312} + p_{123}p_{131}p_{302}p_{012}p_{222} - p_{123}p_{131}p_{302}p_{022}p_{212}\\
&-p_{021}p_{103}p_{112}p_{222}p_{332} + p_{021}p_{103}p_{112}p_{232}p_{322} + p_{021}p_{103}p_{212}p_{122}p_{332}\\
&-p_{021}p_{103}p_{212}p_{132}p_{322} - p_{021}p_{103}p_{312}p_{122}p_{232} + p_{021}p_{103}p_{312}p_{132}p_{222}\\
&+p_{021}p_{113}p_{102}p_{222}p_{332} - p_{021}p_{113}p_{102}p_{232}p_{322} - p_{021}p_{113}p_{202}p_{122}p_{332}\\
&+p_{021}p_{113}p_{202}p_{132}p_{322} + p_{021}p_{113}p_{302}p_{122}p_{232} - p_{021}p_{113}p_{302}p_{132}p_{222}\\
&-p_{021}p_{123}p_{102}p_{212}p_{332} + p_{021}p_{123}p_{102}p_{232}p_{312} + p_{021}p_{123}p_{202}p_{112}p_{332}\\
&-p_{021}p_{123}p_{202}p_{132}p_{312} + p_{023}p_{121}p_{202}p_{132}p_{312} + p_{023}p_{121}p_{302}p_{112}p_{232}\\
&+p_{223}p_{101}p_{012}p_{122}p_{332} + p_{223}p_{101}p_{112}p_{032}p_{322} + p_{223}p_{101}p_{312}p_{022}p_{132}\\
&-p_{223}p_{101}p_{312}p_{032}p_{122} - p_{223}p_{111}p_{002}p_{122}p_{332} + p_{223}p_{111}p_{002}p_{132}p_{322}\\
&+p_{223}p_{111}p_{102}p_{022}p_{332} - p_{223}p_{111}p_{102}p_{032}p_{322} + p_{023}p_{101}p_{112}p_{222}p_{332}\\
&-p_{023}p_{101}p_{112}p_{232}p_{322} - p_{023}p_{101}p_{212}p_{122}p_{332} + p_{023}p_{101}p_{212}p_{132}p_{322}\\
&+p_{023}p_{101}p_{312}p_{122}p_{232} - p_{023}p_{101}p_{312}p_{132}p_{222} - p_{023}p_{111}p_{102}p_{222}p_{332}\\
&+p_{023}p_{111}p_{102}p_{232}p_{322} + p_{023}p_{111}p_{202}p_{122}p_{332} - p_{023}p_{111}p_{202}p_{132}p_{322}\\
&-p_{023}p_{111}p_{302}p_{122}p_{232} + p_{023}p_{111}p_{302}p_{132}p_{222} + p_{023}p_{121}p_{102}p_{212}p_{332}\\
&-p_{023}p_{121}p_{102}p_{232}p_{312} - p_{023}p_{121}p_{202}p_{112}p_{332} - p_{021}p_{123}p_{302}p_{112}p_{232}\\
&+p_{021}p_{123}p_{302}p_{132}p_{212} + p_{021}p_{133}p_{102}p_{212}p_{322} - p_{021}p_{133}p_{102}p_{222}p_{312}\\
&-p_{021}p_{133}p_{202}p_{112}p_{322} + p_{021}p_{133}p_{202}p_{122}p_{312} + p_{021}p_{133}p_{302}p_{112}p_{222}\\
&-p_{021}p_{133}p_{302}p_{122}p_{212} - p_{023}p_{121}p_{302}p_{132}p_{212} - p_{023}p_{131}p_{102}p_{212}p_{322}\\
&+p_{023}p_{131}p_{102}p_{222}p_{312} + p_{023}p_{131}p_{202}p_{112}p_{322} - p_{023}p_{131}p_{202}p_{122}p_{312}\\
&-p_{023}p_{131}p_{302}p_{112}p_{222} + p_{023}p_{131}p_{302}p_{122}p_{212} + p_{223}p_{121}p_{002}p_{112}p_{332}\\
&-p_{223}p_{121}p_{002}p_{132}p_{312} - p_{223}p_{121}p_{102}p_{012}p_{332} + p_{223}p_{121}p_{102}p_{032}p_{312}\\
&+p_{223}p_{121}p_{302}p_{012}p_{132} - p_{223}p_{121}p_{302}p_{032}p_{112} - p_{223}p_{131}p_{002}p_{112}p_{322}\\
&+p_{223}p_{131}p_{002}p_{122}p_{312} + p_{223}p_{131}p_{102}p_{012}p_{322} - p_{223}p_{131}p_{102}p_{022}p_{312}\\
&-p_{223}p_{131}p_{302}p_{012}p_{122} + p_{223}p_{131}p_{302}p_{022}p_{112} - p_{323}p_{101}p_{012}p_{122}p_{232}\\
&+p_{323}p_{101}p_{012}p_{132}p_{222} + p_{323}p_{101}p_{112}p_{022}p_{232} - p_{323}p_{101}p_{112}p_{032}p_{222}\\
&+p_{121}p_{103}p_{012}p_{222}p_{332} - p_{121}p_{103}p_{212}p_{022}p_{332} - p_{121}p_{103}p_{312}p_{032}p_{222}\\
&-p_{121}p_{113}p_{002}p_{222}p_{332} + p_{121}p_{113}p_{002}p_{232}p_{322} + p_{121}p_{113}p_{202}p_{022}p_{332}\\
&-p_{121}p_{113}p_{202}p_{032}p_{322} - p_{121}p_{113}p_{302}p_{022}p_{232} + p_{121}p_{113}p_{302}p_{032}p_{222}
\end{aligned}
$$

Towards a complete understanding of Phylogenetic Invariants for the general Markov model...

Informal interpretation of theorems/conjectures to follow: The phylogenetic ideal I_T for any tree T with GM is determined by the *local topological structure* of T.

Biological interpretation: Phylogenetic invariants may be useful for measuring *support for* a particular *split* (edge) or *tripartition* (node) in a tree.

Example of Theorem: For GM, $\kappa = 2$, consider the tree



The joint distribution tensor P is $2 \times 2 \times 2 \times 2 \times 2$.

P has two natural *flattenings* according to *splits* in the tree:

$$\{\{a_1, a_2\}, \{a_3, a_4, a_5\}\}, \text{ and } \{\{a_1, a_2, a_3\}, \{a_4, a_5\}\}.$$

The corresponding *flattenings* are

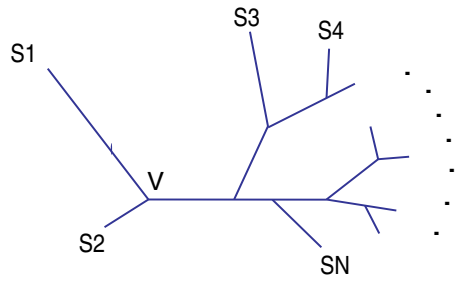
$$\begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} & p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} & p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} & p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} & p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}$$

and

$$\begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} \\ p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} \\ p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} \\ p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} \\ p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}.$$

Theorem: For this 5-leaf tree, I_T is generated by all 3×3 minors of these two matrices. (That is, these matrices have rank ≤ 2.)

Theorem: For $\kappa = 2$, any bifurcating T, GM, the phylogenetic ideal I_T is generated by *edge invariants* (3×3 determinants associated to flattenings on edges).



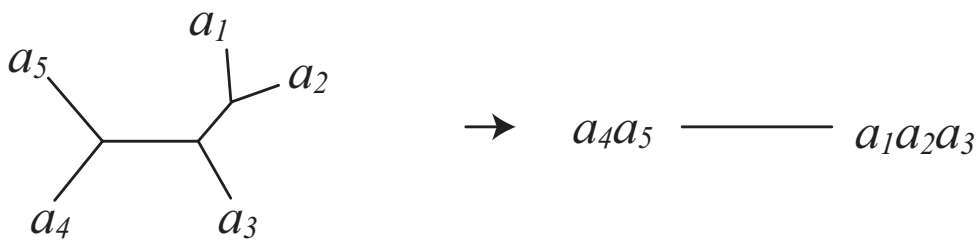
Implications: Via rank of flattenings, we can potentially say something about data's support for a *particular edge* of a phylogenetic tree.

Furthermore,

support for all edges = support for tree

(when $\kappa = 2$).

Indeed, for any κ, we can find invariants from *edge flattenings*:

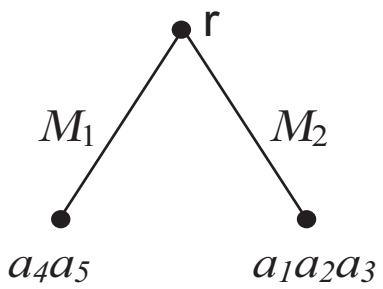


For $P \in V_T$, we can 'flatten' P along an edge e:

$$P \mapsto \mathrm{Flat}_e(P), \quad \text{a } \kappa^{n_1} \times \kappa^{n_2} \text{ matrix},$$

for $n_1 + n_2 = n$.

Coarser model is now



for

M_1, a $\kappa \times \kappa^{n_1}$ matrix

M_2, a $\kappa \times \kappa^{n_2}$ matrix

Thus,

$$\mathrm{Flat}_e(P) = M_1^T \,\mathrm{diag}(\mathbf{p_r}) M_2,$$

and $\mathrm{Flat}_e(P)$ is rank κ matrix (at most).

Therefore,

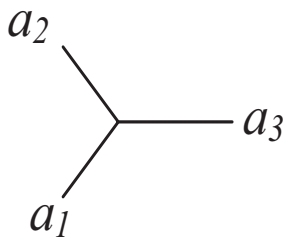
all $(\kappa+1) \times (\kappa+1)$ minors will vanish.

These are the edge invariants for a tree T.

But edge invariants are *not* enough for ideal generation if $\kappa > 2$.

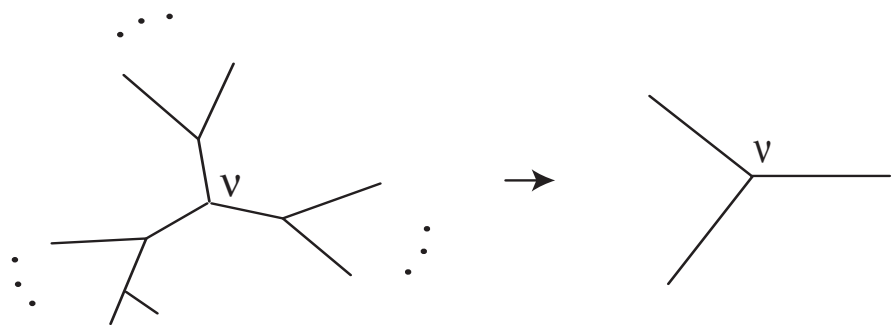
Evidence for this:

Example: $\kappa = 4,\ n = 3,$



Since all the edges of the 3-taxon tree T are terminal, edge invariants do *not* exist, even though there are *many* invariants here.

For an arbitrary tree, focus on a node:



For $P \in V_T$, an n-dimensional tensor, flatten

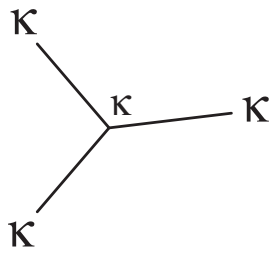
$$P \mapsto \mathrm{Flat}_v(P), \quad \text{a } 4^{n_1} \times 4^{n_2} \times 4^{n_3} \text{ tensor},$$

$$n_1 + n_2 + n_3 = n$$

and determine analogue for edge invariants for a vertex v.

Note: T is bifurcating for simplicity

Focus again on 3-leaf tree, κ-state model, root at center...



Let $V(\kappa;\kappa,\kappa,\kappa) = V_T$ denote the associated variety.

$$p_{ijk} = \sum_l \pi_l M_1(l,i) M_2(l,j) M_3(l,k)$$

But $M_e(l,\cdot) \in \mathbb{P}^{\kappa-1}$, so

$$V(\kappa;\kappa,\kappa,\kappa) \;=\; \mathrm{Sec}^{\kappa}(\mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1})$$

$$\text{“}=\text{”}\ \kappa \times \kappa \times \kappa \text{ tensors of rank} \leq \kappa$$

Defn: A n-dimensional tensor has *rank* 1 if it is the tensor product of n vectors. It has *rank* k if it can be expressed as a sum of k rank 1 tensors, but no fewer.

Viewing the 3-taxon phylogenetic variety as a secant variety:

$$V(\kappa;\kappa,\kappa,\kappa) = \mathrm{Sec}^{\kappa}(\mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1})$$

$$\text{“}=\text{”}\ \kappa \times \kappa \times \kappa \text{ tensors of rank} \leq \kappa$$

makes the problem classical — but doesn't solve it.

Vertex invariants are the analogues of matrix minors for tensors of rank κ.

matrix M of rank κ

$$\leftrightarrow \quad M = \begin{pmatrix} | \\ p_1 \\ | \end{pmatrix} \begin{pmatrix} - & q_1 & - \end{pmatrix} + \cdots + \begin{pmatrix} | \\ p_\kappa \\ | \end{pmatrix} \begin{pmatrix} - & q_\kappa & - \end{pmatrix}$$

$$\leftrightarrow M \in \mathrm{Sec}^\kappa(\mathbb{P}^n \times \mathbb{P}^m)$$

$$\leftrightarrow (\kappa+1) \times (\kappa+1) \text{ minors vanish}$$

3-dim tensor P of rank κ

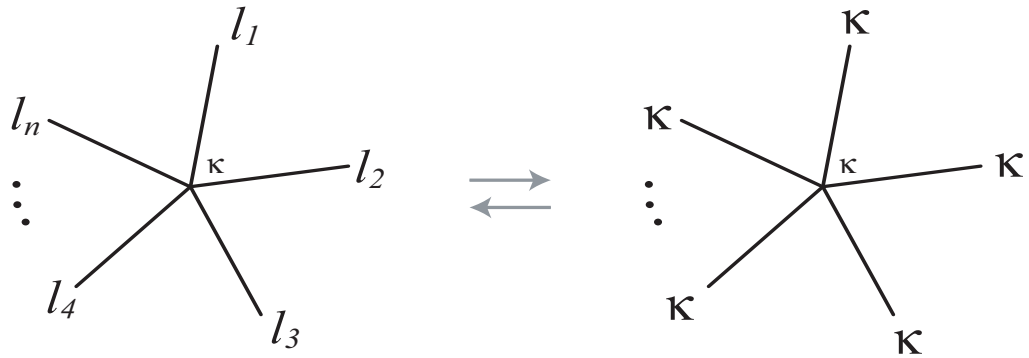
$$\leftrightarrow \quad P = \mathbf{p_1} \otimes \mathbf{q_1} \otimes \mathbf{r_1} + \cdots + \mathbf{p}_\kappa \otimes \mathbf{q}_\kappa \otimes \mathbf{r}_\kappa$$

$$\leftrightarrow P \in \mathrm{Sec}^\kappa(\mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1} \times \mathbb{P}^{\kappa-1})$$

$$\leftrightarrow \text{vertex invariants vanish}$$

Understanding star models

Theorem: If $l_1, l_2, \ldots, l_n \geq \kappa$, then generators of the ideal defining $V(\kappa; l_1, l_2, \ldots, l_n)$ can be explicitly constructed from generators of the ideal defining $V(\kappa; \kappa, \kappa, \ldots, \kappa)$.

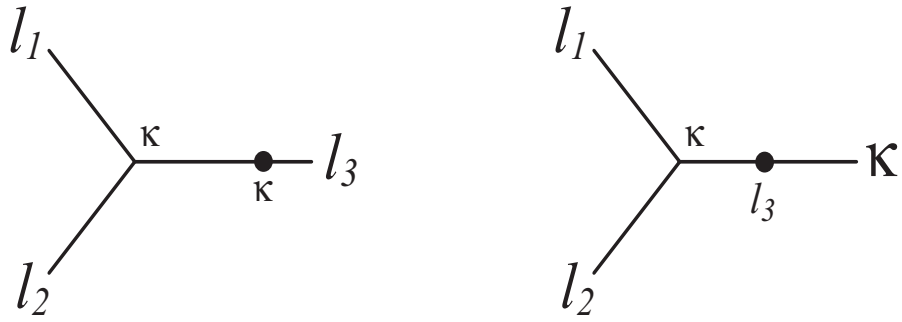


A glimpse of the proof: Observe that

$$V(\kappa; l_1, l_2, \kappa) *_{3,1} M_{\kappa \times l_3} = V(\kappa; l_1, l_2, l_3),$$

$$V(\kappa; l_1, l_2, l_3) *_{3,1} M_{l_3 \times \kappa} = V(\kappa; l_1, l_2, \kappa).$$

Here $M_{m \times n}$ denotes $m \times n$ matrices, and $*_{3,1}$ denotes 'matrix multiplication' in the 3 and 1 indices.

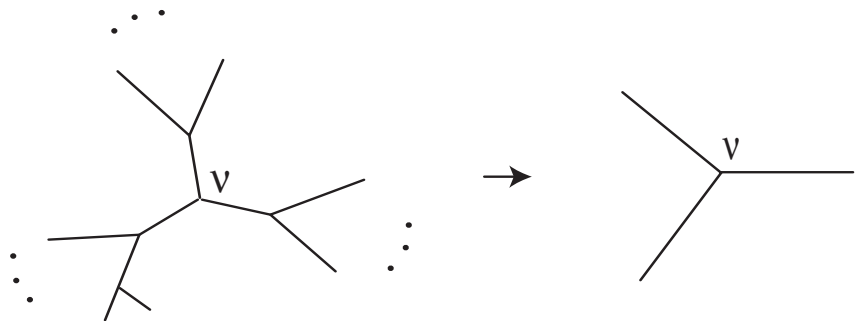


Get maps between ideals, related to $GL(l_3)$-action, and careful use of basic representation theory gives result.

Main result for $\kappa > 2$:

Theorem: For any κ, given *set-theoretic* defining polynomials of $V(\kappa; \kappa, \kappa, \kappa)$, we can explicitly construct *set-theoretic* defining polynomials for V_T for GM model on any trivalent tree T.

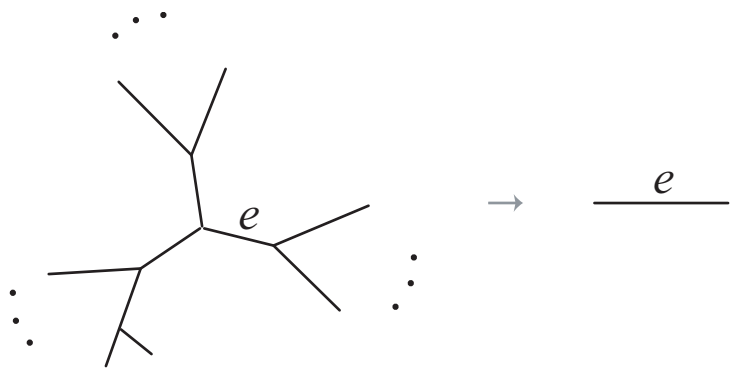
So local structure of the tree determines a collection of phylogenetic invariants *set-theoretically* defining the variety.



n-dim $\kappa \times \cdots \times \kappa$ tensor $P \longrightarrow$ 3-dim $\kappa^{n_1} \times \kappa^{n_2} \times \kappa^{n_3}$ tensor

Do this for all vertices v and ...

and for all edges e:



n-dim $\kappa \times \cdots \times \kappa$ tensor $P \longrightarrow \kappa^{n_1} \times \kappa^{n_2}$ matrix

Implications: Via rank of flattenings, we can potentially say something about data's support for a *particular edge* or *particular node* of a phylogenetic tree.

Furthermore,

support for all edges and nodes = support for tree

New Directions...

What might invariants be good for?

Applications of invariants:

1) Proofs of identifiability of current models

2) Understanding new models

3) Understanding ML

4) Practical inference methods

5) ???

Identifiability of Model Parameters

Question: Given a joint distribution of bases at the leaves from some model, are model parameters (T, stochastic) *identifiable*?

Identifiability is needed to prove the consistency of inference methods such as ML:

> "As sequence length $\longrightarrow \infty$, with probability 1 inference will be correct."

Previously, identifiability of T usually shown by distance + 4-point condition.

New results use invariants to identify T *without* using distances

- GM+I model

No known distance formula, yet T is identifiable by vanishing of 4-leaf invariants.

(For quartet tree, a subset of edge invariants for central edge for GM model are also invariants of GM+I.)

- Covarion model of Tuffley-Steel

  8 states at internal nodes

  $$A^{\text{on}}, A^{\text{off}}, C^{\text{on}}, C^{\text{off}}, G^{\text{on}}, G^{\text{off}}, T^{\text{on}}, T^{\text{off}}$$

  4 observable states at leaves A, C, G, T

  $M_e = \exp(Qt_e)$ where Q is 8×8 rate matrix of special form

This model allows sites to switch between variable/invariable modes in different parts of tree, increases biological realism.

Again, no known distance, but T identifiable through 4-leaf invariants.

(Internal edge flattening according to correct T has rank ≤ 8, flattening according to wrong tree gives higher rank.)

Note:

Covarion result is first proved for more general model, with purely algebraic definition (no common rate matrix).

Specialization to covarion model involves analytic variety.

Also specializes to (re)prove identifiability of other models

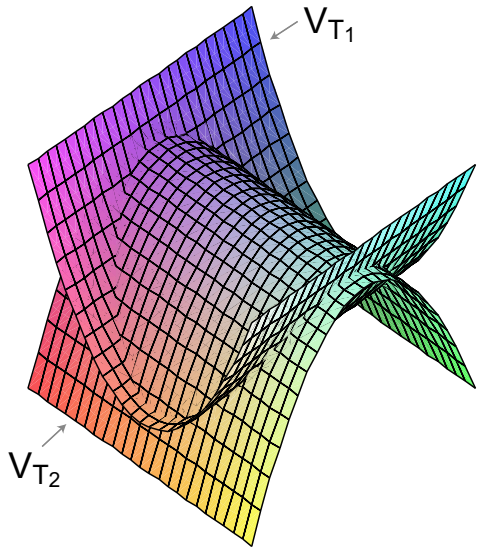
- GM,
- GM+I,
- GM+GM+...+GM model ($< \kappa$ summands),
- rates-across-sites models with $< \kappa$ arbitrary rate classes

Geometric Interpretation of Identifiability:

Tree topology: When is $V_{T_1} = V_{T_2}$?

Since V_{T_1} and V_{T_2} are irreducible varieties of the same dimension, then either

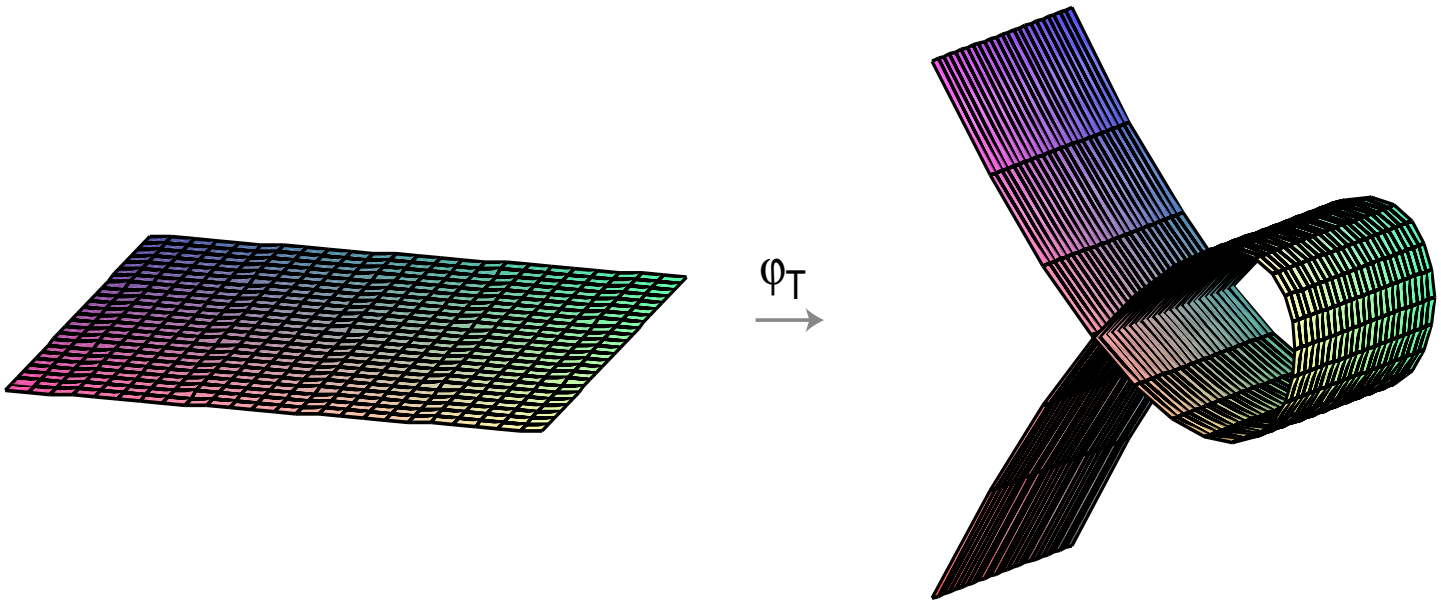
- $V_{T_1} = V_{T_2}$ (tree identifiability fails for all parameter choices) OR
- $V_{T_1} \cap V_{T_2}$ is of lower dimension (tree is identifiable for generic parameters).



Stochastic parameters: Where is V_T non-singular?

For a 'good' parameterization (generically 1-1), all points with non-identifiable parameters lie in singular locus

Singular locus is of lower dimension than variety



New models

Strand Symmetric model (Casanellas–Sullivant, 2005)

a 4-state amalgam of 2-state group-based/2-state GM models

(A, T mutate similarly, C, G mutate similarly)

Analyzing model involves both

- Hadamard-like transform (from group-based model) and
- edge invariants (from GM)

to appear, *Algebraic Statistics for Computational Biology*, CUP

Stable base distribution (AR, 2005)

General model, assuming stable base frequencies throughout tree, has an algebraic definition involving eigenvectors.

- For $\kappa = 2$, variety is rational and have explicit rational inverse to parameterization map
- For $\kappa = 2$, invariant found that checks that internal node distribution agrees with leaves
- For $\kappa > 2$, only partial results

Invariants and Maximum Likelihood

Understanding the geometry of likelihood function: (Chor, Holland, Hendy, Penny, 2000)

Use invariants to show that the likelihood function can have multiple maxima or even continuum of maxima

(Constrained optimization problem for Neyman model, 4-taxa)

Exact ML Optimization: For small trees, use invariants in constrained optimization problem for exact solution of ML problem via computational algebra

(Chor, Khetan, Snir, 2003) Computations for Neyman model.

(Hoşten, Khetan, Sturmfels, 2004) Computations for JC model.

(Chor, Hendy, Snir, 2005) Computations for JC model.

Estimating invariable sites: from understanding invariants/identifiability of parameters for GM+I, can give

explicit rational formulas

for the proportion of Invariable sites. Useful in estimating this parameter for preliminary analysis of data or ML search (?).

Practical inference

- Inefficiency of invariants in simulation studies shown only for linear invariants

(From algebraic geometry perspective, this is not surprising. Identify smallest linear space L so that $L \supset V_T$.)

- Since

distance + 4-point condition '=' higher-degree invariant

there is evidence of potential for better use

- But invariants are best understood for group-based and general models; these are not common-rate-matrix models

Many parameters $\rightsquigarrow$ high variance in inferred trees ??

How to best use invariants for inference needs development

Tree construction using GM invariants/SVD (Eriksson, 2005).

- Focus on edge invariants = matrix rank conditions
- Matrix rank can be computed well via SVD,
- Algorithm constructs tree via joining neighbors from outside

to appear, *Algebraic Statistics for Computational Biology*, CUP

To improve this,

- How can we use vertex invariants? (No SVD analogue yet)
- Rather than iteratively work outside $\longrightarrow$ inside like NJ, try inside $\longrightarrow$ outside
- Test for 'all' splits?

Even if tree is inferred by another method, can invariants put

measure of support on edges/nodes?

- On edges, use measure of 'excess' rank (via SVD) of edge flattening of data as deviation from perfect fit

For 2-state model, 0 deviation on all edges = perfect fit by GM

- For nodes, unclear computationally how to measure excess rank, but...

For 4-state model 0 deviation on all edges, nodes = perfect fit by GM

???

References:

Background Papers on Invariants and Phylogenetic Inference:

Introduction

J. A. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *J. of Class.*, 4:57–71, 1987.

J. A. Lake. A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Bio. Evol.*, 4:167–191, 1987.

Hadamard Methods

M. D. Hendy. The relationship between simple evolutionary tree models and observable sequence data. *Systematic Zoology*, 38:310–321, 1989.

M. D. Hendy and D. Penny. A framework for the quantitative study of evolutionary trees. *Systematic Zoology*, 38:297–309, 1989.

S. N. Evans and T. P. Speed. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.*, 21(1):355–377, 1993.

M. Steel, L. Székely, P. L. Erdös, and P. Waddell. A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. *N.Z. J. Botany*, 31(31):289–296, 1993.

Computational methods/Simulation studies

V. Ferretti and D. Sankoff. The empirical discovery of phylogenetic invariants. *Adv. in Appl. Probab.*, 25(2):290–302, 1993.

V. Ferretti and D. Sankoff. Phylogenetic invariants for more general evolutionary models. *J. Theor. Biol.*, 173:147–162, 1995.

T. R. Hagedorn. Determining the number and structure of phylogenetic invariants. *Adv. in Appl. Math.*, 24(1):1–21, 2000.

J. P. Huelsenbeck. Performance of phylogenetic methods in simulation. *Sys. Biol.*, 44(1):17–48, 1995.

Recent Work

GM and other models

E. S. Allman and J. A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186:113–144, 2003.

E. S. Allman and J. A. Rhodes. Phylogenetic invariants for stationary base composition. *J. Symbolic Comp.*, 2004. to appear, `arXiv:q-bio.PE/0407035`.

E. S. Allman and J. A. Rhodes. Phylogenetic ideals and varieties for the general Markov model. 2004. preprint, `arXiv:math.AG/0410604`.

E. S. Allman and J. A. Rhodes. Phylogenetic invariants and parameter recovery for the general Markov plus invariable sites model. 2005. in preparation.

M. Casanellas and S. Sullivant. The strand symmetric model. In Lior Pachter and Bernd Sturmfels, editors, *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005. to appear.

B. Sturmfels and S. Sullivant. Toric ideals of phylogenetic invariants. *J. Comput. Biol.*,

2004. to appear, arXiv:q-bio.PE/0402015.

Identifiability

E. S. Allman and J. A. Rhodes. The identifiability of a general phylogenetic model, with application to the covarion model. 2005. in preparation.

J. T. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.*, 137(1):51–73, 1996.

M. A. Steel, L.A. Székely, and M. D. Hendy. Reconstructing trees from sequences whose sites evolve at variable rates. *J. Comput. Biol.*, 1(2):153–163, 1994.

New Directions

B. Chor, M. D. Hendy, B. R. Holland, and D. Penny. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Mol. Bio. and Evol.*, 17:1529–1541, 2000.

B. Chor, M. D. Hendy, and S. Snir. Maximum likelihood Jukes-Cantor triplets: analytic solutions. 2005. preprint, arXiv:q-bio.PE/0505054.

N. Eriksson. Tree construction using singular value decomposition. In Lior Pachter and Bernd Sturmfels, editors, *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005. to appear.

S. Hosten, A. Khetan, and B. Sturmfels. Solving the likelihood equations. 2004. preprint, arXiv:math.ST/0408270.

L. Pachter and B. Sturmfels, editors. *Algebraic statistics for computational biology*. Cambridge University Press, 2005. to appear.