Semi-algebraic descriptions of the general Markov model

John A. Rhodes
UAF UNIVERSITY OF ALASKA FAIRBANKS

Phylomania 2010
Hobart, Tas.
November 4-5

Thanks to my collaborators:

Elizabeth Allman,

Mathematics and Statistics, UAF

Amelia Taylor,

Mathematics and Computer Science, Colorado College

GM(k) Model on T:

$k =$ size of some alphabet (state space);

e.g., $k = 2$ (0=R,1=Y), $k = 4$ (A,C,T,G)

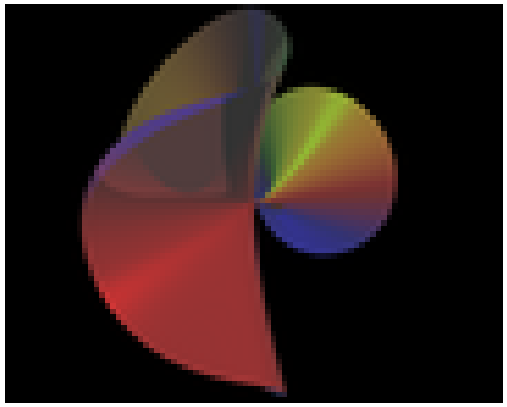
T a rooted tree, n leaves

Pick a vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k) \in [0,1]^k, \sum \pi_i = 1$ to specify a distribution of states at the root of T.

For each edge of T directed away from the root, pick a $k \times k$ stochastic matrix M_e of conditional probabilities of state changes between the endpoints.

These choices determine a joint probability distribution $P \in \mathbb{R}^{k^n}$ of states at the leaves (the pattern distribution)

The General Markov (GM) model on T is the collection of such P for all choices of $\boldsymbol{\pi}$, M_e



Variants:

Require $\boldsymbol{\pi}$, M_e to have positive entries

(To statisticians, this is a very important difference.)

Require M_e to be non-singular

(Standard assumption of GTR and submodels.)

For fixed T with n leaves, the GM model is the image of a polynomial map

$$\phi_T : \Theta_T \to \mathbb{R}^{k^n}.$$

with domain $\Theta_T \subseteq \mathbb{R}^L$ defined by

equalities (e.g., $\sum \pi_i = 1$, $M_e \mathbf{1} = \mathbf{1}$), and

inequalities (e.g., $\pi_i \geq 0$, $\det(M_e \neq 0)$).

Definition. A subset of $\mathbb{R}^m$ defined by polynomial equalities and inequalities is said to be semi-algebraic set.

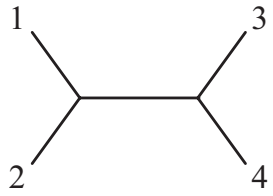
Theorem. *(Tarski-Seidenberg) The polynomial image of a semi-algebraic set is semi-algebraic.*

Thus the GM model on T is a semialgebraic set

Problem: Give an explicit semialgebraic description of the k-state GM model on a tree T.

- *Equalities* (called phylogenetic invariants) have been much studied.
- *Inequalities* are more elusive.

Example: (Cavender-Felsenstein)



T a 4-leaf tree.

The 4-point condition with log-det distance, can be exponentiated and expressed as

$$f_1(P) = f_2(P) > f_3(P)$$

where the f_i are polynomials.

Thus semi-algebraic considerations underlie much theory,

and practical algorithms (NJ).

Precursors to our work:

S. Klaere: $k = 2$, Trees with 3 and 4 leaves (preprint soon!?!)

uses natural coordinates, parameterization,

clever, detailed arguments

P. Zwiernik, J. Smith: $k = 2$, any number of leaves, preprints on arXiv

different coordinates, parameterization

approach seems to require $k = 2$

Goal: Understand the $k = 2$ semialgebraic description in a way that generalizes (at least partially) to $k > 2$.

Approach:

Trees: 3-leaves, then 4-leaves, then more leaves

Parameters: $\mathbb{C}$, then $\mathbb{R}$, then stochastic

Background:

The $2 \times 2 \times 2$ hyperdeterminant (tangle) Δ:

For $P = (p_{ijk})$ a $2 \times 2 \times 2$ array, a distribution from 3-leaf tree

$$\begin{aligned}\Delta(P) = &(p_{000}^2 p_{111}^2 + p_{001}^2 p_{110}^2 + p_{010}^2 p_{101}^2 + p_{011}^2 p_{100}^2) \\ &- 2(p_{000}p_{001}p_{110}p_{111} + p_{000}p_{010}p_{101}p_{111} + p_{000}p_{011}p_{100}p_{111} \\ &+ p_{001}p_{010}p_{101}p_{110} + p_{001}p_{011}p_{110}p_{100} + p_{010}p_{011}p_{101}p_{100}) \\ &+ 4(p_{000}p_{011}p_{101}p_{110} + p_{001}p_{010}p_{100}p_{111}).\end{aligned}$$

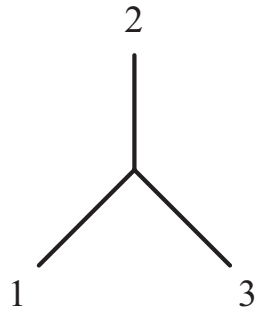
One way to think of Δ (Schläfli):

For a column vector $\mathbf{v} = (x, y)$ of indeterminates,

- $P *_3 \mathbf{v}$ is the sum of matrix slices of P weighted by x and y,
- ... so $\det(P *_3 \mathbf{v})$ is a homogeneous quadratic polynomial in x, y, of form $ax^2 + bxy + cy^2$, with a, b, c quadratic in the entries of P,
- ... so the discriminant, $b^2 - 4ac$, is a quartic in the entries of P, and is in fact $\Delta(P)$.

So $\Delta(P) \neq 0 \Leftrightarrow$ there are exactly two non-zero $\mathbf{v} \in \mathbb{C}^2$ (up to scaling) for which $P *_3 \mathbf{v}$ has rank ≤ 1.

Application of Δ to GM(2) on a 3-leaf tree:



Parameters $\boldsymbol{\pi}, M_1, M_2, M_3$

$$P = (((\mathrm{Diag}(\boldsymbol{\pi})) *_1 M_1) *_2 M_2) *_3 M_3$$

Let $\mathbf{v}$ be the first column of M_3^{-1}.

$$\begin{aligned}
\text{Then } P *_3 \mathbf{v} &= (((\mathrm{Diag}(\boldsymbol{\pi})) *_1 M_1) *_2 M_2) *_3 M_3 *_3 \mathbf{v} \\
&= (((\mathrm{Diag}(\boldsymbol{\pi})) *_1 M_1) *_2 M_2) *_3 (1, 0) \\
&= (((\mathrm{Diag}(\boldsymbol{\pi})) *_3 (1, 0)) *_1 M_1) *_2 M_2 \\
&= M_1^T \,\mathrm{diag}(\pi_0, 0) M_2
\end{aligned}$$

which has rank at most 1.

Proposition 1. *A tensor P is in the image of the complex parameterization map for the GM(2) model on the 3-leaf tree iff its entries sum to 1 and either*

(a) $\Delta(P) \neq 0$, *and* $\det(P *_i \mathbf{1}) \neq 0$ *for* $i = 1, 2, 3$, *or*

(b) $\Delta(P) = 0$, *and all* 2×2 *minors of at least one of the flattenings* $P_{1,23}, P_{2,13}, P_{3,12}$ *are zero.*

In case (a), P is the image of a unique (up to label swapping) choice of non-singular parameters; in case (b), P's preimage is larger.

Note: Only invariant for GM(2) on 3-leaf tree is trivial.

Moreover, since the sign of the discriminant of a quadratic determines whether roots are real or complex, the connnection between Δ and the discrimnant yields...

Proposition 2. *A tensor P is in the image of the real parameterization map for the GM(2) model on the 3-leaf tree if, an only if, it is real, its entries sum to 1, and either*

(a) $\Delta(P) > 0$, *and* $\det(P *_i \mathbf{1}) \neq 0$ *for* $i = 1, 2, 3$, *or*

(b) $\Delta(P) = 0$, *and all* 2×2 *minors of at least one of the flattenings* $P_{1,23}, P_{2,13}, P_{3,12}$ *are zero.*

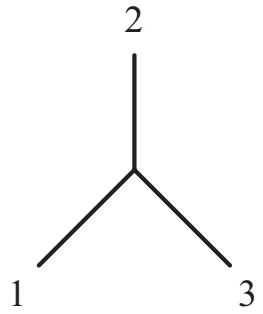
In case (a), P is the image of a unique (up to label swapping) choice of non-singular parameters; in case (b), P's preimage is larger.

Note:

It is not (yet) clear how to generalize the preceding to $k > 2$.

But what follows holds for $k \geq 2$.

Positivity of parameters:



Note the marginalizations of P from 3 to 2 taxa are

$$P_{..+} = P *_3 (1,1) = M_1^T \operatorname{diag}(\boldsymbol{\pi}) M_2$$

$$P_{.+.} = P *_2 (1,1) = M_1^T \operatorname{diag}(\boldsymbol{\pi}) M_3$$

$$P_{+..} = P *_1 (1,1) = M_2^T \operatorname{diag}(\boldsymbol{\pi}) M_3$$

so

$$P_{+..}(P_{.+.})^{-1} P_{..+} = M_2^T \operatorname{diag}(\boldsymbol{\pi}) M_2$$

is a symmetric matrix.

(This was a construction of invariants given in Allman-R 2003.)

But

$$P_{+\cdot\cdot}(P_{\cdot+\cdot})^{-1}P_{\cdot\cdot+} = M_2^T \operatorname{diag}(\boldsymbol{\pi}) M_2$$

is the matrix of a positive definite quadratic form if, and only if, $\pi_0, \pi_1 > 0$.

There are known semialgebraic descriptions of matrices of such forms (and also positive semdefinite ones).

Theorem. *(Sylvester) A symmetric matrix defines a positive definite quadratic form if, and only if, its leading principal minors are positive.*

lth leading principal minor= $l \times l$ subdeterminant in upper left

Theorem. *A tensor P is in the image of the positive parameterization map for the GM(2) model on the 3-leaf tree if, an only if, its entries are positive, its entries sum to 1, and either*

(a) $\Delta(P) > 0$, $\det(P *_i \mathbf{1}) \neq 0$ *for* $i = 1, 2, 3$, *and the 1,1-entries, and the determinants of the following seven matrices are positive:*

$$\det(P_{..+})P_{+..}\operatorname{Cof}(P_{..+})^T P_{.+.},$$

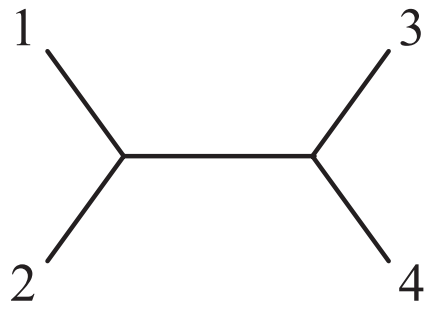
$$\det(P_{..+})P_{i..}^T\operatorname{Cof}(P_{..+})^T P_{.+.},$$

$$\det(P_{..+})P_{+..}^T\operatorname{Cof}(P_{..+})^T P_{.i.},$$

$$\det(P_{+..})P_{.+.}^T\operatorname{Cof}(P_{+..})^T P_{..i}^T,$$

(b) $\Delta(P) = 0$, *and all* 2×2 *minors of at least one of the flattenings* $P_{1,23}, P_{2,13}, P_{3,12}$ *are zero.*

4-leaf Tree:



P is $2 \times 2 \times 2 \times 2$.

First, marginalizing out any taxon i, gives $2 \times 2 \times 2$ array

$$P *_i (1, 1)$$

which arises from a 3-leaf tree, and hence earlier theorems apply.

Second, all non-trivial invariants for GM(2) on trees with 4 or more leaves are known (Allman-R, 2007).

The key ones are edge invariants:

If T has split $12|34$, the 4×4 flattening $P_{12,34}$ has rank 2, so all its 3×3 subdeterminants are 0.

Theorem. *Let P be a complex $2 \times 2 \times 2 \times 2$ with entries summing to 1. Then P arises from complex non-singular parameters on a 4-leaf T iff*

1. *All marginalizations of P to 3-taxon sets arise from complex non-singular parameters on 3-leaf trees, and*
2. *The edge invariants are satisfied by P.*

For real parameters, replace all occurances of 'complex' by 'real'.

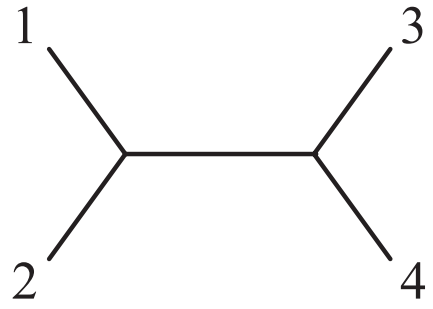
Positivity of parameters:

For root distribution $\boldsymbol{\pi}$ and stochastic matrices M_e on pendant edges, follows from 3-leaf case.

Matrices on internal edges require more. . .

We first 'adjust' P

If $T = 12|34$, and P arises from matrices M_1, M_2, M_3, M_4 on pendant edges, M_5 on internal



Let

$$N_{32} = P^T_{+\cdot\cdot+} = M_3^T M_5^T \operatorname{diag}(\boldsymbol{\pi}) M_2$$

$$N_{31} = P^T_{\cdot+\cdot+} = M_3^T M_5^T \operatorname{diag}(\boldsymbol{\pi}))\ M_1$$

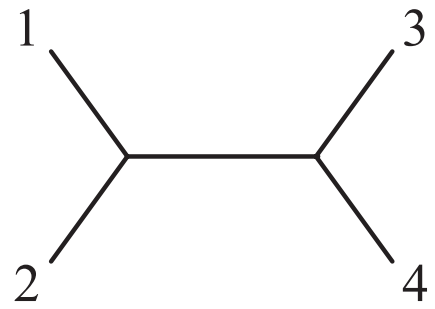
so $N_{32}^{-1} N_{31} = M_2^{-1} M_1$. Then

$$\hat{P} = P *_2 N_{32}^{-1} N_{31}$$

arises from same parameters but with M_1 replacing M_2.

A similar trick produces $\hat{\hat{P}}$ from parameters with

$$M_1 = M_2, \quad M_3 = M_4.$$



Now flatten $\hat{\hat{P}}$ to a 4×4 the wrong way according to $13|24$.

Then

$$\hat{\hat{P}}_{13,24} = A^T D A$$

where A depends on M_1, M_3, and

D is 4×4 diagonal with entries of $\mathrm{diag}(\boldsymbol{\pi}) M_5$

So the entries of M_5 are positive iff $\hat{\hat{P}}_{13,24}$ has positive leading principal minors.

A bit more work extends this to 5 or more taxa.

Summary:

This yields *one form* of a complete semialgebraic description of GM(2).

But this is not the only possibility, or necessarily the best,

Currently working out an alternate approach based on Sturm sequences, likely to lower degree of constraints

For GM(k), this shows how to impose positivity, or non-negativity constraints on parameters.

But large gaps remain in describing GM(k) model for complex or real parameters