Invariants in Phylogenetic Inference

MATH285K - Spring 2010 *Presenter: Neil Katuna*

Reference: [2]

1 The Problem

As we have learned throughout the course, it can be quite difficult to reconstruct the parameters of a phylogenetic model from the marginal distribution on the leaves of its tree. We have studied several estimators, but none have allowed us to precisely determine the shape or structure of such a tree without first understanding the nucleotide substitution process.

Using phylogenetic invariants, one can determine this tree structure without having to estimate any mutation model parameters. We say that a set of phylogenetic invariants, a phylogenetic ideal, fits a tree model if all polynomials in the ideal evaluate to zero on all possible distributions of bases for that tree. Determining whether mutations exhibit a certain tree structure becomes a simple problem of evaluating a set of polynomials against the observed frequencies of these mutations.

Remarkably, Evans and Speed [2] determine all possible phylogenetic invariants for trees with certain continuous-time mutation models: they require that the distribution of bases at the root be uniform, and that the infinitesimal generator matrix for the stochastic transition matrices at interior vertices admit a particular group structure. These assumptions enable some very clever Fourier analysis, analysis which allows for the explicit construction of the desired invariants.

2 The Set-Up

Consider a finite rooted tree $\mathbf{T}$ with root ρ. To each vertex $v \in \mathbf{T} \setminus \{\rho\}$ there corresponds a neighbor vertex $\sigma(v)$ closest to ρ in the usual graph-theoretic sense. Let $\mathbf{L}$ denote the set of leaves of $\mathbf{T}$.

Associate to each element $\ell \in \mathbf{L}$ an $\{A, G, C, T\}$-valued random variable Y_ℓ with a distribution induced by the following mutation model. Let π be a probability distribution on $\{A, G, C, T\}$, and let $P^{(v)}$ be the stochastic substitution matrix for the edge $(\sigma(v), v)$, $v \in \mathbf{V} \setminus \rho$. The natural probability distribution μ on

$\{A, G, C, T\}^{\mathbf{V}}$ is

$$\mu((i_v)_{v\in\mathbf{V}}) = \pi(i_\rho) \prod_{v\in\mathbf{V}\setminus\{\rho\}} P^{(v)}(i_{\sigma(v)}, i_v),$$

and the most obvious the marginal distribution on $\{Y_\ell\}_{\ell\in\mathbf{L}}$ is

$$P\left[(Y_\ell)_{\ell\in\mathbf{L}} = (i_\ell)_{\ell\in\mathbf{L}}\right] = \sum_{v\in\mathbf{V}\setminus\mathbf{L}} \sum_{i_v\in\{A,G,C,T\}} \mu((i_v)_{v\in\mathbf{V}\setminus\mathbf{L}}, (i_\ell)_{\ell\in\mathbf{L}}).$$

One could stop here and, using some extremely simple linear algebraic techniques, construct invariants for some basic trees whose substitution matrices are all equal. In fact, this was one of the first approaches to the invariant construction problem; see [1] for further reading.

Instead, we assume that our stochastic transition matrices arise from a continuous time Markov chain on the state space $\{A, G, C, T\}$. The infinitesimal generator matrix for our models will have the form

$$Q = \begin{pmatrix} -(\alpha+\beta+\gamma) & \alpha & \beta & \gamma \\ \alpha & -(\alpha+\beta+\gamma) & \gamma & \beta \\ \beta & \gamma & -(\alpha+\beta+\gamma) & \alpha \\ \gamma & \beta & \alpha & -(\alpha+\beta+\gamma) \end{pmatrix}.$$

That is, $e^{t_e Q}$ represents the transition matrix $P^{(v)}$ where e is the edge $(\sigma(v), v)$ and the time between mutations is given by t_e. One can show that this process corresponds to a discrete time Markov model with a rate $-(\alpha+\beta+\gamma)$ Poisson process on the inter-mutation times. We label rows and columns of the Q-matrix and its corresponding stochastic matrices in the order A, G, C, T.

You may have learned that Adenine and Guanine are chemically very similar, as are Cytosine and Guanine. The matrix Q expresses that there is a certain mutation transitivity between the elements in these pairs A, G and C, T. Such a Q with α, β, γ distinct corresponds to the *Kimura three-parameter model*. When $\beta = \gamma$, Q is said to generate the *Kimura two-parameter model*. When $\alpha = \beta = \gamma$, we obtain the *Jukes-Cantor model*.

Moreover, Q has the property that its $(i, j)^{\text{th}}$ entry depends upon $j - i$, where $i, j \in \{A, G, C, T\}$ and addition between bases is given by the following table:

+	A	G	C	T
A	A	G	C	T
G	G	A	T	C
C	C	T	A	G
T	T	C	G	A

One can easily check that this table specifies a group $\mathbb{G}$ isomorphic to the Klein four-group $\mathbb{Z}^2 \oplus \mathbb{Z}^2$, pairs of binary integers under addition modulo 2. One possible isomorphism is

$$A \leftrightarrow (0,0), \quad G \leftrightarrow (0,1), \quad C \leftrightarrow (1,0), \quad T \leftrightarrow (1,1).$$

Mutations along tree edges can be described as group actions. For the simple three-pronged tree with one root and three leaves, let Z_0, Z_1, Z_2, Z_3 be random variables on $\{A, G, C, T\} = \mathbb{G}$ such that

$$\begin{aligned} Y_1 &= Z_0 + Z_1, \\ Y_2 &= Z_0 + Z_2, \\ Y_3 &= Z_0 + Z_3, \end{aligned}$$

where Z_0 has the same distribution as the root and Z_1, Z_2, Z_3 have the appropriate transition distributions.

Moreover, there exists a function q which acts on elements of $\mathbb{G}$ such that $Q(i,j) = q(j^{-1}i)$, where $i, j \in \mathbb{G}$. This is precisely the condition that enables us to transform our continuous-time Markov process into a random walk on a group. After reviewing some Fourier analysis, we will show how this observation allows for the construction of polynomial invariants for a simple tree.

3 Fourier Analysis

Since $\mathbb{G}$ is abelian, any function on $\mathbb{G}$ has a corresponding Fourier transform, a function acting on the dual group $\hat{\mathbb{G}}$. Recall that the dual group consists of a collection of group homomorphisms mapping $\mathbb{G}$ into the unit circle in the complex plane. That is, χ is a character in $\hat{\mathbb{G}}$ if $\chi(g_1 + g_2) = \chi(g_1)\chi(g_2)$ for all $g_1, g_2 \in \mathbb{G}$. It can be shown that a group and its dual are isomorphic.

The dual group $\hat{\mathbb{G}}$ of our group $\mathbb{G} \cong \mathbb{Z}^2 \otimes \mathbb{Z}^2$ consists of characters $\{1, \phi, \psi, \phi\psi\}$, each its own inverse. The following table describes the action of $\hat{\mathbb{G}}$ on $\mathbb{G}$:

	A	G	C	T
$\langle \cdot, \cdot \rangle$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
1	1	1	1	1
ϕ	1	-1	1	-1
ψ	1	1	-1	-1
$\phi\psi$	1	-1	-1	1

Though the Fourier transform does not crop-up in what follows, it plays a role in Evans and Speed's classification of all linear invariants. Recall that for a function

$f : \mathbb{G} \to \mathbb{C}$,

$$\hat{f}(\chi) = \sum_{g\in\mathbb{G}} \langle g, \chi \rangle f(g), \qquad f(g) = \frac{1}{|\mathbb{G}|} \sum_{\chi\in\hat{\mathbb{G}}} \langle g, \chi \rangle \hat{f}(\chi).$$

It can be shown that convolutions become products in the transform domain. Finding Q^k—as is necessary for finding transition probabilities—is equivalent to convolving q with itself k times, a task easily accomplished using this Fourier analysis.

4 An Example

Consider the three-pronged tree of height 1 as described in the second section. Observe that

$$\begin{aligned}
&\mathbb{E}\left[\langle Y_1, \phi\rangle \langle Y_2, \psi\rangle \langle Y_3, \phi\psi\rangle\right] \\
&= \mathbb{E}\left[\langle Z_0 + Z_1, \phi\rangle \langle Z_0 + Z_2, \psi\rangle \langle Z_0 + Z_3, \phi\psi\rangle\right] \\
&= \mathbb{E}\left[\langle Z_0, \phi\rangle \langle Z_0, \psi\rangle \langle Z_0, \phi\psi\rangle \langle Z_1, \phi\rangle \langle Z_2, \psi\rangle \langle Z_3, \phi\psi\rangle\right] \\
&= \mathbb{E}\left[\langle Z_0 + Z_0, \phi\psi\rangle \langle Z_1, \phi\rangle \langle Z_2, \psi\rangle \langle Z_3, \phi\psi\rangle\right] \\
&= \mathbb{E}\left[\langle Z_1, \phi\rangle\right] \left[\langle Z_2, \psi\rangle\right] \left[\langle Z_3, \phi\psi\rangle\right],
\end{aligned}$$

where the second and third equalities follow from the fact that characters are homomorphisms. The last line may be deduced from the independence of the Z_i. Arguing similarly, one can show equalities for permutations of the indices of the Y_i and Z_i to conclude

$$\begin{aligned}
\Big(\mathbb{E}\left[\langle Y_1, \phi\rangle \langle Y_2, \psi\rangle \langle Y_3, \phi\psi\rangle\right] \mathbb{E}\left[\langle Y_1, \phi\psi\rangle \langle Y_2, \phi\rangle \langle Y_3, \psi\rangle\right] \\
\times \mathbb{E}\left[\langle Y_1, \psi\rangle \langle Y_2, \phi\psi\rangle \langle Y_3, \phi\rangle\right]\Big)^2 \\
- \prod_{1\le i<j\le 3} \prod_{\theta\in\{\phi,\psi,\phi\psi\}} \mathbb{E}\left[\langle Y_i, \theta\rangle \langle Y_j, \theta\rangle\right] = 0,
\end{aligned}$$

a polynomial in the observed distributions at the leaves. Denoting $t_{1,A} = P(Z_1 = A)$ and likewise for the other three leaves and four bases, we obtain a ninth degree polynomial in all twelve variables. Note that the expression above does not depend upon the distribution at the root.

Evans and Speed show that this argument generalizes for all trees. They characterize all polynomial invariants for these trees, provided that they have an infinitesimal Q matrix of the desired form and exhibit a uniform distribution at the root. For details, including the explicit form for these polynomials, see [2]. Allman and Rhodes [1] give an excellent summary of other invariant construction techniques. See their paper for other references.

References

[1] Allman, E.S. and J.A. Rhodes. Phylogenetic Invariants. Chapter 4 in *Reconstructing Evolution: New mathematical and computational advances,* eds. O. Glascuel and M. Steel, 2007.

[2] Evans, S.N. and T.P. Speed. Invariants of Some Probability Models Used in Phylogenetic Inference. *Annals of Statistics,* **21**(1): 355–377, 1993.