Tree reconstruction from statistical perspectives

Lam Si Tung Ho

Department of Mathematics and Statistics

DALHOUSIE UNIVERSITY

Contact: Lam.Ho@dal.ca

Tree reconstruction

Statistical inference

- **Data:** $(Y_i)_{i=1}^n$
- **Model:** $(Y_i)_{i=1}^n$ follow a distribution $\mathcal{P}_{\theta^*}$ where $\theta^* \in \Theta \subset \mathbb{R}^d$
- **Estimation method:** approximate θ^*

Tree reconstruction

- **Data:** sequences
- **Model:** a substitution model along a true tree $\mathbb{T}$
- **Reconstruction method:** Maximum likelihood, Bayesian, ...

However, $\mathbb{T} \notin \mathbb{R}^d$, and the tree topology is a discrete object

Applying statistical theory is not straightforward

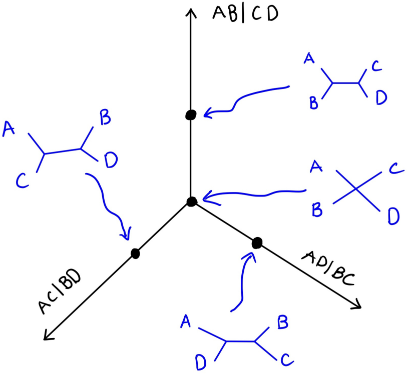
Standard statistical theory: $\hat{\theta}_{\text{MLE}} \to \theta^*$

- Model identification
- Parameter space Θ is compact
- The log likelihood function $\ell(\theta \mid Y)$ is continuous in θ for almost all Y
- $E\left[\sup_\theta |\ell(\theta \mid Y)|\right] < \infty$

"*Several workers ... concerned that the discrete, unordered nature of a tree topology variable prevents it from being the sort of parameter required ...*"

(Rogers, 2001)

Continuous tree space (Billera-Holmes-Vogtmann)



Embedding

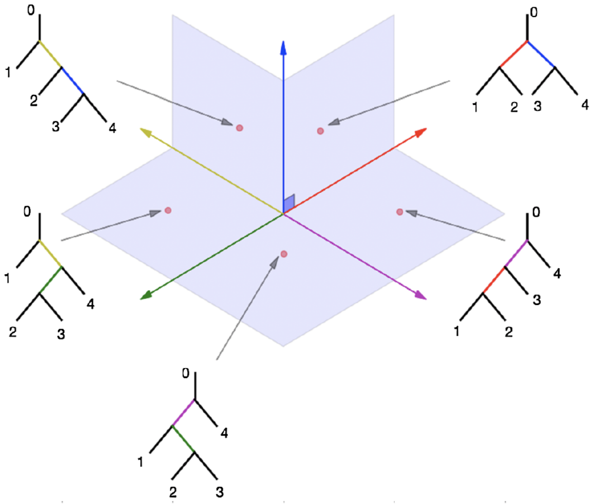
$$\mathbb{T} \hookrightarrow \sum_{s \in \mathcal{S}} e_s \zeta_s$$

- $\mathcal{S}$: set of all tree splits
- e_s: edge length
- ζ_s: basis vector

Distance:

- Branch score distance
- Geodesic distance

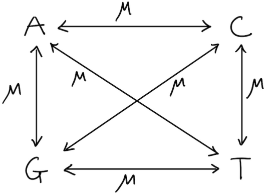
Continuous tree space (Billera-Holmes-Vogtmann)



Sufficient condition for consistency

- Model identification
  - well-studied
- Parameter space $\mathcal{T} \times \Theta$ is compact
  - bounded model parameters
  - bounded branch lengths
  - external branch lengths are bounded away from 0
- The log likelihood function $\ell(\mathbb{T}, \theta \mid Y)$ is continuous in $\mathbb{T}, \theta$
  - often true
- $E\left[\sup_{\mathbb{T},\theta} |\ell(\mathbb{T}, \theta \mid Y)|\right] < \infty$

Jukes-Cantor model

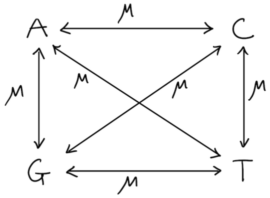


$$P = \begin{pmatrix} \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} \end{pmatrix}$$

- Model identification
- Parameter space $\mathcal{T} \times \Theta$ is compact
- The log likelihood function $\ell(\mathbb{T}, \theta \mid Y)$ is continuous in $\mathbb{T}, \theta$

$$P(Y \mid \mathbb{T}) = \frac{1}{4} \sum_{(x,y)} \prod_{(u,v) \in E} P[v = y \mid u = x, t = e_{(u,v)}]$$

Jukes-Cantor model



$$P = \begin{pmatrix} \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} \end{pmatrix}$$

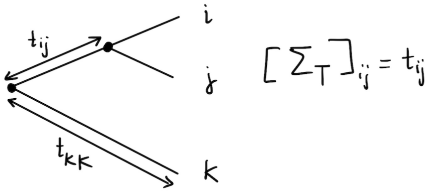
$E\left[\sup_{\mathbb{T},\theta} |\ell(\mathbb{T},\theta \mid Y)|\right] < \infty$

- Bound $P(Y \mid \mathbb{T})$ away from 0 by setting all internal nodes to A
- Probability of transition $A \to A$ is at least $1/4$
- Done since all external edges are bounded away from 0

$$P(Y \mid \mathbb{T}) = \frac{1}{4} \sum_{(x,y)} \prod_{(u,v)\in E} P[v = y \mid u = x, t = e_{(u,v)}]$$

Frequency model

- Rooted trees
- Observe the frequency of alleles
- $Y_i \mid \mathbb{T} \sim_{iid} \mathcal{N}(\kappa 1, \Sigma_{\mathbb{T}})$ (Brownian motion model)



MLE is a consistent tree reconstruction method

- Use the continuous representation of tree space
- Verify the conditions of Wald (1949) in the form given by Redner (1981)

(RoyChoudhurya et al., 2015)

Sufficient condition for consistency

Frequency model:

- Model identification
- Parameter space $\mathcal{T} \times \Theta$ is compact
  - Without loss of generality, set $\kappa = 0$.
- The log likelihood function $\ell(\mathbb{T}, \theta \mid Y)$ is continuous in $\mathbb{T}, \theta$

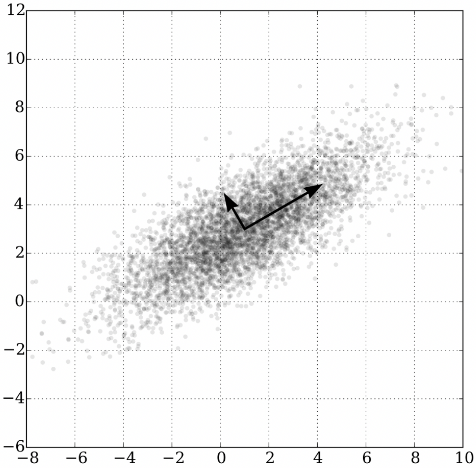
$$\ell(\mathbb{T} \mid Y) = -\frac{1}{2} \sum_{i=1}^{n} Y_i^T \Sigma_{\mathbb{T}}^{-1} Y_i - \frac{n}{2} \log |\Sigma_{\mathbb{T}}|$$

- $E\left[\sup_{\mathbb{T},\theta} |\ell(\mathbb{T}, \theta \mid Y)|\right] < \infty$
  - upper bound $Y_i^T \Sigma_{\mathbb{T}}^{-1} Y_i$
  - external edges are bounded away from 0 implies $\Sigma_{\mathbb{T}} \geq cI$ for some $c > 0$
  - $Y_i^T \Sigma_{\mathbb{T}}^{-1} Y_i \leq \frac{1}{c} Y_i^T Y_i$ and $E(Y_i^T Y_i) = \Sigma_{\mathbb{T}^*}$

Beyond Consistency

- Principal component analysis
- Hamiltonian Monte Carlo
- Regularized Estimation Methods

Principal component analysis



Principal component analysis

- Given trees $\{T_i\}_{i=1}^{n}$, construct a central point T_0:

$$T_0 = \arg\min_{T} \sum_{i=1}^{n} d(x, T_i)^2$$

- For a geodesic line L through T_0, find the projection:

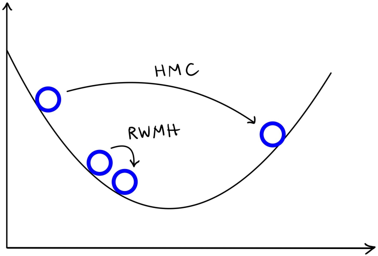
$$T_i^{(L)} = \arg\min_{T \in L} d(T, T_i)^2$$

- Find the line L_{opt} that optimizes an objective function:

$$L_{\text{opt}} = \arg\max_{L} \sum_{i=1}^{n} d(T_0, T_i^{(L)})^2$$

(Nye, 2011)

Hamiltonian Monte Carlo (HMC)

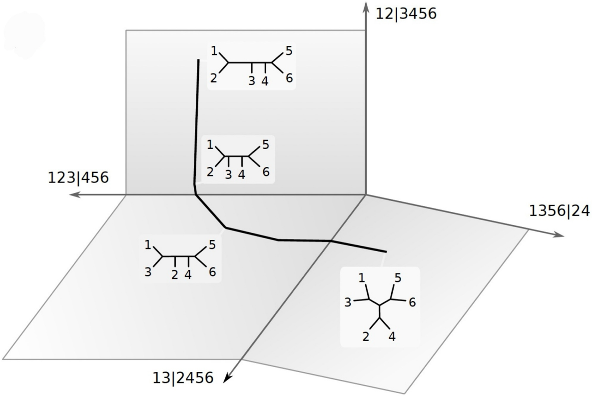


Hamiltonian's equations

$$\frac{dx_i}{dt} = \frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i},$$

where $H(x,p) = U(x) + K(p)$, with $U(x) = -\log f(x)$ and $K(p) = \|p\|_2^2/2$

Hamiltonian Monte Carlo for sampling trees



(Dinh et al., 2017)

Regularized Estimation Method

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \underbrace{\ell(\theta \mid Y)}_{\text{log likelihood}} - \lambda \underbrace{R(\theta)}_{\text{penalty}} = \arg\min_{\theta \in \Theta} - \underbrace{\ell(\theta \mid Y)}_{\text{log likelihood}} + \lambda \underbrace{R(\theta)}_{\text{penalty}}$$

- Ridge regression (L2 regularization)

$$R(\theta) = \|\theta - \theta_0\|_2^2 = \sum_{i=1}^{d} (\theta_i - \theta_0)^2$$

- Lasso (L1 regularization)

$$R(\theta) = \|\theta\|_1 = \sum_{i=1}^{d} |\theta_i|$$

Ridge estimator for tree reconstruction

$$\hat{\mathbb{T}}_{\text{ridge}} = \underset{\mathbb{T}\in\mathcal{T}}{\arg\min} -\frac{1}{k}\ell(\mathbb{T} \mid Y) + \lambda_k[d_{\text{geodesic}}(\mathbb{T}, \mathbb{T}_0)]^2$$

- insufficient signal in the gene sequences
- introduce extra information ($\mathbb{T}_0$)

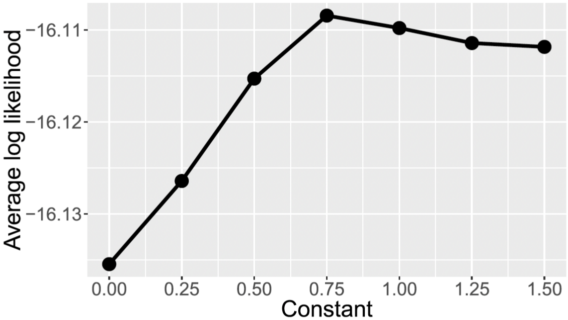
Convergence rate (Jukes-Cantor)

$$d_{\text{geodesic}}(\hat{\mathbb{T}}_{\text{ridge}}, \mathbb{T}^*) = \mathcal{O}\left(\frac{\log k}{\lambda_k\sqrt{k}} + \lambda_k\right)^{1/2}$$

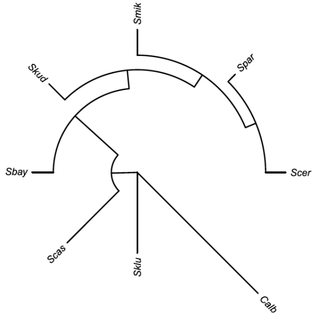
Yeast gene-tree reconstruction (YKL120W)

$$\hat{\mathbb{T}}_{\text{ridge}} = \arg\min_{\mathbb{T}\in\mathcal{T}} -\frac{1}{k}\ell(\mathbb{T} \mid Y) + \frac{C}{k^{1/4}}[d_{\text{geodesic}}(\mathbb{T}, \mathbb{T}_0)]^2$$

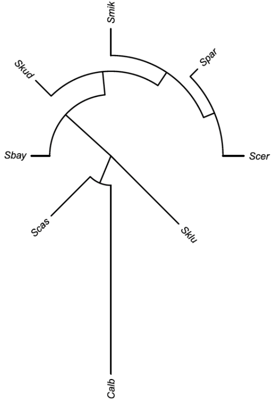
- $\mathbb{T}_0$: concatenated gene tree
- $C = 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5$



Yeast gene-tree reconstruction (YKL120W)

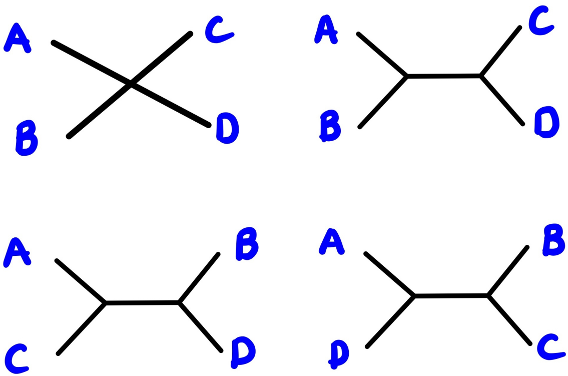


(a) Regularized method

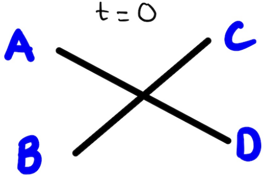


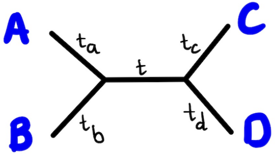
(b) MLE method

Nonbifurcating tree



Tree topology is known





Lasso

$$(\hat{t}_a, \hat{t}_b, \hat{t}_c, \hat{t}_d, \hat{t}) = \arg\min -\frac{1}{k}\ell(t_a, t_b, t_c, t_d, t) + \lambda_k(t_a + t_b + t_c + t_d + t)$$

Adaptive Lasso

$$(\tilde{t}_a, \tilde{t}_b, \tilde{t}_c, \tilde{t}_d, \tilde{t}) = \arg\min -\frac{1}{k}\ell(t_a, t_b, t_c, t_d, t) + \eta_k\left(\frac{t_a}{\hat{t}_a^\gamma} + \frac{t_b}{\hat{t}_b^\gamma} + \frac{t_c}{\hat{t}_c^\gamma} + \frac{t_d}{\hat{t}_d^\gamma} + \frac{t}{\hat{t}^\gamma}\right)$$

(Zhang et al., 2021)

Tree topology is unknown

Embedding

$$\mathbb{T} \hookrightarrow \sum_{s \in \mathcal{S}} e_{\mathbb{T},s} \zeta_s$$

Adaptive Lasso

- Step 1: MLE

$$\hat{\mathbb{T}} = \arg\max_{\mathbb{T} \in \mathcal{T}} \ell_k(\mathbb{T})$$

  where $\ell_k(\mathbb{T})$ is the log likelihood function
- Step 2: regularization

$$\hat{\mathbb{T}}_{\mathrm{AL}} = \arg\min_{\mathbb{T} \in \mathcal{T}} -\frac{1}{k}\ell_k(\mathbb{T}) + \lambda_k \left(\sum_{s \in \mathcal{S}} \frac{e_{\mathbb{T},s}}{e^{\gamma}_{\hat{\mathbb{T}},s}} \right)$$

Consistency

- $e_{\hat{\mathbb{T}}_{\mathrm{AL}},s} \to_p e_{\mathbb{T}^*,s}$
- If $e_{\mathbb{T}^*,s} = 0$, then $e_{\hat{\mathbb{T}}_{\mathrm{AL}},s} = 0$ with high probability

Sketch of Proof

Lemmas

- Convergence rate of MLE

$$d(\hat{\mathbb{T}}, \mathbb{T}^*) \leq \left(\frac{\log k}{\sqrt{k}}\right)^{1/\beta}$$

- Lojasiewicz inequality

$$\phi(\mathbb{T}) - \phi(\mathbb{T}^*) \geq c_{\mathcal{T}} d(\mathbb{T}, \mathbb{T}^*)_2^{\beta}, \quad \forall \mathbb{T} \in \mathcal{T}$$

- Concentration inequality

$$\left|\frac{1}{k}\ell_k(\mathbb{T}) - \phi(\mathbb{T})\right| \leq c\frac{\log k}{\sqrt{k}}, \quad \forall \mathbb{T} \in \mathcal{T}$$

where $\phi(\mathbb{T}) = E[\ell_1(\mathbb{T})]$

Sketch of Proof

Define

$$M(\mathbb{T}) = \sum_{s \in \mathcal{S}} \frac{e_{\mathbb{T},s}}{e^{\gamma}_{\hat{\mathbb{T}},s}}$$

$$\begin{aligned}
c_{\mathcal{T}} d(\hat{\mathbb{T}}_{\mathrm{AL}}, \mathbb{T}^*)^{\beta} &\leq \phi(\mathbb{T}^*) - \phi(\hat{\mathbb{T}}_{\mathrm{AL}}) \\
&\leq c \frac{\log k}{\sqrt{k}} + \frac{1}{k}\ell_k(\mathbb{T}^*) - \frac{1}{k}\ell_k(\hat{\mathbb{T}}_{\mathrm{AL}}) \\
&= c \frac{\log k}{\sqrt{k}} + \frac{1}{k}\ell_k(\mathbb{T}^*) - \lambda_k M(\mathbb{T}^*) \\
&\quad - \frac{1}{k}\ell_k(\hat{\mathbb{T}}_{\mathrm{AL}}) + \lambda_k M(\hat{\mathbb{T}}_{\mathrm{AL}}) + \lambda_k M(\mathbb{T}^*) - \lambda_k M(\hat{\mathbb{T}})_{\mathrm{AL}} \\
&\leq c \frac{\log k}{\sqrt{k}} + \lambda_k M(\mathbb{T}^*) \to 0
\end{aligned}$$

Sketch of Proof

- Assume that $e_{\mathbb{T}^*,s} = 0$ and $e_{\hat{\mathbb{T}}_{\text{AL}},s} > 0$ for some s
- $\mathbb{T}'$ is the same as $\hat{\mathbb{T}}_{\text{AL}}$, except $e_{\hat{\mathbb{T}}_{\text{AL}},s} = 0$

$$\lambda_k \frac{e_{\hat{\mathbb{T}}_{\text{AL}},s}}{e_{\hat{\mathbb{T}},s}} \leq \frac{1}{k}\ell_k(\hat{\mathbb{T}}_{\text{AL}}) - \frac{1}{k}\ell_k(\mathbb{T}') \leq c_{\mathcal{T}} d(\hat{\mathbb{T}}_{\text{AL}}, \mathbb{T}') = c_{\mathcal{T}} e_{\hat{\mathbb{T}}_{\text{AL}},s}$$

On the other hand,

$$e_{\hat{\mathbb{T}},s} \leq d(\hat{\mathbb{T}}, \mathbb{T}^*) \leq \left(\frac{\log k}{\sqrt{k}}\right)^{1/\beta}$$

Contradiction!

Summary

- Continuous tree space is helpful if you want to study tree reconstruction from a statistical viewpoint
- Consistency of MLE
- Regularized estimation methods can be good alternatives for MLE

Future directions

- Stein's Paradox
- "Large p, small n"
- Space of phylogenetic networks

Contact: Lam.Ho@dal.ca

Acknowledgement

NSERC
CRSNG