Variational Bayesian Phylogenetic Inference

Discussant: Derek Hansen

Stats 701
University of Michigan Department of Statistics

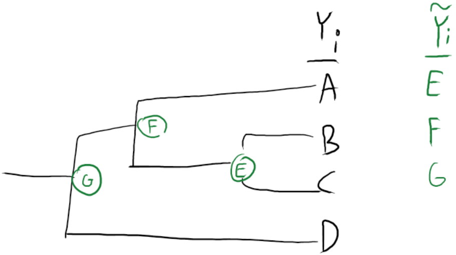
November 17, 2020

Inference on Phylogenetic Trees

- A phyologenetic tree has tree topology τ and associated branch lengths q.
- We observe the traits of M individuals/specices at N difference sites. This is expressed as the data matrix $Y \in \Omega^{N \times M}$, where Ω is the state-space of different possible traits.
- E.g. $\Omega = \{A, G, C, T\}$.
- The ith column of Y, Y_i, are the observed traits in each individual at site i.
- Given a tree (τ, q), we want to calculate the likelihood $\mathcal{P}(Y|\tau, q)$. We assume that the traits are conditionally independent, i.e.

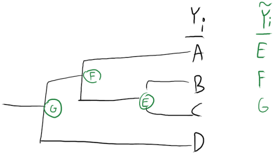
$$\mathcal{P}(Y|\tau, q) = \prod_{i=1}^{M} \mathcal{P}(Y_i|\tau, q)$$

Inference on Phylogenetic Trees



- There is a continuous time markov process P where $P_{ij}(q)$ is the probability of starting at state i and ending at state j for an edge-length of q.
- Let p_i be the limiting distribution of P_{ij} as $q \to \infty$.
- Conditional on the observed traits Y_i and the traits of the unobserved ancestors $\tilde{Y}_i$, we can calculate the probability of this graph.

Inference on Phylogenetic Trees



- Let $a^i = Y_i, \tilde{Y}_i$ be the states of all nodes in graph (their notation)
- Let a^i_0 be the state of the root node. (Its probability is the stationary distribution of $P_{i,j}$).

$$\begin{aligned}\mathcal{P}(Y_i|\tau,q) &= \sum_{\tilde{Y}_i} \mathcal{P}(\underbrace{Y_i, \tilde{Y}_i}_{\triangleq a^i}|\tau,q) \\ &= \sum_{a_i} \mathcal{P}(a^i|\tau,q) 1[a^i_{\text{leaf}} = Y_i] \\ &= \sum_{a_i} p_{a^i_0} \prod_{(u,v)\in E(\tau)} P_{a^i_u, a^i_v}(q_{u,v}) 1[a^i_{\text{leaf}} = Y_i]\end{aligned}$$

Inference on Phylogenetic Trees

- $\mathcal{P}(Y_i|\tau, q)$ can be calculated efficiently with established methods in the literature.
- With a prior $\mathcal{P}(\tau, q)$, the posterior density is proportional to:

$$\mathcal{P}(\tau, q|Y) \propto \prod_{i=1}^{M} \mathcal{P}(Y_i|\tau, q)\mathcal{P}(\tau, q)$$

- The authors don't mention this, but presumably the prior for τ, q could be a generative process like those we have discussed in class (e.g. birth-death process).
- In practice, the prior they use in experiments is a uniform prior over a discrete support found via bootstrap.

Subsplit Bayesian Networks

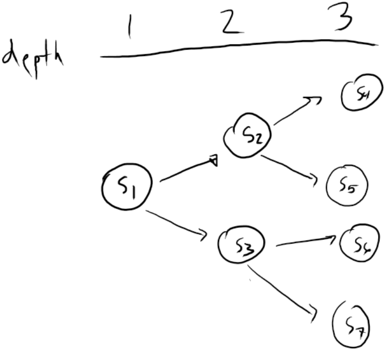
- A significant challenge to inference over phylogenetic trees is that the topology of the tree is a random variable.
- In a previous paper, the authors developed an equivalent representation of phylogenetic trees called "Subsplit Bayesian Networks" (SBN).
- An SBN has a fixed topology; it's a binary tree of depth N-1.
- Each node of an SBN is not an individual, but a representation of a split in the population.
- Before formally describing what these networks are, we'll need some definitions.

Subsplit Bayesian Networks - Definitions

- Let $\mathcal{X}$ be the set of N individuals. For example $\mathcal{X} = \{A, B, C, D\}$
- A **clade** is a non-empty subset of $\mathcal{X}$ (i.e. $X \in 2^{\mathcal{X}} \setminus \varnothing$).
  - $X = \{A, B\}$
- We can equip clades with an ordering $\succ$ which is lexigraphical, meaning that we order clades as if they were words.
  - $\{A, C\} \succ \{B, D\}$.
- A **subsplit** of clade C is a pair of disjoint clades $S = (W, Z)$ such that:
  1. $W \cup Z = C$
  2. $W \succ Z$
- The second condition ensures that every subsplit has a unique representation.
- $\{A, C, D\}, \{B\}$ is a subsplit of $\{A, B, C, D\}$.

Subsplit Bayesian Networks

- Consider $\mathcal{X}$ of size N.
- Let $\mathcal{B}^*_{\mathcal{X}}$ be a complete binary tree of depth $N - 1$.



$$\mathcal{B}^*_{\mathcal{X}}$$

$$\mathcal{X} = \{A, B, C, D\}$$

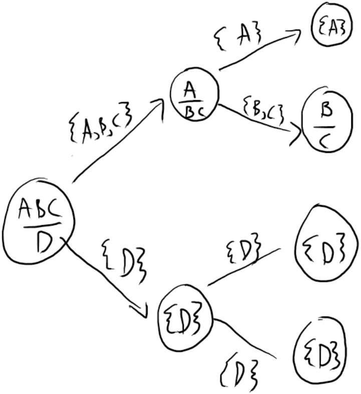
$$N = |\mathcal{X}| = 4$$

Subsplit Bayesian Networks

- Every node in $\mathcal{B}^*_{\mathcal{X}}$ is one of two things:
  - A clade subsplit (X, Y)
  - A singleton clade $\{A\}$
- The root node of $\mathcal{B}^*_{\mathcal{X}}$, S_0, is a subsplit of $\mathcal{X}$.
- All child nodes S_i, $i \geq 1$, inherit a clade from their parent:
  - If the parent is a subsplit (X, Y), then the child inherits X if it is the first child and Y if it is second.
  - If the parent is a singleton clade $\{A\}$, the child inherits $\{A\}$.
- The value of node S_i depends on the clade C_i it inherits.
  - If $|C_i| > 1$, S_i is a subsplit of C_i.
  - If $|C_i| = 1$, $S_i = C_i$.

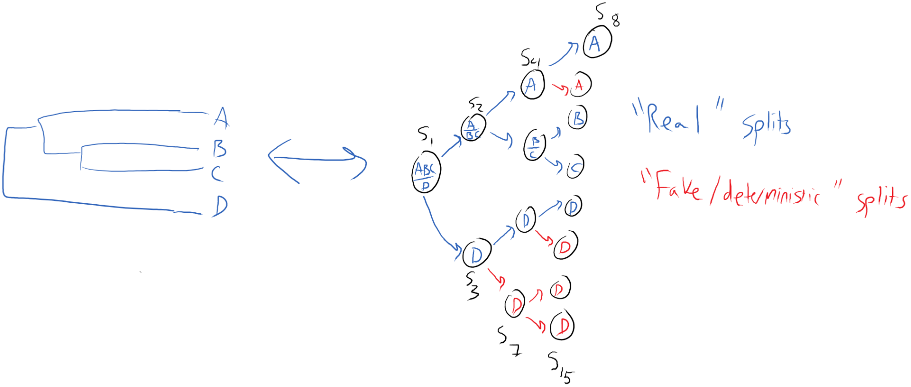
Subsplit Bayesian Networks

- An example of one particular choice of $S_1, S_2, \ldots, S_7$ on the tree $\mathcal{B}^*_{\mathcal{X}}$



Subsplit Bayesian Networks

- Any phylogenetic tree of the individuals in $\mathcal{X}$ can be written as a particular choice of node values on $\mathcal{B}^*_{\mathcal{X}}$.
- The original phylogenetic tree can be recovered by picking one terminal edge per individual in the expanded tree.

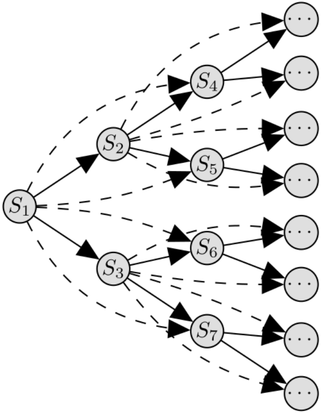


Subsplit Bayesian Networks

- With $\mathcal{B}_{\mathcal{X}}^*$ properly defined, we are ready to define a Subsplit Bayesian Network.
- A subsplit Bayesian network $\mathcal{B}_{\mathcal{X}}$ (SBN) is a graph of random nodes such that there exists a $\mathcal{B}_{\mathcal{X}}^*$ where
  - The nodes of $\mathcal{B}_{\mathcal{X}}$ are the same as $\mathcal{B}_{\mathcal{X}}^*$.
  - The edges of $\mathcal{B}_{\mathcal{X}}$ contain those of $\mathcal{B}_{\mathcal{X}}^*$
- $\mathcal{B}_{\mathcal{X}}^*$ itself is the simplest SBN
- A subsplit Bayesian network allows for the distribution of subsplits to have arbitrary dependencies, so long as they do not violate the graph structure.

Subsplit Bayesian Networks

- An example SBN $B_{\mathcal{X}}$, where each node is conditionally dependent on both its parents and "grandparents"



Subsplit Bayesian Networks

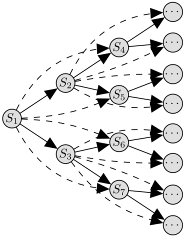
- If T is the random topology of the phylogenetic tree and τ a particular realization, we write the probability in terms of $S_1, \ldots, S_{2^{N-1}-1}$, where $\pi(i)$ are the dependencies of S_i given by the edges of $\mathcal{B}_{\mathcal{X}}$.

$$P(T = \tau) = p(S_1 = s_1) \prod_{i>1} p(S_i = s_i | S_{j \in \pi(i)})$$

Subsplit Bayesian Networks

What are the advantages of this structure?

- More flexible than a "clade"-conditional model
- Unrooted trees can be accounted for by integrating out S_1.
- The topology of the SBN is fixed



Variational Phylogenetic Inference via SBNs

- Recall that our goal is to infer the posterior distribution $\mathcal{P}(\tau, q|Y)$.
- Variational Inference approaches this problem by attempting to find a variational distribution Q_ϕ from a family $\mathcal{Q}$ parameterized by ϕ which minimizes the KL-divergence between them.

$$\begin{aligned}\hat{\phi} &= \text{argmin}_\phi \text{KL}(Q_\phi(\tau, q)||\mathcal{P}(\tau, q|Y)) \\ &= \text{argmin}_\phi E_{Q_\phi(t,q)}\left(\log \frac{\mathcal{P}(\tau, q|Y)}{Q_\phi(t, q)}\right)\end{aligned}$$

- This formulation is not directly usable, since $\mathcal{P}(\tau, q|Y)$ is intractable.

Variational Phylogenetic Inference via SBNs

- We can calculate the evidence lower bound (ELBO) instead; we write this as $L(\phi)$.
- Maximizing the ELBO with respect to ϕ is equivalent to minimizing the KL term on the previous slide.

$$\begin{aligned} L(\phi) &\triangleq E_{Q_\phi(\tau,q)}(\log \mathcal{P}(\tau, q, Y) - \log Q_\phi(\tau, q)) \\ &= E_{Q_\phi(\tau,q)}(\log \mathcal{P}(Y) + \log \mathcal{P}(\tau, q|Y) - \log Q_\phi(\tau, q)) \\ &= \log \mathcal{P}(Y) + E_{Q_\phi(\tau,q)}(\log \mathcal{P}(\tau, q|Y) - \log Q_\phi(\tau, q)) \\ &= \underbrace{\log \mathcal{P}(Y)}_{\text{Evidence}} + \underbrace{\mathrm{KL}(Q_\phi(\tau, q)||\mathcal{P}(\tau, q|Y))}_{\text{Non-negative divergence}} \end{aligned}$$

- The integrand of the ELBO is tractable, although the expectation still is not. Hence we use stochastic optimization.

Variational Phylogenetic Inference via SBNs

- An unbiased estimate of the ELBO $L(\phi)$ is $\log \mathcal{P}(\tau^i, q^i, Y) - \log Q_\phi(\tau^i, q^i)$ for $\tau^i, q^i \sim Q_\phi(\cdot, \cdot)$.
- A more accurate lower bound to the evidence can be achieved by using multiple iid samples and weighting by their importance:

$$L^K(\phi) \triangleq E_{Q_\phi(\tau^i, q^i)} \left(\log \left(\frac{1}{K} \sum_{i=1}^{K} \frac{\mathcal{P}(\tau^i, q^i, Y)}{Q_\phi(\tau^i, q^i)} \right) \right)$$

- An unbiased estimate of $L^K(\phi)$ is $\log \left(\frac{1}{K} \sum_{i=1}^{K} \frac{\mathcal{P}(\tau^i, q^i, Y)}{Q_\phi(\tau^i, q^i)} \right)$.
- $L^1(\phi) = L(\phi)$, $\lim L^K(\phi) \to \mathcal{P}(Y)$.

Variational Phylogenetic Inference via SBNs

- The variational distribution $Q_\phi(\tau, q)$ is parameterized as $Q_\phi(\tau)Q_\phi(q|\tau)$.
- The variational distribution of τ is based on the SBN parameterization.

$$Q_\phi(\tau) = Q_\phi(S_1, \dots) = Q_\phi(S_1 = s_i) \prod_{i>1} Q_\phi(S_i = s_i | s_{j\in\pi(i)} = s_{j\in\pi(i)})$$

Variational Phylogenetic Inference via SBNs

- For every possible parent-child subsplit pair, there is an associated parameter $\phi_{s|t}$.
- This problem is made tractable by restricting the support of possible pairs over high-probability areas of the posterior (found via bootstrap)
- $\mathbb{S}_1$ and $\mathbb{S}_i(s_{j\in\pi(i)})$ represent these restricted supports.

$$Q_\phi(S_1 = s_1) = \frac{\exp(\phi_{s_1})}{\sum_{s_1^r \in \mathbb{S}_1} \exp\left(\phi_{s_1^r}\right)}$$

$$Q_\phi(S_i = s_i | S_{j\in\pi(i)} = s_{j\in\pi(i)}) = \frac{\exp(\phi(s_i | s_{j\in\pi(i)}))}{\sum_{s_i^r \in \mathbb{S}_i(s_{j\in\pi(i)})} \exp(\phi(s_i^r | s_{j\in\pi(i)}))}$$

Variational Phylogenetic Inference via SBNs

- Conditional on the topology τ, branch lengths v have a log-Normal distribution.

$$Q_\phi(q|\tau) = \prod_{e \in E(\tau)} f_{\mathrm{LN}}(q_e; \mu_\phi(D(e,\tau)), \sigma_\phi(D(e,\tau))$$

- $e = (A_i \rightarrow A_j)$ is an edge between two individuals on the phylogenetic tree topology τ (not the SBN!)
- Rather than parametrize every possible edge in every possible tree τ, μ_ϕ and σ_ϕ depend only on the subsplit of individuals $\mathcal{X}$ induced by the edge $D(e,\tau) \subseteq \mathcal{X}$.
- $D(e,\tau) = \{x \in \mathcal{X} : \exists \text{ path from } A_j \text{ to } x\}$
- In their first "Simple Independent Approximation", they parameterize each function value as $\psi^{\mu}_{e/\tau} \equiv \mu_\phi(D(e,t))$

Variational Phylogenetic Inference via SBNs

- The authors also consider a formulation called “Primary Subsplit Pair” (PSP) where the branch length can also depend on splits which occur later in the subtree.
- Let $e = (S_i \to S_j)$ and $e^C \triangleq (S_i \to S_k)$ be the other edge connected to the same parent. If S_j or S_j are not leaf nodes, define e_1 and e_1^C to be their first descendent edges respectively.

$$\begin{aligned}\tilde{\mu}_\phi(e, t) = \psi^{\mu}_{e/\tau} \\ &+ 1[S_j \text{ not leaf}](\psi_{D(e_1,\tau)}) \\ &+ 1[S_j \text{ not leaf}](\psi_{D(e_1^c,\tau)})\end{aligned}$$

- In other words, the branch length of e can depend not just on the clade it is splitting, but also on the actual split that occured underneath it or in its complement edge.

Variational Phylogenetic Inference via SBNs - VIMCO

- $L^K(\phi)$ is optimized using stochastic gradient descent (SGD).
- In general, $\nabla_\phi L^K(\phi) \neq E_{Q_\phi(\tau^i, q^i)} \left(\nabla_\phi \log \frac{1}{K} \sum_{i=1}^K \frac{\mathcal{P}(\tau^i, q^i, Y)}{Q_\phi(\tau^i, q^i)} \right)$.
- The branch lengths q can be reparameterized so that the expectation does not depend on the ψ parameters; this is the reparameterization trick.

$$\log q_e = \mu(e, \tau) + \epsilon_e \sigma(e, \tau)$$
$$\epsilon_e \sim N(0, 1)$$

- Then the gradient can be passed inside the expectation.
- The reparameterization trick doesn't work for discrete variables like τ, and the naive estimator is very noisy.
- The authors use the VIMCO gradient estimator for parameters associated with τ; this reduces the variance by leveraging the multiple samples in L^K.

Variational Phylogenetic Inference via SBNs - RWS

- The authors also consider an alternative objective which seeks to minimize the KL divergence in the opposite direction.
- This is equivalent to maximizing the likelihood of the variational approximation over samples from the true posterior.

$$\begin{aligned}\hat{\phi} &= \text{argmin}_{\phi}\text{KL}\left(\mathcal{P}(\tau, q|Y)||Q_{\phi}(\tau, q)\right)\\ &= \text{argmax}_{\phi}E_{\mathcal{P}(\tau, q|Y)}\log Q_{\phi}(\tau, q)\end{aligned}$$

Experiments

- UFBoot is used to bootstrap the support of possible trees.
- Two different optimization schemes
  - ELBO objective with VIMCO gradient estimator
  - Reversed KL objective with RWS gradient estimator
- Two different sample sizes in calculating bounds: $20, 50$
- Tested with and without the Primary Subsplit Pair (PSP) modification for branch length q variational distribution.

Experiments - Simulated Trees

- Sample space of τ is all unrooted phylogenetic trees with 8 leaves (sample size 10395)
- Branch lengths are not considered
- A "posterior" $p_0(\tau)$ is generated by drawing from a symmetric Dirichlet distribution with concentration parameter $\beta = 0.008$.
- The model log evidence is $\log \mathcal{P}(Y) = 0$ (since there is no data).

Experiments - Simulated Trees

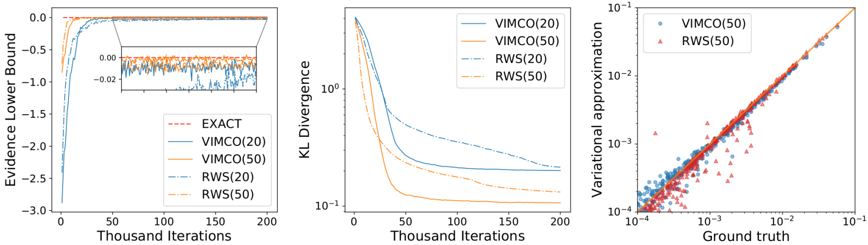


Figure 3: Comparison of multi-sample objective on approximating a challenging distribution over unrooted phylogenetic trees with 8 leaves using VIMCO and RWS gradient estimators. **Left:** Evidence lower bound. **Middle:** KL divergence. **Right:** Variational approximations vs ground truth probabilities. The number in brackets specifies the number of samples used in the training objective.

- Focusing on KL, the VIMCO with 50 samples in training objective learns the best variational distribution.
- From the right plot, VIMCO slightly underestimates higher-probability areas and over-estimates lower-probability areas. RWS underestimates lower-probability areas.

Experiments - Real Data

- Goal is to learn posterior distribution of unrooted phylogenetic trees on 8 real datasets.
- $\mathcal{P}(\tau)$ is uniform over support taken from UFBoot.
- Each branch length q has an iid exponential prior with $\lambda = 10$.
- Ground truth is calculated from an extremely long MCMC run (10 billion iterations) using MrBayes.
- Also compare performance to MrBayes with less iterations

Experiments - Real Data

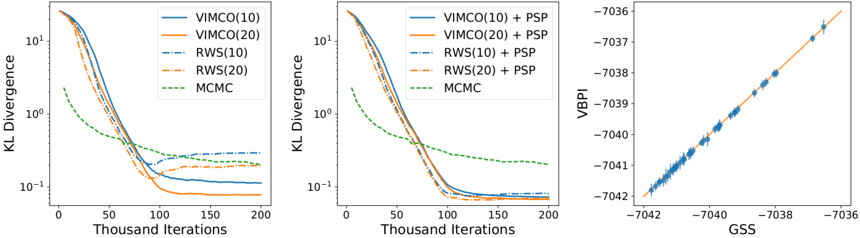


Figure 4: Performance on DS1. **Left:** KL divergence for methods that use the simple split-based parameterization for the branch length distributions. **Middle:** KL divergence for methods that use PSP. **Right:** Per-tree marginal likelihood estimation (in nats): VBPI vs GSS. The number in brackets specifies the number of samples used in the training objective. MCMC results are averaged over 10 independent runs. The results for VBPI were obtained using 1000 samples and the error bar shows one standard deviation over 100 independent runs.

- The variational methods converge much faster than MCMC
- Adding PSP modification for branch length led to lower KL.
- Log-likelihood estimates compare well to state-of-art GSS algorithm

Conclusion

- VBPI enables Variational Inference by utilizing Subsplit Bayesian Networks (SBNs).
- The random topology of a phylogenetic tree, whose nodes are individuals, is mapped to a fixed-topology SBN whose nodes are subsplits of clades
- This enables learning of a variational distribution that can quickly sample trees.
- The resulting variational methods show promising performance in both simulations and in real data.