Geometry of Phylogenetic Inference

Matilde Marcolli
CS101: Mathematical and Computational Linguistics

Winter 2015
References


Hidden Markov Models

• $n$ observed states $Y_1, \ldots, Y_n$, each taking $\ell$ possible values
• $n$ hidden states $X_1, \ldots, X_n$, each taking $k$ possible values
• conditional independence

\[
P(X_i|X_1, \ldots, X_{i-1}) = P(X_i|X_{i-1})
\]

\[
P(Y_i|X_1, \ldots, X_i, Y_1, \ldots, Y_{i-1}) = P(Y_i|X_i)
\]

• special case: all transitions $X_{i-1} \mapsto X_i$ same $k \times k$-stochastic matrix $P = (p_{ij})$; all transitions $X_i \mapsto Y_i$ same $k \times \ell$-stochastic matrix $T = (t_{ij})$
• a HMM described by the image of a polynomial map

\[ \Phi : \mathbb{R}^{k(k+1)} \rightarrow \mathbb{R}^{\ell^n} \]

of degree \( n - 1 \) bi-homogeneous in the coordinates \( p_{ij} \) and \( t_{ij} \)

• plus added positivity and normalization conditions (stochastic matrices and probability distributions)

• Example with \( k = \ell = 2 \) and \( n = 3 \), \( \Phi = (\Phi_{ijk}) : \mathbb{R}^8 \rightarrow \mathbb{R}^8 \)

\[
\Phi_{ijk} = p_{00}p_{00}t_{0i}t_{0j}t_{0k} + p_{00}p_{01}t_{0i}t_{0j}t_{1k} + p_{01}p_{10}t_{0i}t_{1j}t_{0k} + p_{01}p_{11}t_{0i}t_{1j}t_{1k} \\
+ p_{10}p_{00}t_{1i}t_{0j}t_{0k} + p_{10}p_{01}t_{1i}t_{0j}t_{1k} + p_{11}p_{10}t_{1i}t_{1j}t_{0k} + p_{11}p_{11}t_{1i}t_{1j}t_{1k}
\]
• **invariants** of the HMM: polynomial functions on $\mathbb{R}^{\ell n}$ that vanish on the image of $\Phi$
• ideal $\mathcal{I}_\Phi$ generated by invariants? small $k, \ell, n$ Gröbner bases; larger computationally hard

**Questions**

• Viterbi sequence: find the most likely hidden data given observed data
• find all parameter values for a model that result in the same observed distribution
• find what parameter-independent relations hold between the observed probabilities $P_{i_1, \ldots, i_n} = \Phi_{i_1, \ldots, i_n}$
Phylogenetic Algebraic Geometry

- \( T \) a rooted binary tree with \( n \) leaves (hence \( 2n - 2 \) edges)
- At each vertex a binary random variable (e.g. one of the syntactic parameters)
- Probability distribution at the root vertex \( \pi = (p, 1 - p) \)
- Along each edge \( e \) transition matrix: stochastic matrix \( P_e = (p_{ij}^{(e)}) \) with \( \sum_i p_{ij}^{(e)} = 1 \)
- these represent the probabilities that a mutation in the parameter happens along that edge
Model Parameters

• the random variables at the leaves of the tree are \textit{observed}; the random variables at the interior nodes are \textit{hidden} (assuming no direct knowledge of the “ancient languages” in the family)

• matrix entries of transition matrices $P_e$ and probability $\pi$ at root vertex are \textit{model parameters}

• number of parameters $N = (2n - 2)k^2 + k$
  (binary variable $k = 2$)
Polynomial Map

• at the $n$ leaves there are $k^n = 2^n$ possible observations
• the probability of an observation at the leaves is a polynomial function of the parameters
• can view this as a complex polynomial

$$\Phi: \mathbb{C}^N \rightarrow \mathbb{C}^{2^n}$$

plus some (real) normalization conditions

• polytope $\Delta \subset \mathbb{R}_+^N \subset \mathbb{C}^N$ determined by the conditions

$$\pi_1 + \pi_2 = 1$$
$$\sum_i p^{(e)}_{ij} = 1$$

with $\pi_i \geq 0$ and $p^{(e)}_{ij} \geq 0$

• $\Phi$ should map $\Delta$ to a cube $\mathcal{I}^n$ in $\mathbb{C}^{2^n}$ where $[0, 1] \simeq \mathcal{I} \subset \mathbb{C}^2$ is

$$\mathcal{I} = \{(p_1, p_2) | p_1 + p_2 = 1, p_i \geq 0\}$$
Example

\[ \Phi_{ijk} = \pi_0 a_i b_{00} c_0 d_{0k} + \pi_0 a_{0i} b_{01} c_{1j} d_{1k} + \pi_1 a_{1i} b_{10} c_0 d_{0k} + \pi_1 a_{1i} b_{11} c_{1j} d_{1k} \]

there are 8 such polynomials: \( i, j, k \in \{0, 1\} \)
• polynomial $\Phi$ is **homogeneous** in the parameters
• can view $\Phi$ as a map of **projective spaces**
• in the previous example

\[ \Phi : \mathbb{C}^4 \times \mathbb{C}^4 \times \mathbb{C}^4 \times \mathbb{C}^4 \times \mathbb{C}^2 \rightarrow \mathbb{C}^8 \]

\[ \Phi : \mathbb{P}^3(\mathbb{C}) \times \mathbb{P}^3(\mathbb{C}) \times \mathbb{P}^3(\mathbb{C}) \times \mathbb{P}^3(\mathbb{C}) \times \mathbb{P}^1(\mathbb{C}) \rightarrow \mathbb{P}^7(\mathbb{C}) \]

homogeneous with respect to each group of variables $a, b, c, d, \pi$

• the **fibers** of this morphism give all possible values of parameters (before imposing real normalization conditions) that give a certain probability at the leaves
Algebraic varieties occurring in these models

- Toric varieties (including Segre varieties and Veronese varieties)
- Determinantal varieties: the tree structure imposes rank constraints on matrices built starting from observed probabilities at the leaves
- Example: Segre embedding

\[
p_{ijkl} = u_i v_j w_k x_l \quad i, j, k, l \in \{0, 1\}
\]
• Prime ideal defining this variety: generated by $2 \times 2$ minors of $4 \times 4$-matrices

\[
\begin{pmatrix}
p_{0000} & p_{0001} & p_{0010} & p_{0011} \\
p_{0100} & p_{0101} & p_{0110} & p_{0111} \\
p_{1000} & p_{1001} & p_{1010} & p_{1011} \\
p_{1100} & p_{1101} & p_{1110} & p_{1111}
\end{pmatrix},
\begin{pmatrix}
p_{0000} & p_{0001} & p_{0100} & p_{0101} \\
p_{0010} & p_{0011} & p_{0110} & p_{0111} \\
p_{1000} & p_{1001} & p_{1100} & p_{1101} \\
p_{1010} & p_{1011} & p_{1110} & p_{1111}
\end{pmatrix},
\begin{pmatrix}
p_{0000} & p_{0010} & p_{0100} & p_{0110} \\
p_{0001} & p_{0011} & p_{0101} & p_{0111} \\
p_{1000} & p_{1010} & p_{1100} & p_{1110} \\
p_{1001} & p_{1011} & p_{1101} & p_{1111}
\end{pmatrix}
\]

corresponding to

\[
\begin{array}{cccc}
\text{u} & \text{v} & \text{w} & \text{x} \\
\text{u} & \text{w} & \text{v} & \text{x} \\
\text{u} & \text{x} & \text{v} & \text{w}
\end{array}
\]
Secant variety of the Segre variety

• $X$ nine-dimensional subvariety of $\mathbb{P}^{15}$ given by all $2 \times 2 \times 2 \times 2$-tensors of rank at most 2

\[
p_{ijkl} = \pi_0 u_0 v_0 w_0 x_0 + \pi_1 u_1 v_1 w_1 x_1
\]

• Ideal generated by all $3 \times 3$-minors of previous matrices

\[
X = X_{(12)(34)} \cap X_{(13)(24)} \cap X_{(14)(23)}
\]
Determinantal variety

- each determinantal variety corresponds to a Markov model on one of the binary trees: $X_{(12)(34)}$ is defined by

\[
p_{ijkl} = \pi_0 (a_{00} u_0 v_0 + a_{01} u_1 v_1) (b_{00} w_0 x_0 + b_{01} w_1 x_1) + \pi_1 (a_{10} u_0 v_0 + a_{11} u_1 v_1) (b_{10} w_0 x_0 + b_{11} w_1 x_1)
\]

this corresponds to vanishing of all $3 \times 3$-minors in first matrix

- stratification of $\mathbb{P}^{2n-1}$ by phylogenetic models $X$
Special case: Jukes-Cantor model

- special case where all the edge matrices $P_e$ have the form

\[ P_e = \begin{pmatrix} p_0 & p_1 \\ p_1 & p_0 \end{pmatrix} \]

- it is known that in this case an explicit change of coordinates describes it as a toric variety.

General Idea of Phylogenetic Algebraic Geometry

- generators of the ideal defining the complex variety = phylogenetic invariants

- which phylogenetic invariants suffice to distinguish between different Markov models?

- parameter inference from tropicalization of the algebraic variety
Tropical Semiring

- min-plus (or tropical) semiring $\mathbb{T} = \mathbb{R} \cup \{\infty\}$, with operations $\oplus$ and $\odot$ given by
  
  $$x \oplus y = \min\{x, y\},$$

  with $\infty$ the identity element for $\oplus$ and with

  $$x \odot y = x + y,$$

  with 0 the identity element for $\odot$

- operations $\oplus$ and $\odot$ satisfy associativity and commutativity and distributivity of the product $\odot$ over the sum $\oplus$

- addition is no longer invertible and is idemponent

  $$x \oplus x = \min\{x, x\} = x$$
Tropical polynomials

• function $\phi : \mathbb{R}^n \to \mathbb{R}$ of the form

$$\phi(x_1, \ldots, x_n) = \bigoplus_{j=1}^m a_j \odot x_1^{k_{j1}} \odot \cdots \odot x_n^{k_{jn}}$$

$$= \min\{ \ a_1 + k_{11}x_1 + \cdots + k_{1n}x_n, \ a_2 + k_{21}x_1 + \cdots + k_{2n}x_n, \ \cdots \ a_m + k_{m1}x_1 + \cdots + k_{mn}x_n \ \}.$$

• tropicalization: algebraic varieties become piecewise linear spaces
• can recover information about a variety from its tropicalization
• in the previous HMM example with $n = 3$ and $k = \ell = 2$ the tropicalization of the polynomials $\Phi_{ijk}$

\[
\Phi_{ijk} = p_{00}p_{00}t_{0i}t_{0j}t_{0k} + p_{00}p_{01}t_{0i}t_{0j}t_{1k} + p_{01}p_{10}t_{0i}t_{1j}t_{0k} + p_{01}p_{11}t_{0i}t_{1j}t_{1k} \\
+ p_{10}p_{00}t_{1i}t_{0j}t_{0k} + p_{10}p_{01}t_{1i}t_{0j}t_{1k} + p_{11}p_{10}t_{1i}t_{1j}t_{0k} + p_{11}p_{11}t_{1i}t_{1j}t_{1k}
\]

is given by

\[
\tau_{ijk} = \min\{u_{h_1h_2} + u_{h_2h_3} + v_{h_1i} + v_{h_2j} + v_{h_3k} \mid (h_1, h_2, h_3) \in \{0, 1\}^3\}
\]

where $u_{ab} = -\log(p_{ab})$ and $v_{ab} = -\log(t_{ab})$

• Viterbi sequence: $(h_1, h_2, h_3)$ realizing minimum, given observed $(i, j, k)$ is the Viterbi sequence of hidden data
Newton polytope

- polynomial $f = \sum_{\omega \in \mathbb{Z}^n} a_{\omega} x^\omega$ with $x^\omega = x_1^{\omega_1} \cdots x_n^{\omega_n}$

- Newton polytope

$$\mathcal{N}(f) = \text{Convex Hull}\{\omega \in \mathbb{Z}^n \mid a_\omega \neq 0\} \subset \mathbb{R}^n$$

- $\mathcal{N}(f + g) = \mathcal{N}(f) \cup \mathcal{N}(g)$ and $\mathcal{N}(f \cdot g) = \mathcal{N}(f) + \mathcal{N}(g)$

(Minkowski sum of polytopes $\mathcal{P} + \mathcal{Q} = \{x + y \mid x \in \mathcal{P}, y \in \mathcal{Q}\}$

- normal fan $\mathcal{C}(\mathcal{N}(f))$: normal cones of all faces $\mathcal{C}_F(\mathcal{N}(f))$

$$\mathcal{C}_F(\mathcal{N}(f)) = \{w \in \mathbb{R}^n \mid F = F_w(\mathcal{N}(f))\}$$

$$F_w(\mathcal{N}(f)) = \{x \in \mathcal{N}(f) \mid (x - y) \cdot w \leq 0 \ \forall y \in \mathcal{N}(f)\}$$
• the set of parameters $U = (u_{ab})$, $V = (v_{ab})$ in tropicalization $\tau_{ijk}$ of $\Phi_{ijk}$ that determine the Viterbi sequence $(h_1, h_2, h_3)$ is the normal cone to a vertex of the Newton polygon $\mathcal{N}(\Phi_{ijk})$

• given observed data $(i, j, k)$ and hidden data $(h_1, h_2, h_3)$ the normal cones of $\mathcal{N}(\Phi_{ijk})$ give all parameter values for which $(h_1, h_2, h_3)$ is the most likely explanation for the observed $(i, j, k)$

• domains of linearity of the piecewise linear tropical $\tau_{ijk}$ are the cones in the normal fan $C_F(\mathcal{N}(\Phi_{ijk}))$; each maximal cone corresponds to one set of hidden data $(h_1, h_2, h_3)$ maximizing probability

$\tau_{ijk} = -\log \mathbb{P}((X_1, X_2, X_3) = (h_1, h_2, h_3) \mid (Y_1, Y_2, Y_3) = (i, j, k))$

• each vertex of the Newton polygon $\mathcal{N}(\Phi_{ijk})$ determines an inference function: $(i, j, k) \mapsto (h_1, h_2, h_3)$ that realize min $\tau_{ijk}$