

What is Algebraic Statistics?



Seth Sullivant

Society of Fellows, Harvard University

Outline

- ◆ “Crash Course” in Algebraic Geometry
- ◆ The Philosophy of Algebraic Statistics
- ◆ Three Examples:
 - ◆ Warm up: Independent Random Variables
 - ◆ Markov Bases of Graphical Models
 - ◆ Phylogenetic Invariants

Introduction to Algebraic Geometry

Let $g_1(p_1, \dots, p_m), \dots, g_k(p_1, \dots, p_m) \in \mathbb{C}[p_1, \dots, p_m]$.

Definition: $V(g_1, \dots, g_k) = \left\{ (a_1, \dots, a_m) \in \mathbb{C}^m \mid \right.$
 $\left. g_1(a_1, \dots, a_m) = \dots = g_k(a_1, \dots, a_m) = 0 \right\}$

is the **variety defined by** g_1, \dots, g_k
or the **zero set of** g_1, \dots, g_k .

Given $S \subseteq \mathbf{C}^m$ form

$$I(S) = \left\{ g(p_1, \dots, p_m) \in \mathbf{C}[p_1, \dots, p_m] \mid \right. \\ \left. g(a_1, \dots, a_m) = 0 \text{ for all } (a_1, \dots, a_m) \in S \right\}$$

Note: If f and $g \in I(S)$ and h arbitrary then $f + g \in I(S)$ and $hf \in I(S)$.

So $I(S)$ is an **ideal**: the ideal of polynomial functions vanishing on S .

Given $g_1, \dots, g_k \in \mathbf{C}[p_1, \dots, p_m]$

$$\langle g_1, \dots, g_k \rangle = \left\{ \sum_{i=1}^k h_i g_i \mid h_1, \dots, h_k \in \mathbf{C}[p_1, \dots, p_m] \right\}$$

is the **ideal generated by** g_1, \dots, g_k .

Hilbert Basis Theorem: If $I \subseteq \mathbf{C}[p_1, \dots, p_m]$ is an ideal, there exist g_1, \dots, g_k such that

$$\langle g_1, \dots, g_k \rangle = I.$$

Corollary: $I(\mathcal{S})$ has a finite generating set.

Let $f_1(\theta_1, \mathbf{K}, \theta_d), \mathbf{K}, f_m(\theta_1, \mathbf{K}, \theta_d) \in \mathbf{C}[\theta_1, \mathbf{K}, \theta_d]$ and define

$$\mathcal{S} = \left\{ \begin{pmatrix} f_1(a_1, \mathbf{K}, a_d) \\ \mathbf{M} \\ f_m(a_1, \mathbf{K}, a_d) \end{pmatrix} \in \mathbf{C}^m \mid (a_1, \mathbf{K}, a_d) \in \mathbf{C}^d \right\}$$

\mathcal{S} is called a **rational parameterization (r.p.)**.

Theorem: If \mathcal{S} is a **r.p.** and $I(\mathcal{S}) = \langle g_1, \mathbf{K}, g_k \rangle$ then

\mathcal{S} and $V(g_1, \mathbf{K}, g_k)$ differ by a set of dimension

less than the dimension of \mathcal{S} .

Example: Hardy-Weinberg Equilibrium

Consider a gene with two alleles: A and B. Let $X = \#$ of times A appears in a pair of matching chromosomes.

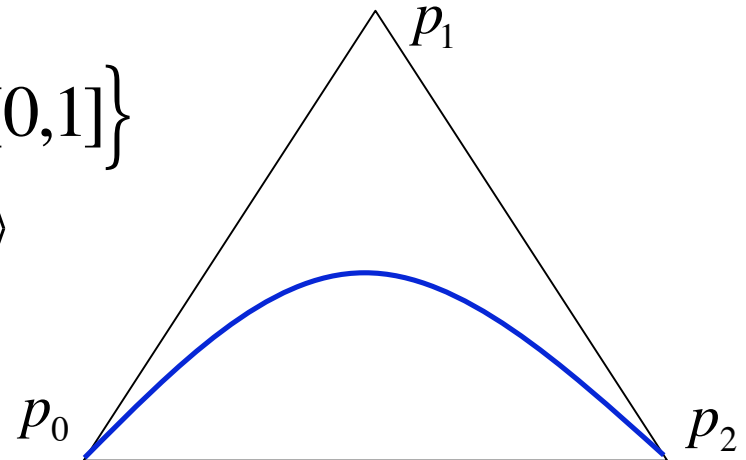
If A and B are in Hardy-Weinberg equilibrium, the alleles are selected **independently** on each chromosome, that is, there exists θ such that $P(X = 0) = \theta^2$

$$P(X = 1) = 2\theta(1 - \theta) \quad \text{and} \quad P(X = 2) = (1 - \theta)^2$$

$$S = \{(\theta^2, 2\theta(1 - \theta), (1 - \theta)^2) \mid \theta \in [0, 1]\}$$

$$I(S) = \langle p_0 + p_1 + p_2 - 1, p_1^2 - 4p_0p_2 \rangle$$

$$\subset \mathbf{C}[p_0, p_1, p_2]$$



The Philosophy of Algebraic Statistics

- ◆ Many natural families of probability distributions on discrete random variables are (parameterized) varieties.
- ◆ Understanding the algebro-geometric structure of these varieties provides new insight for problems in probability and statistics.
- ◆ Knowing the ideal generators for these varieties is potentially useful for making statistical inferences.

Independent Random Variables

X, Y discrete random variables with
 $X \in \{1, 2, \dots, d_1\}$ and $Y \in \{1, 2, \dots, d_2\}$

Definition: X is independent of Y if and only if

$$P(X = i, Y = j) = P(X = i)P(Y = j)$$

for all i and j .

Parametric Representation

The set of **all** independent distributions for X and Y is the set of $d_1 \times d_2$ matrices:

$$S = \left\{ \left(\begin{array}{cccc} \eta_1 \theta_1 & \eta_1 \theta_2 & \text{L} & \eta_1 \theta_{d_2} \\ \eta_2 \theta_1 & \eta_2 \theta_2 & \text{L} & \eta_2 \theta_{d_2} \\ \text{M} & \text{M} & \text{O} & \text{M} \\ \eta_{d_1} \theta_1 & \eta_{d_1} \theta_2 & \text{L} & \eta_{d_1} \theta_{d_2} \end{array} \right) \left| \begin{array}{l} 0 \leq \eta_i \leq 1 \quad \forall i \quad \text{and} \quad \sum_{i=1}^{d_1} \eta_i = 1 \\ 0 \leq \theta_j \leq 1 \quad \forall j \quad \text{and} \quad \sum_{j=1}^{d_2} \theta_j = 1 \end{array} \right. \right\}$$

S is a rational parametrization with $f_{ij}(\eta, \theta) = \eta_i \theta_j$

Implicit Representation

X and Y are independent if they satisfy the well-known odds-ratio conditions:

$$\frac{p_{ij}p_{kl}}{p_{il}p_{kj}} = 1 \quad \text{for all } i,j,k,l.$$

$$\begin{pmatrix} p_{11} & p_{12} & \text{L} & p_{1d_2} \\ p_{21} & p_{22} & \text{L} & p_{2d_2} \\ \text{M} & \text{M} & \text{O} & \text{M} \\ p_{d_11} & p_{d_12} & \text{L} & p_{d_1d_2} \end{pmatrix}$$

This is equivalent to saying all 2x2 minors are zero. In the language of ideals:

$$I(S) = \left\langle p_{ij}p_{kl} - p_{il}p_{kj} \mid i,k \in \{1, \dots, d_1\}, j,l \in \{1, \dots, d_2\} \right\rangle$$

Segre Embedding = Independence

$N =$ Birth Month

	J	F	M	A	M	J	J	A	S	O	N	D
J	1	0	0	0	1	2	0	0	1	0	1	0
F	1	0	0	1	0	0	0	0	0	1	0	2
M	1	0	0	0	2	1	0	0	0	0	0	1
A	3	0	2	0	0	0	1	0	1	3	1	1
M	2	1	1	1	1	1	1	1	1	1	1	0
J	2	0	0	0	1	0	0	0	0	0	0	0
J	2	0	2	1	0	0	0	0	1	1	1	2
A	0	0	0	3	0	0	1	0	0	1	0	2
S	0	0	0	1	1	0	0	0	0	0	1	0
O	1	1	0	2	0	0	1	0	0	1	1	0
N	0	1	1	1	2	0	0	2	0	1	1	0
D	0	1	1	0	0	0	1	0	0	0	0	0

Are birth month and death month independent?

Pearson's χ^2 test

$$U = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{(N_{ij} - \hat{N}_{ij})^2}{\hat{N}_{ij}}$$

Asymptotically:

$$U \sim \chi_{(d_1-1)(d_2-1)}^2$$

This data is too sparse for the asymptotics to be valid.

Markov Bases of Graphical Models

Let $\mathbf{X}_1, \mathbf{K}, \mathbf{X}_n$ be discrete random variables, each \mathbf{X}_i takes values in $\mathbf{X}_i \in \{1, \mathbf{K}, d_i\}$.

A graphical model encodes special factorizations of the joint distribution.

$$P(\mathbf{X}_1 = u_1, \mathbf{K}, \mathbf{X}_n = u_n) = P(u_1, \mathbf{K}, u_n) = p_{u_1 \perp u_n}$$

Example:

$$P(u_1, u_2) = P(u_1)P(u_2)$$

“ \mathbf{X} independent of \mathbf{Y} ”

1 ● ● 2

Let $\Delta = \{C_1, \dots, C_m\}$ where $C_i \subset \{1, \dots, n\}$.

The **graphical model** consists of the joint distributions

$$P(u_1, \dots, u_n) = \frac{1}{Z} \prod_{C_i \in \Delta} \Psi_{C_i}(u_{C_i}).$$

As the Ψ_{C_i} range over all their possible values this gives a rationally parameterized family of probability distributions!

Problem: Find the ideal of functions I_Δ that vanish on this family of probability distributions.

Example: $\Delta = \{\{1\}, \{2\}\}$ 1 ● ● 2

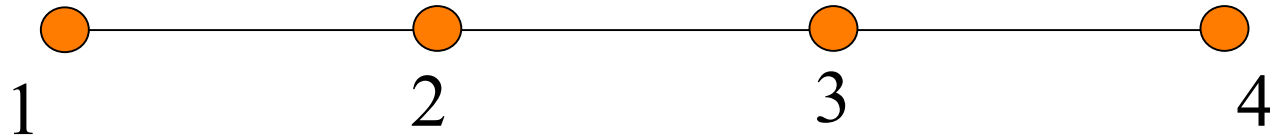
$$P(u_1, u_2) = \Psi(u_1)\Phi(u_2)$$

$$p_{ij} = \eta_i \theta_j$$

$$\begin{pmatrix} p_{11} & p_{12} & \vdots & p_{1d_2} \\ p_{21} & p_{22} & \vdots & p_{2d_2} \\ \vdots & \vdots & \vdots & \vdots \\ p_{d_11} & p_{d_12} & \vdots & p_{d_1d_2} \end{pmatrix} \quad \text{has rank 1.}$$

$$I_\Delta = \left\langle p_{ij}p_{kl} - p_{il}p_{kj} \mid i, k \in \{1, \dots, d_1\}, j, l \in \{1, \dots, d_2\} \right\rangle$$

Example: $\Delta = \{\{1,2\}, \{2,3\}, \{3,4\}\}$



$$p_{ijkl} = \eta_{ij} \theta_{jk} \gamma_{kl}$$

$$I_{\Delta} = \left\langle \begin{array}{l} p_{i_1 j k_1 l_1} p_{i_2 j k_2 l_2} - p_{i_1 j k_2 l_2} p_{i_2 j k_1 l_1} \\ p_{i_1 j_1 k l_1} p_{i_2 j_2 k l_2} - p_{i_1 j_1 k l_2} p_{i_2 j_2 k l_1} \end{array} \right\rangle$$

Testing Contingency Tables

	Black	Brown	Red	Blonde	Total
Brown	68	119	26	7	220
Blue	20	84	17	94	215
Hazel	15	54	14	10	93
Green	5	29	14	16	64
Total	108	286	71	127	592

Are hair and eye color independent?

Generally: Does the model fit the data?

Compute p-values of exact test: generate random draws from the set of all nonnegative integral tables with the same sufficient statistics (i.e. row and column sums) and compare χ^2 statistics.

How to generate random tables?

Use MCMC to take a random walk over the set of nonnegative integral tables with fixed sufficient statistics. Use “moves” that preserve sufficient statistics to add/subtract from a starting table.

e.g.

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \longrightarrow p_{22}p_{34} - p_{24}p_{32}$$

Definition: A set of “moves” M (tables with zero sufficient statistics) is called a Markov basis for Δ if for any choice of sufficient statistics one can run an irreducible (connected) Markov chain over the set of tables with those fixed sufficient statistics.

Notation: For $\mathbf{m} \in M$ write $\mathbf{m} = \mathbf{m}^+ - \mathbf{m}^-$.

$$\mathbf{p}^{\mathbf{m}} = p_{11 \llcorner 1}^{m_{11 \llcorner 1}} p_{11 \llcorner 2}^{m_{11 \llcorner 2}} \llcorner p_{d_1 \llcorner d_n}^{m_{d_1 \llcorner d_n}}$$

Theorem: (Diaconis-Sturmfels) M is a Markov basis for

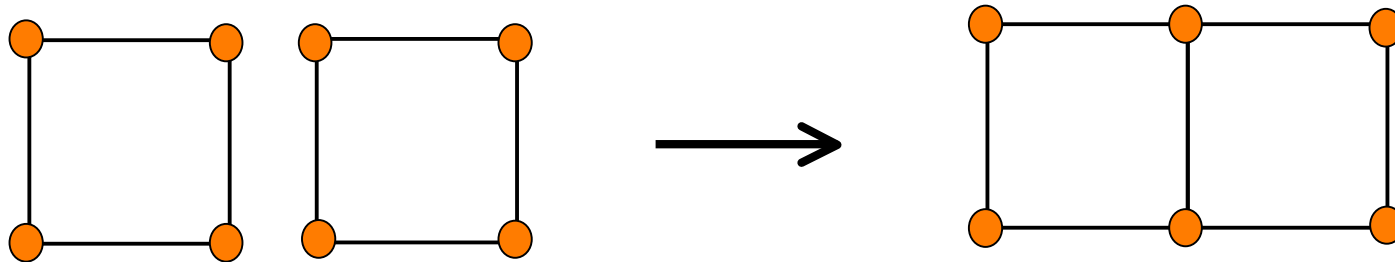
Δ if and only if $\langle \mathbf{p}^{\mathbf{m}^+} - \mathbf{p}^{\mathbf{m}^-} \mid \mathbf{m} \in M \rangle = I_{\Delta}$.

Theorems about Markov Bases

Theorem: (Takken 1999 ; Dobra, 2001)

Decomposable models have degree 2 Markov bases.

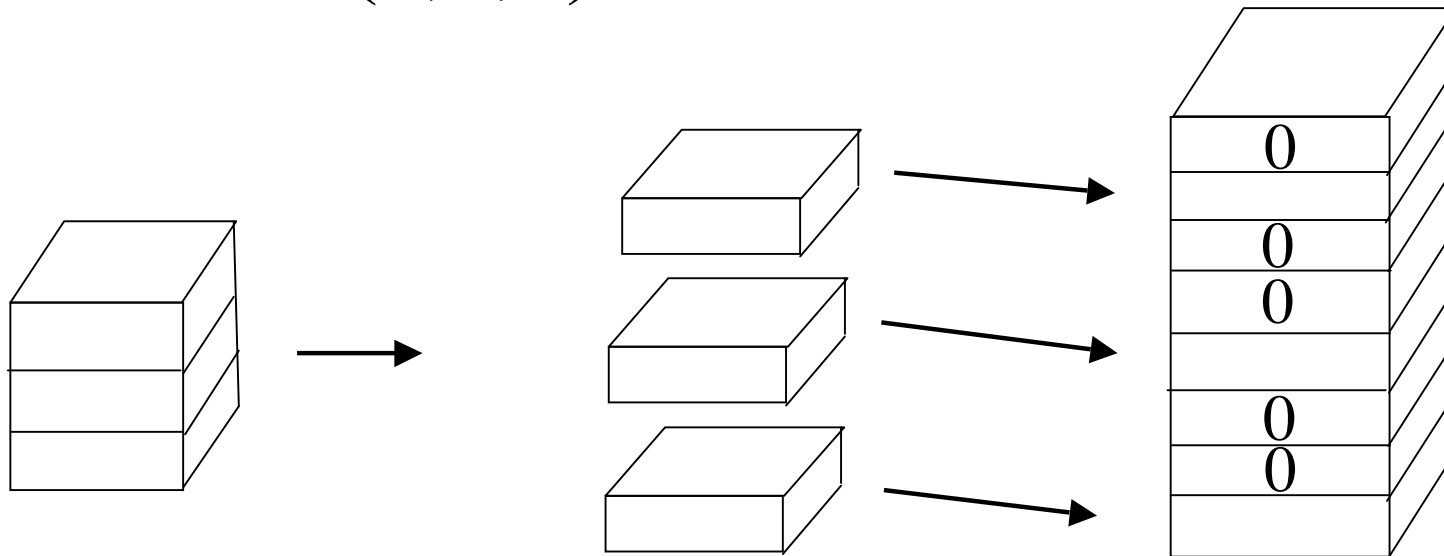
Theorem: (Dobra, Hosten, --- 2002) Can answer Markov basis questions by restricting attention to “prime subgraphs”.



Theorem: (Hosten, ---) Fix d_1, \mathbf{K}, d_{n-1} and let d_n vary. There exists a constant $m = m(d_1, \mathbf{K}, d_{n-1})$ such that the Markov bases for d_n large are determined by the Markov bases when $d_n = m$.

Example: $m(3,4) = 8$ (Aoki, Takemura)

$m(2,2,2) \leq 12$ (Hosten, ---)



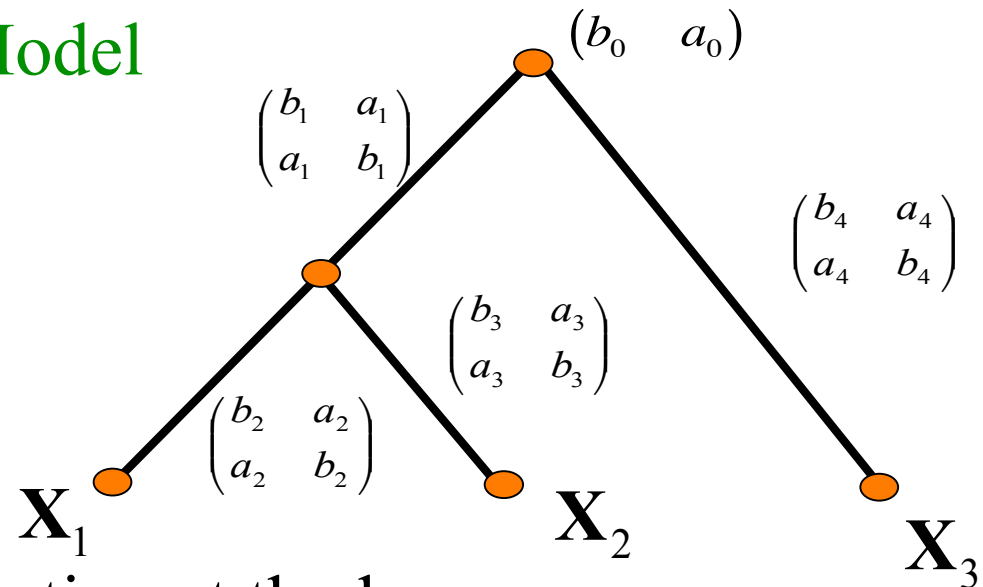
Phylogenetic Invariants

Jukes-Cantor binary Model

$(b_0 \ a_0)$ root dist

a_i “swap” prob.

$$b_i = 1 - a_i$$



Determines a joint distribution at the leaves, e.g.

$$P(\mathbf{X}_1 = 0, \mathbf{X}_2 = 1, \mathbf{X}_3 = 0) =$$

$$b_0 b_1 b_2 a_3 b_4 + b_0 a_1 a_2 b_3 b_4 + a_0 a_1 b_2 a_3 a_4 + a_0 b_1 a_2 b_3 a_4$$

This family of distributions is a poly. param. set!

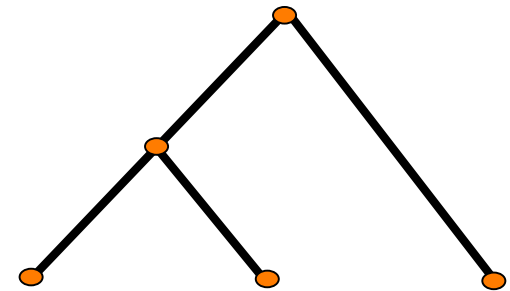
How to determine the tree that best fits the data?

Cavender-Felsenstein (1987) and Lake (1987) proposed finding the polynomial functions that vanish on the family of probability distributions for each tree.

Can evaluate the “phylogenetic invariants” to determine the best fit.

Example:

$$p_{001}p_{010} - p_{000}p_{011} - p_{101}p_{110} + p_{100}p_{111}$$
$$p_{001}p_{100} - p_{000}p_{101} - p_{011}p_{110} + p_{010}p_{111}$$



Advantages: Can find a good fit without estimating parameters. Exact solution of likelihood equations.

Jukes-Cantor DNA model

$$\begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix}$$

b transition prob

Kimura 2-parameter model

$$\begin{pmatrix} a & b & c & c \\ b & a & c & c \\ c & c & a & b \\ c & c & b & a \end{pmatrix}$$

b transition prob

c transversion prob

Kimura 3-parameter model

$$\begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix}$$

b transition prob

c, d transversion prob

Group Structure

Evans and Speed (1993) and Szekeley et al (1993) noticed that all these models have an underlying group structure.

$$P(X = g_2 | Y = g_1) = f(g_1 - g_2) \quad \begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix}$$

Using this fact, can apply discrete Fourier transform to make a linear change of coordinates and simplify the parameterization.

Theorems about Phylogenetic Invariants (with B. Sturmfels)

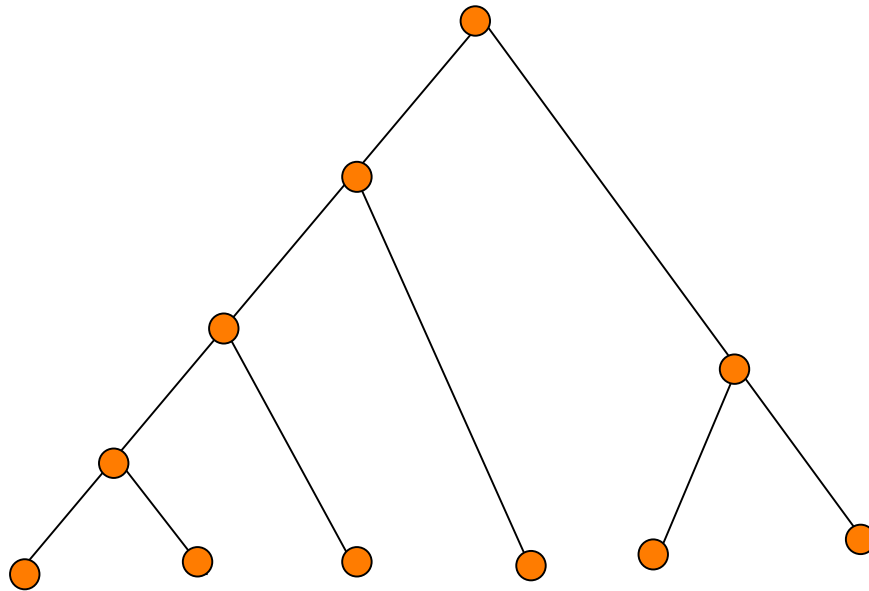
Theorem: The Jukes-Cantor binary model has ideal of invariants generated in degree 2 for any tree.

Theorem: The Jukes-Cantor DNA model has ideal of invariants generated in degree 3 or less for binary trees.

Theorem: The Kimura 2-parameter model has ideal of invariants generated in degree 4 or less for binary trees.

Theorem: The Kimura 3-parameter model has ideal of invariants generated in degree 4 or less for binary trees.

Even more is true!



Phylogenetic invariants correspond to **local features** of the tree:

Splits give degree 2 invariants

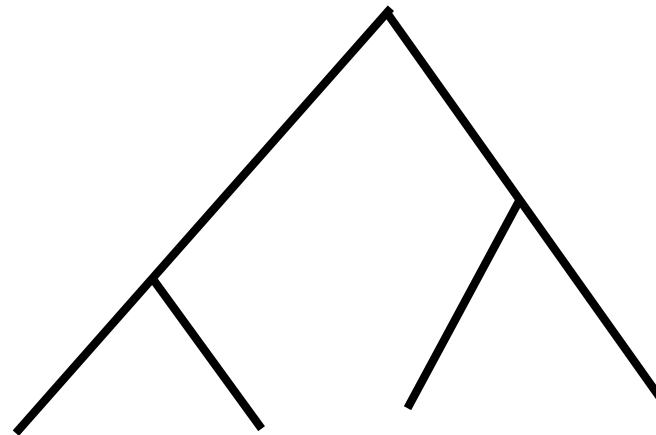
Interior nodes give higher degree invariants

Leads to new tree construction algorithms:
Eriksson and Garcia

Small Phylogenetic Trees Website

<http://www.math.tamu.edu/~lgp/small-trees/>

Google: “Small phylogenetic trees”



Summary

- ◆ Algebraic varieties appear frequently in probabilistic and statistical models.
- ◆ Understanding the algebraic structure provides new insights into these areas.
- ◆ Knowing the defining equations can be useful for making statistical inferences.