Metric Learning for Phylogenetic Invariants

Yuan Yao

Department of Mathematics
Stanford University

\(^a\) joint work with Nick Eriksson
Outline of the talk

- Model Selection by Invariants (structure learning)
- Running example: JC DNA model on Quartet Trees
- Metric Learning for Invariant Aggregation
- Ongoing Studies or Future Directions
Motivation: Model Structure Learning by Invariants

- Structure Learning: infer topology/structure of graphical models from observations
- Bayesian Average: a probabilistic model $p(\theta, s)$ ($\theta$ parameter, $s$ structure) and a priori distribution $p_0(\theta)$,

$$\max_s \int_\theta p(\theta, s)p_0(\theta)$$

challenge: super-exponential complexity for $s$.

- Attack: Friedman-Koller’03, order-space sampling
- Question: are model invariants helpful?
Running Example: JC DNA Model on Quartet Tree

Given DNA sequences of 4 taxa:

\[ s_i = \text{ATACGTACGTCAAA...} \quad , \quad i = 1, \ldots, 4, \]

generated from one of the following 3 trees:

i.e. in matching representation tree \([12][34], [13][24], \text{and} [14][23]\), (Note: structures are invariant to permutations within and between matched pairs.)

\( \text{identify the correct Tree structure among the three.} \)
Jukes-Cantor (Neyman) DNA Model

To each edge, associate a Markov chain with transition matrix

\[
M_e = \frac{1}{4} \begin{bmatrix}
1 + 3e^{-\frac{4}{3}t_e} & 1 - e^{-\frac{4}{3}t_e} & 1 - e^{-\frac{4}{3}t_e} & 1 - e^{-\frac{4}{3}t_e} \\
1 - e^{-\frac{4}{3}t_e} & 1 + 3e^{-\frac{4}{3}t_e} & 1 - e^{-\frac{4}{3}t_e} & 1 - e^{-\frac{4}{3}t_e} \\
1 - e^{-\frac{4}{3}t_e} & 1 - e^{-\frac{4}{3}t_e} & 1 + 3e^{-\frac{4}{3}t_e} & 1 - e^{-\frac{4}{3}t_e} \\
1 - e^{-\frac{4}{3}t_e} & 1 - e^{-\frac{4}{3}t_e} & 1 - e^{-\frac{4}{3}t_e} & 1 + 3e^{-\frac{4}{3}t_e}
\end{bmatrix}
\]

\(t_e\) is the length of the edge, i.e. \(t_e \in \{a, b\}\).
Setting $\theta_1 = 1 - e^{-4t_e/3}$ and $\theta_2 = 1 + 3e^{-4t_e/3}$, probability distribution $p_{ijkl}(\theta)$ ($i, j, k, l \in \{A, G, C, T\}$) is a polynomial map $\Theta \subseteq \mathbb{R}^2 \to \mathbb{R}^{256}$.

For each tree, there are 33 generators for the the ideal $< h : \mathbb{R}^{256} \to \mathbb{R} >=: I_p$, s.t.

$$h(p_{AAAA}(\theta), \ldots, p_{TTTT}(\theta)) = 0, \quad \forall \theta \in \Theta$$

See "www.math.tamu.edu/~lgp/small-trees/"
The 33 generators depend on the order of leaves.

Tree [12][34], [13][24], [14][23] are invariant under 3 permutation groups on the order space: \(<(12)\>, \(<(34)\>, \text{ and } <(13)(24)> (\simeq \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2)\).

These permutations induce 16 more invariants, such that the \(\mathbb{Z}_2^3\)-action orbits are closed in this 49-invariant space.

e.g. let \(\pi_1 = (12)\), then \(\pi_1\) induces a signed permutation on the 49 invariants (e.g. \(h_{10} \mapsto -h_{10}\)), but not on 33 invariants.
for each structure, evaluate the invariants associated with it, e.g. in JC-model

\[ h^s(\hat{\phi}) = (h_1^s(\hat{\phi}), \ldots, h_{49}^s(\hat{\phi})) \]

where \( s \in \{ [12][34], [13][24], [14][23] \} \), and \( \hat{\phi} \) is empirical distribution

choose the structure with smallest “norm”: \( \min_s \| h^s \|_* \)

\( l_1 \)-norm (Casanellas et al. 2006): \( \| h \|_1 = \sum_i |h_i| \)

\( l_0 \)-norm: \( \| h \|_0 = \#\{ i : |h_i| > 0 \} \)

\( l_2 \)-norm: \( \| h \|_2 = \sqrt{\sum_i |h_i|^2} \)

Mahalanobis-norm: \( \| h \|_A = \sqrt{h^T A h} \), \( A \succeq 0 \)
The best is $h_{10}$, (comparable to Neighbor-Joining).
Invariant Prediction Rates on Parametric Plane

---

Percent Correct: Invariant 10

Percent Correct: Invariant 14

Percent Correct: Invariant 19

Percent Correct: Invariant 30

---

Metric Learning for Phylogenetic Invariants – p. 10/?
Note: The best orbit-4 just contains $\{\pm h_{10}\}$. 
Aggregation by Metric Learning

Algorithm I:

\[
\begin{align*}
\min & \quad \sum_j \xi_j \\
\text{s.t.} & \quad \|v_j^{s_0}\|_A^2 \leq \|v_j^s\|_A^2 + \xi_j, \quad s \neq s_0 \\
& \quad \xi_j \geq 0 \\
& \quad \pi_k A = A \pi_k, \quad k = 0, \ldots, 7
\end{align*}
\]

for all \( j \) examples and \( \pi_k \) are permutation matrices (including identity) for the \( \mathbb{Z}_2^3 \)-action on \( (1, \ldots, 49) \).

\( \text{SDP: } 49 \times 49 \text{ dense SD, } 2k \text{ linear constraints} \)
Algorithm II:

\[
\min \sum_j \xi_j + \lambda \text{trace}(A)
\]

s.t. \[\|v_j^{\mathbf{s}_0}\|_A^2 + \gamma \leq \|v_j^{\mathbf{s}}\|_A^2 + \xi_j\]

\[\xi_j \geq 0, \quad \pi_k A = A \pi_k.\]

- \(\gamma\): margin parameter, e.g. \(\gamma = 0.05\).
- \(\lambda\): regularization parameter, keep \(A\) as low rank as possible, e.g. \(\lambda = 0.0001\), prevent \(A\) being too large.
Explanations of Algorithm I and II

In stead of searching the whole ideal $I_p$, the SDP-based algorithms search in the convex cone,

$$\sum_{i,j} A_{ij} h_i h_j, \quad A \succeq 0$$

some “good” invariants which are closer to the origin for the generative tree structure/topology than other trees.
The top 15 eigenvectors contribute most.
<table>
<thead>
<tr>
<th></th>
<th>$l_2$</th>
<th>$l_1$</th>
<th>$NJ$</th>
<th>A-norm</th>
<th>$NJ_\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.724</td>
<td>.800</td>
<td>.844</td>
<td>.929</td>
<td>.931</td>
<td></td>
</tr>
</tbody>
</table>

- NJ uses typical method in literature with $\infty$-distance being counted as “failure”
- $NJ_\infty$ is a modified version with a careful treatment on $\infty$-distance
- Typically $l_0$-norm (voting) is slightly worse than $l_1$, better than $l_2$. 
Comparisons on a-b plane

Percent correct: NJ (infinite distances)

Percent correct: A-norm

Percent correct: NJ

Percent correct: L1-norm

Percent correct: L2-norm
Ongoing/future works...

- Full Kimura DNA model: $11k$ extended invariants, $1k$ orbits
- Large-scale dense SDP is intractable, use diagonal weights with LP or block structure.
- Understanding the geometry of invariants, e.g. #10
  #10 is the equality test of four-point condition
  NJ is the inequality test of four-point condition
- Developing *sign invariants* for inequality test?