Notes 13 : Eigenvector-based estimation

MATH 833 - Fall 2012 *Lecturer: Sebastien Roch*

References: [MP03], [Roc10].

1 Extending majority to GTR models and general trees

The following natural generalization of the CFN model is commonly used in evolutionary biology.

DEF 13.1 (GTR model) *Fix C with $|C| \geq 2$. Let $0 < \pi \in \Delta_C$ and Q a $|C| \times |C|$ rate matrix reversible w.r.t. π, that is:*

- (Infinitesimal Generator) *Q has nonnegative off-diagonal entries and each row sums to 0.*
- (Reversibility) *For all $i, j \in C$, $\pi_i Q_{ij} = \pi_j Q_{ji}$.*

Let δ be a tree metric on $X = [n]$ with corresponding tree metric representation $(\mathcal{T}, \{w_e\}_{e \in E})$. Then a GTR model *on $\mathcal{T}$ (rooted at an arbitrary node ρ) with rate matrix Q is an MCT $(\mathcal{T}, \mathcal{P}, \pi_\rho)$ such that:*

- (Stationarity) $\pi_\rho \equiv \pi$.
- (Transition matrix) *$\mathcal{P} = \{P_e\}_{e \in E}$ is of the form*

$$P_e = e^{-w_e Q}.$$

Recall that for a matrix A the matrix exponential *is defined as*

$$e^A = \sum_{i=0}^{+\infty} \frac{A^i}{i!}.$$

For instance with

$$Q = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix},$$

we recover the CFN model above.

Because the matrix $(\pi_i^{1/2} Q_{ij} \pi_j^{-1/2})_{ij}$ is symmetric by reversibility, it is easily seen (check!) that Q is diagonalizable. Further, by the infinitesimal generator assumption, all eigenvalues are nonpositive with the largest being 0. We normalize Q as follows: let $\nu^{(1)} = \mathbf{1} = (1, \ldots, 1), \ldots, \nu^{(|C|)}$ be orthonormal eigenvectors of Q corresponding to eigenvalues $0 = \lambda_1 > \lambda_2 = -1 \geq \cdots \lambda_{|C|}$ where we assume further that

$$\sum_{\alpha \in C} \pi_\alpha (\nu_\alpha^{(i)})^2 = 1,$$

for all $i = 1, \ldots, |C|$. The second eigenvector $\nu^{(2)}$ will play a special role and we denote it simply by ν.

Given a realization $\{\Xi_v\}_{v \in V}$ of the GTR model, we let

$$\sigma_v = \nu_{\Xi_v}.$$

The appropriate generalization of majority for GTR models is then as follows: let $\{\mu_e\}_{e \in E}$ be a unit flow from ρ to $\phi(X)$ and let $\{\mu_x\}_{x \in X}$ be the flow reaching $\phi(X)$, then we let

$$Z_\mu = \sum_{x \in X} \frac{\mu_x \sigma_x}{e^{-\delta(\rho, \phi(x))}}.$$

See [MP03] and [Roc10] for a proof of the following theorem.

THM 13.2 *It holds that*

$$\mathbb{E}[Z_\mu \,|\, \sigma_\rho] = \sigma_\rho,$$

and

$$\mathrm{Var}[Z_\mu] = 1 + \sum_{e=(u,v) \in E} (1 - e^{-2w_e}) e^{2\delta(\rho, v)} \mu_e^2, \tag{1}$$

where the sum above assumes that v is furthest away from the root.

Note that minimizing the variance of Z_μ over μ is a convex quadratic optimization problem.

2 Kesten-Stigum Phase

In the Kesten-Stigum phase, a good choice of flow turns out to be the following.

THM 13.3 (Kesten-Stigum Phase) *Assume that $\mathcal{T}$ is a rooted binary phylogenetic tree with $w_e \leq g < g_* \equiv \ln \sqrt{2}$ for all e. Let μ be the flow that splits itself equally at each branching. Then,*

$$\mathrm{Var}[Z_\mu] \leq \mathcal{V} < +\infty,$$

where $\mathcal{V}$ is an absolute constant (independent of $\mathcal{T}$).

Proof: Assume the largest graphical distance between the root and the leaf set is H. Then summing the edges level by level in (1)

$$\begin{aligned}
\mathrm{Var}[Z_\mu] &\leq 1 + \sum_{h=1}^{H} 2^h (1 - e^{-2g}) e^{2hg} 2^{-2h} \\
&\leq 1 + \sum_{h=1}^{H} e^{2gh} e^{-(\ln 2)h} \\
&\leq 1 + \sum_{h=1}^{H} e^{-2(g_* - g)h} \\
&\leq 1 + \frac{1}{1 - e^{-2(g_* - g)}} < +\infty.
\end{aligned}$$

■

3 Eigenvector-based metrics

Suppose now we have k i.i.d. samples $\{\Xi_X^i\}_{i=1}^k$ from a GTR model. As before, let $\{\sigma_X^i\}_{i=1}^k$ be the corresponding eigenvector mapped states. For convenience, assume that the underlying metric is an ultrametric (although this is not needed here). Notice that in that case $1 - e^{-\delta(a,b)}$ is also an ultrametric since

$$1 - e^{-\delta(a,b)} \leq \max\{1 - e^{-\delta(a,c)}, 1 - e^{-\delta(b,c)}\} \iff \delta(a,b) \leq \max\{\delta(a,c), \delta(b,c)\}.$$

In fact, we will work with the *similarity map* $\varphi(a,b) = e^{-\delta(a,b)}$.

We consider the following similarity estimator

$$\hat{\varphi}(a,b) = \frac{1}{k} \sum_{i=1}^{k} \sigma_a^i \sigma_b^i.$$

LEM 13.4 (Unbiasedness) *It holds that*

$$\mathbb{E}[\hat{\varphi}(a,b)] = \varphi(a,b).$$

Proof: Letting

$$\hat{F}^{ab}_{\alpha,\beta} = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\{\Xi^i_a = \alpha, \Xi^i_b = \beta\},$$

note that

$$\hat{\varphi}(a,b) = \nu^{\perp}\hat{F}^{ab}\nu,$$

and therefore

$$\mathbb{E}[\hat{\varphi}(a,b)] = \nu^{\perp}\left[\pi_\alpha(e^{-\delta(a,b)Q})_{\alpha,\beta}\right]_{\alpha,\beta}\nu = e^{-\delta(a,b)}\nu^{\perp}\left[\pi_\alpha\nu_\alpha\right]_{\alpha} = e^{-\delta(a,b)}.$$

■

Further reading

Work on Steel's conjecture was initiated in the seminal paper of Mossel [Mos04]. See also [DMR06].

References

[DMR06] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Optimal phylogenetic reconstruction. In *STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 159–168, New York, 2006. ACM.

[Mos04] E. Mossel. Phase transitions in phylogeny. *Trans. Amer. Math. Soc.*, 356(6):2379–2404, 2004.

[MP03] E. Mossel and Y. Peres. Information flow on trees. *Ann. Appl. Probab.*, 13(3):817–844, 2003.

[Roc10] Sebastien Roch. Toward Extracting All Phylogenetic Information from Matrices of Evolutionary Distances. *Science*, 327(5971):1376–1379, 2010.