Theory of SVDQuartets

Jed Chou

CS598 AGB

April 13, 2015

Overview

1. Motivation
2. Background
3. Main Theorems
4. Applying the Theory
5. Further Research

Competing Approaches

Two competing approaches to species tree inference:

- Summary methods: estimate a tree on each gene alignment then combine gene trees.
  - MP-est
  - NJst
  - ASTRAL-II
- Concatenation: concatenate all gene alignments and estimate species tree on resulting supermatrix.
  - CA-ML (RAxML)

Challenges for Summary and Concatenation

Challenges for summary methods:

- Long alignments are unreliable due to recombination
- Gene trees estimated on short alignments have estimation error
- Summary methods are sensitive to gene tree estimation error

Challenges for concatenation:

- Ignores incomplete lineage sorting
- Can be statistically inconsistent under the coalescent model (Roch and Steel, 2014)

κ-state GTR model

The κ-state analytic General Time-Reversible model $C_{GTR(\kappa)}$ has parameters:

- species tree topology S
- vector of speciation times $\tau = (\tau_1, \tau_2, ..., \tau_{n-1})$
- effective population size θ
- substitution matrix P

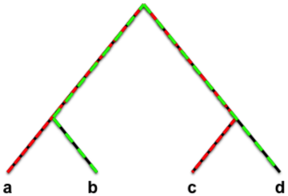
Notation

- parameter space $U_S \subseteq \mathbb{R}^M$
- probability simplex $\Delta^{\kappa^4-1} = \{(p_1, ..., p_{\kappa^4}) \in \mathbb{R}^{\kappa^4} | \sum_{i=1}^{\kappa^4} p_i = 1 \text{ and } p_i \geq 0\}$
- parameterization map $\psi_S : U_S \to \Delta^{\kappa^4-1}$

Splits

Definition

A **split** $L_1|L_2$ of a set L of taxa is a bipartition of L into two non-overlapping subsets L_1 and L_2; it is **valid** for a tree T if it is induced by an edge, i.e. if $L_1|L_2 \in C(T)$. Alternatively, $L_1|L_2$ is valid if the subtrees on L_1 and L_2 do not intersect.



ac|bd is not a valid split

Flattenings

Definition

Let $L_1|L_2$ be a split and $P \in \psi_S(U_S)$, a $\kappa \times \ldots \kappa$ tensor. A **flattening** of P, $Flat_{L_1|L_2}(P)$, is a $\kappa^{|L_1|} \times \kappa^{|L_2|}$ matrix whose rows are indexed by possible states for the leaves in L_1 and columns by possible states in L_2.

Example

For a 4-taxon tree and split $ad|bc$

$$Flat_{ad|bc}(P^*_{(S,\tau)}) = \begin{pmatrix} p^*_{AAAA} & p^*_{AACA} & \cdots & p^*_{ATTA} \\ p^*_{AAAC} & p^*_{AACC} & \cdots & p^*_{ATTC} \\ \vdots & \vdots & \ddots & \vdots \\ p^*_{TAAT} & p^*_{TACT} & \cdots & p^*_{TTTT} \end{pmatrix}$$

Invariants

Definition

An **invariant** is a function in the site pattern probabilities that vanishes when evaluated on any distribution arising from the model.

Examples

- For any species tree (S, τ) on n taxa under a κ-state GTR model, $\sum_{i_j \in [\kappa]} p^*_{i_1 \dots i_n | (S,\tau)} - 1 = 0$.
- For a species tree $S = ((a, (b, (c, d)))$, $p^*_{**ij|(S,\tau)} - p^*_{**ji|(S,\tau)} = 0$.

A Lemma

Definition

Let $R = \{f_1, ..., f_n\}$ be a set of analytic functions on a connected, open set $D \subseteq \mathbb{R}^m$. The **analytic variety** $V(R)$ is the set of common zeros of $f_1, ..., f_n$:

$$V(R) = \{x \in D | f_i(x) = 0, 1 \leq i \leq n\}$$

Definition

Fix a coalescent phylogenetic model $\psi_S : U_S \to \Delta^{\kappa^4 - 1}$. An analytic function f is called a **coalescent phylogenetic invariant** if $f(x) = 0$ for all $x \in \psi_S(U_S)$.

Main Theorems

Let (S, τ) be a 4-taxon symmetric or asymmetric species tree with cherry (c, d). Identify the parameter space U_S with a full dimensional subset of $\mathbb{R}^M$.

Theorem 1

If $L_1|L_2$ is a valid split for S, then for all distributions $P^*_{(S,\tau)}$,

$$rank(Flat_{L_1|L_2}(P^*_{(S,\tau)})) \leq \binom{\kappa+1}{2}$$

Theorem 2

If $L_1|L_2$ is not a valid split for S, then generically

$$rank(Flat_{L_1|L_2}(P^*_{(S,\tau)})) > \binom{\kappa+1}{2}$$

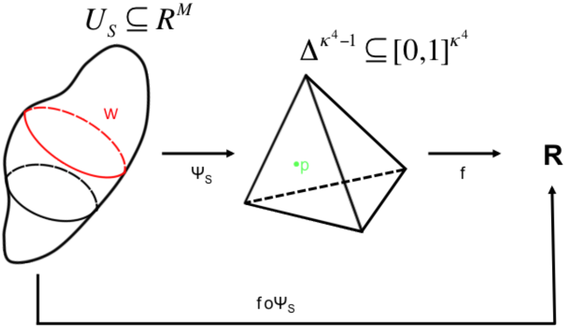
Sketch of Proof Theorem 1

1. Suppose $L_1|L_2$ is a valid split for S.
2. Both symmetric and asymmetric 4-taxon trees have cherry (c,d), so columns in $Flat_{L_1|L_2}(P^*_{(S,\tau)})$ labeled by the cd-indices kl and lk are identical for $k \neq l \in [\kappa]$.
3. There are $\binom{\kappa}{2}$ such pairs, so $rank(Flat_{L_1|L_2}(P^*_{(S,\tau)})) \leq \kappa^2 - \binom{\kappa}{2} = \binom{\kappa+1}{2}$.

Sketch of Proof Theorem 2

1. Suppose $L_1|L_2$ is not a valid split for S.
2. Let $X_{ac|bd}$ and $X_{ad|bc}$ be the sets of all $(\binom{\kappa+1}{2}) + 1)$-minors of $Flat_{ac|bd}$ and $Flat_{ad|bc}$ respectively and let $V_{ac|bd} = V(X_{ac|bd})$, $V_{ad|bc} = V(X_{ad|bc})$.
3. Note: The rank of a matrix is the maximal order of a non-zero minor.
4. Let $W = V(\{f \circ \psi_S\})$ where $f : \Delta^{\kappa^4-1} \to \mathbb{R}$ are analytic functions that vanish on $V_{ac|bd} \cup V_{ad|bc}$.
5. W is an analytic subvariety of $U_S \subseteq \mathbb{R}^M$ and in fact it is **proper**.
6. So $dim(W) < dim(U_S)$ and W has measure 0 in U_S.

Sketch of Proof Theorem 2



Frobenius Norm

Definition

The **Frobenius Norm** of an $n \times m$ matrix A is

$$||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$$

Singular Values

The **singular values** of a square matrix A are the square roots of the eigenvalues of A^*A where A^* is the conjugate transpose of A.

Some Linear Algebra

Theorem

$$||A||_F = \sqrt{\sum_{i=1}^{p} \sigma_i^2}$$

where $\sigma_1 \geq \sigma_2 \geq ...\sigma_p$ are the singular values of A and $p = \min\{m, n\}$.

Theorem (Eckart-Young, 1936)

For any $k \leq p$,

$$\min_{rank(B)=k} ||A - B||_F = \sqrt{\sum_{i=k+1}^{p} \sigma_i^2}$$

SVD score

Definition

The **SVD score** of a split $L_1|L_2$ for a 4-state GTR model species tree on 4 taxa is

$$SVD(L_1|L_2) = \sqrt{\sum_{i=11}^{16} \sigma_i^2}$$

Consequences

- For a valid split $L_1|L_2$, $rank(Flat_{L_1|L_2}(P^*_{(S,\tau)})) \leq \binom{5}{2}$, so $\sigma_{11} = ... = \sigma_{16} = 0$ and $SVD(L_1|L_2) = 0$.
- For a non-valid split $L_1|L_2$, $rank(Flat_{L_1|L_2}(P^*_{(S,\tau)})) > \binom{5}{2}$, so $\sigma_{11} \neq 0$ and $SVD(L_1|L_2) > 0$.

Inferring the Species Tree

To infer the species tree on n taxa given a collection of gene alignments:

1. For a set $L = \{a, b, c, d\}$ in S, estimate the flattening matrices of site pattern probabilities for the 3 possible splits by picking a single site from each gene alignment and counting the frequencies of all 256 site patterns ($AAAA$, $AAAT$, etc.) among the selected sites.
2. Compute the SVD score of each split $L_1|L_2$ from these flattening matrices.
3. Pick the split with the SVD score closest to 0 and return the associated quartet tree. Do this for every set of 4 taxa in S to get a collection of quartet trees.
4. Combine all quartet trees with a quartet method.

Further Research

We are currently studying SVDQuartets combined with quartet methods QMC and wQMC. Some directions for further research include:

- Rigorous experiments under a variety of conditions
- Comparison to summary methods and concatention
- Quartet subsampling to reduce running time
- Better approaches to select sites from each gene alignment
- Statistical consistency in the presence of HGT or ILS+HGT

References

Julia Chifman and Laura Kubatko (2014)
Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes
arXiv:1406.4811v1

Julia Chifman and Laura Kubatko (2014)
Quartet Inference from SNP Data Under the Coalescent Model
Bioinformatics Advance Access