Notes 8 : Markov Models on Trees

MATH 833 - Fall 2012 *Lecturer: Sebastien Roch*

References: [SS03, Chapter 8].

1 Markov Chain on a Tree

We describe a standard model of nucleotide substitution. Let C be a finite character state space, e.g., $C = \{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$. Let $\mathbb{T}_n$ be the set of rooted phylogenetic trees on $X = [n]$ and $\mathbb{M}_C$ be the set of all transition matrices on C, i.e., $|C| \times |C|$ non-negative matrices whose rows sum to 1. The set of probability distributions on C is denoted by Δ_C.

DEF 8.1 (Markov Chain on a Tree (MCT)) *Let $\mathcal{T} = (T, \phi) \in \mathbb{T}_n$ with $T = (V, E)$ rooted at ρ, $\mathcal{P} = \{P^e\}_{e \in E} \in \mathbb{M}_C^E$, and $\mu_\rho \in \Delta_C$. A* Markov chain on a tree *$(\mathcal{T}, \mathcal{P}, \mu_\rho)$ is the following stochastic process $\Xi_V = \{\Xi_v\}_{v \in V}$:*

- *Pick a state Ξ_ρ for ρ according to μ_ρ.*
- *Moving away from the root toward the leaves, apply to each edge $e = \{u, v\} \in E$ with $u \leq_T v$ the transition matrix P^e independently from everything else.*

We denote by μ_V the distribution on V so obtained.

For $\xi_V = \{\xi_v\}_{v \in V} \in C^V$, the distribution μ_V can be written explicitly as

$$\begin{aligned} \mu_V(\xi_V) &= \mu_\rho(\xi_\rho) \prod_{e=\{u,v\} \in E, u \leq_T v} P^e_{\xi_u, \xi_v} \\ &= \mathbb{P}[\Xi_\rho = \xi_\rho] \prod_{e=\{u,v\} \in E, u \leq_T v} \mathbb{P}[\Xi_v = \xi_v \,|\, \Xi_u = \xi_u], \end{aligned} \tag{1}$$

where the second line follows from the construction of the process. For $W \subseteq V$, $\xi_W = \{\xi_w\}_{w \in W} \in C^W$ and $W^c = V \backslash W$, we let the *marginal* μ_W at W be

$$\mu_W(\xi_W) = \sum_{\xi_{W^c} \in C^{W^c}} \mu_V((\xi_W, \xi_{W^c})).$$

With a slight abuse of notation, for $v \in V$ and $a, b \in X$ we let $\mu_v = \mu_{\{v\}}$, $\mu_{a,b} = \mu_{\{\phi(a), \phi(b)\}}$, $\mu_{a,b,c} = \mu_{\{\phi(a), \phi(b), \phi(c)\}}$, and $\mu_X = \mu_{\phi(X)}$.

EX 8.2 (GTR Model) *A special case of the previous model that commonly arises in biology is the General Time-Reversible (GTR) model. Let π be a distribution on C satisfying $\pi(\alpha) > 0$ for all $\alpha \in C$. The $|C| \times |C|$ matrix Q is a* rate matrix *if $Q_{\alpha\beta} > 0$ for all $\alpha \neq \beta$ and $\sum_{\beta \in C} Q_{\alpha\beta} = 0$, for all $\alpha \in C$. We assume further that Q is normalized by requiring that the trace satisfies $\mathrm{tr}(Q) = \sum_{\alpha \in C} Q_{\alpha\alpha} = -1$. The rate matrix Q is* reversible with respect to π *if $\pi_\alpha Q_{\alpha\beta} = \pi_\beta Q_{\beta\alpha}$, for all $\alpha, \beta \in C$. Fix π and Q as above. Assume that in addition to $\mathcal{T}$ we are given an positive edge weight function $\tau : E \to \mathbb{R}_{++}$. Define the MCT by $\mu_\rho = \pi$ and $P^e = e^{\tau(e)Q}$ for all $e \in E$. (This corresponds to a continuous-time Markov subsititution process with rate matrix Q running for time $\tau(e)$ along edge e.)*

MCTs satisfy a natural generalization of the Markov property of discrete-time Markov processes. For $W, Z \subseteq V$, we denote by $\mu_{W|Z}$ the conditional distribution of Ξ_W given Ξ_Z.

THM 8.3 (Markov Property) *Fix $u \in \mathring{V}$ and v, a child of u, i.e., $u \leq_T v$ and $\{u, v\} \in E$. For any $W, Z \subseteq V \backslash \{v\}$ satisfying*

$$\forall w \in W,\ v \leq_T w \text{ and } \forall z \in Z,\ v \not\leq_T z.$$

Then,

$$\mu_{W|Z\cup\{u\}}(\xi_W|\xi_{Z\cup\{u\}}) = \mu_{W|\{u\}}(\xi_W|\xi_u),$$

for all $\xi_V \in C^V$. In other words, Ξ_W and Ξ_Z are conditionally independent given Ξ_u.

Proof: Clear from the construction. ∎

The main statistical problem we are interested in is the following. We will see below that reconstructing the root of the model is not in general possible. We denote by $\mathcal{T}^{-\rho}$ the phylogenetic tree $\mathcal{T}$ without its root.

DEF 8.4 (Reconstruction Problem) *Let $\boldsymbol{\Xi} = \{\Xi_X^1, \ldots, \Xi_X^k\}$ be i.i.d. samples from an MCT $(\mathcal{T}, \mathcal{P})$. Given $\boldsymbol{\Xi}$, the* tree reconstruction problem (TRP) *consists in finding a phylogenetic X-tree $\hat{\mathcal{T}}$ such that $\hat{\mathcal{T}} = \mathcal{T}^{-\rho}$. Fix $\varepsilon > 0$. Given $\boldsymbol{\Xi}$, the* full reconstruction problem (FRP) *consists in finding an MCT $(\hat{\mathcal{T}}, \hat{\mathcal{P}}, \hat{\mu}_{\hat{\rho}})$ such that the corresponding distribution $\hat{\mu}_X$ is satisfies*

$$\|\mu_X - \hat{\mu}_X\|_1 \equiv \sum_{\xi_X \in C^X} |\mu_X(\xi_X) - \hat{\mu}_X(\xi_X)| \leq \varepsilon.$$

2 Issues with Identifiability

For the reconstruction to be well-posed, it must be that the model is identifiable.

DEF 8.5 (Identifiability) *Let $\Lambda = \{\lambda_\theta\}_{\theta\in\Theta}$ be a family of distributions parametrized by $\theta \in \Theta$. We say that Λ is* identifiable *if: $\nu_\theta \sim \nu_{\theta'}$ if and only if $\theta = \theta'$.*

We will show below that the tree reconstruction problem is identifiable under the following conditions:

$$\mu_\rho > 0,$$

and

$$\forall e \in E,\ \det P^e \neq 0, \pm 1.$$

We denote by Θ_n the family of MCT on $X = [n]$ satisfying the conditions above.

Re-rooting. We begin by explaining why we restrict ourselves to reconstructing unrooted trees.

THM 8.6 (Re-rooting an MCT) *Let $(\mathcal{T}, \mathcal{P}, \mu_\rho) \in \Theta_n$ with $\mathcal{T} = (T, \phi)$ and $T = (V, E)$ rooted at ρ. Then for any $u \in \mathring{V}$ we can re-root the MCT at u without changing the distribution of the process.*

Proof: It suffices to show that the model can be re-rooted at a neighbour u of ρ. Let $\bar{T}$ be the tree T re-rooted at u. Let $e = \{\rho, u\}$ and define

$$\bar{P}^e = U_u^{-1}(P^e)^t U_\rho,$$

where U_v is the $|C| \times |C|$ diagonal matrix with diagonal μ_v. Note that the assumptions in Θ_n imply that $\mu_u > 0$. Indeed, since $\mu_u = \mu_\rho P^e$ (interpreting μ_v as a row vector) and $\mu_\rho > 0$, a zero component in μ_u would imply the existence of a zero column in P^e in which case we would have $\det P^e = 0$—a contradiction.

Let $(\bar{\mathcal{T}}, \bar{\mathcal{P}}, \bar{\mu}_{\bar{\rho}})$ be the MCT with $\bar{\mathcal{T}} = (\bar{T}, \bar{\phi})$ and $\bar{T} = T$ rooted at $\bar{\rho} = u$, $\bar{\mathcal{P}}$ being the same as $\mathcal{P}$ except for the matrix along e which is now $\bar{P}^e$, and $\bar{\mu}_{\bar{\rho}} = \mu_u$. Let $\bar{\mu}_X$ be the corresponding distribution. We claim that $\mu_X \sim \bar{\mu}_X$. Note that by Bayes' rule, for $\alpha, \beta \in C$,

$$\bar{P}^e_{\alpha,\beta} = \mathbb{P}[\Xi_\rho = \beta \,|\, \Xi_u = \alpha],$$

where $\Xi_V \sim (\mathcal{T}, \mathcal{P}, \mu_\rho)$. The result then follows from (1). ∎

Determinantal conditions. We show through an example that the determinantal conditions above are natural.

EX 8.7 *Fix $C = \{0,1\}$ and $X = \{a,b,c,d\}$. Let $q_1 = ab|cd$ and $q_2 = ac|bd$ be two quartet trees on X. Assign to each edge of q_1 and q_2 the same transition matrix P. Denote by μ^1 and μ^2 the corresponding distributions. We consider two cases:*

- *Suppose* $\det P = 0$. *Then, it must be that*

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

  for some $0 < p < 1$. But then the endpoints of any edge are independent. (Notice that, in general, this is not the case when $|C| > 2$.) In particular, the states at the leaves are independent and $\mu^1 \sim \mu^2$. Therefore, the model is not identifiable in that case.

- *Suppose P is the identity matrix, in which case* $\det P = 1$. *Then all states are equal and, again, $\mu^1 \sim \mu^2$.*

3 Log-Det Distance

Our main result is the following.

THM 8.8 (Tree Identifiability) *Let $(\mathcal{T}, \mathcal{P}, \mu_\rho) \in \Theta_n$ with corresponding distribution μ_V. Then $\mathcal{T}^{-\rho}$ can be obtained from μ_X. In fact, pairwise marginals $\{\mu_{a,b}\}_{a,b\in X}$ suffice to derive $\mathcal{T}^{-\rho}$.*

Proof: The proof relies on the Uniqueness of the Tree Metric Representation through the following metric:

DEF 8.9 (Logdet Distance) *For $a, b \in X$, let P^{ab} be defined as follows*

$$\forall \alpha, \beta \in C,\ P^{ab}_{\alpha,\beta} = \mathbb{P}[\Xi_{\phi(b)} = \beta \,|\, \Xi_{\phi(a)} = \alpha].$$

The logdet distance *between a and b is the dissimilarity map*

$$\delta(a,b) = -\frac{1}{2} \log \det[P^{ab} P^{ba}].$$

We return to the proof. We claim that δ is a tree metric on $\mathcal{T}$ with edge weights

$$w_e = -\frac{1}{2}\log\det[P^e\bar{P}^e] > 0,$$

where $\bar{P}^e$ is defined as above. Let u be the most recent common ancestor of a and b under $\leq_T$. Let $e'_1, \ldots, e'_{m'}$ (resp. $e_1, \ldots, e_m$) be the path between u and a (resp. b). Then re-rooting at a and b respectively we get

$$P^{ab} = \bar{P}^{e'_{m'}} \cdots \bar{P}^{e'_1} P^{e_1} \cdots P^{e_m},$$

and

$$P^{ba} = \bar{P}^{e_m} \cdots \bar{P}^{e_1} P^{e'_1} \cdots P^{e'_{m'}}.$$

The result then follows from the multiplicativity of the determinant.

■

EX 8.10 (GTR case) *Let π and Q as in the example above. Then*

$$\begin{aligned} w_e &= -\frac{1}{2}\log\det[P^e\bar{P}^e] \\ &= -\frac{1}{2}\log\det[e^{\tau(e)Q}e^{\tau(e)Q}] \\ &= -\frac{1}{2}\log\left[\det[e^{\tau(e)Q}]^2\right] \\ &= -\log[e^{\tau(e)\mathrm{tr}(Q)}] \\ &= \tau(e), \end{aligned}$$

where we used reversibility on the second line, properties of the trace of exponentials on the fourth line, and the normalization of Q on the fifth line. As a result,

$$\delta(a,b) = -\frac{1}{2}\log\det[P^{ab}P^{ba}] = \sum_{e\in\mathrm{Path}(a,b)} \tau(e).$$

4 Chang's Eigenvector Decomposition

We showed that the tree structure can be deduced from pairwise distributions at the leaves. Interestingly, this is not the case for transition matrices. For an example, see [Cha96]. Instead, one must use three-way distributions. To avoid the possibility of "relabeling" the states at interior vertices, we also require the following assumption.

DEF 8.11 (Reconstructibility from Rows) *A class of transition matrices $\mathcal{M}$ is* reconstructible by rows *if for each $P \in \mathcal{M}$ and each permutation matrix $\Pi \neq I$ (i.e., a stochastic matrix with exactly one 1 in each row and column) we have $\Pi P \notin \mathcal{M}$.*

For instance, matrices such that the diagonal is striclty largest in each column is reconstructible by rows.

THM 8.12 (Full Identifiability) *Let Θ_n^* be a subclass of Θ_n wehere $\mathcal{P} = \{P^e\}$ and $\overline{\mathcal{P}} = \{\bar{P}^e\}$ are in a class $\mathcal{M}$ reconstructible by rows. Let $(\mathcal{T}, \mathcal{P}, \mu_\rho) \in \Theta_n^*$ with corresponding distribution μ_V. Then up to the placement of the root $(\mathcal{T}, \mathcal{P}, \mu_\rho)$ can be derived from μ_X. In fact, three-way marginals $\{\mu_{a,b,c}\}_{a,b,c \in X}$ suffice for this purpose.*

Proof: From the previous theorem, we can derive the unrooted tree structure. Because the the transition matrix on a path is the product of the transition matrices on the corresponding edges, it can be shown that it suffices to recover the transition matrices for the case $n = 3$. See [Cha96] for the full details.

Let $\mathcal{T}$ be a tree on three leaves $\{a, b, c\}$ rooted at the interior vertex m. Denote by Ξ_V the state vector. By abuse of notation, we denote $\Xi_x = \Xi_{\phi(x)}$ for $x \in X$. Chang's clever solution to the identifiability problem relies on the following calculation [Cha96]. Fix $\gamma \in C$. Note that

$$\begin{aligned}
\mathbb{P}[\Xi_c = \gamma, \Xi_b = j \,|\, \Xi_a = i] &= \sum_{k \in C} \mathbb{P}[\Xi_m = k, \Xi_c = \gamma, \Xi_b = j \,|\, \Xi_a = i] \\
&= \sum_{k \in C} \mathbb{P}[\Xi_m = k \,|\, \Xi_a = i] \\
&\qquad \times \mathbb{P}[\Xi_c = \gamma \,|\, \Xi_m = k, \Xi_a = i] \\
&\qquad \times \mathbb{P}[\Xi_b = j \,|\, \Xi_c = \gamma, \Xi_m = k, \Xi_a = i] \\
&= \sum_{k \in C} \mathbb{P}[\Xi_m = k \,|\, \Xi_a = i] \\
&\qquad \times \mathbb{P}[\Xi_c = \gamma \,|\, \Xi_m = k] \\
&\qquad \times \mathbb{P}[\Xi_b = j \,|\, \Xi_m = k].
\end{aligned}$$

Writing $P^{ab,\gamma}$ for the matrix with elements $\mathbb{P}[\Xi_c = \gamma, \Xi_b = j \,|\, \Xi_a = i]$, we obtain in matrix form

$$P^{ab,\gamma} = P^{am} \mathrm{Diag}(P^{mc}_{\cdot\gamma}) P^{mb},$$

or, since $(P^{ab})^{-1} = (P^{mb})^{-1}(P^{am})^{-1}$,

$$(P^{ab})^{-1} P^{ab,\gamma} = (P^{mb})^{-1} \mathrm{Diag}(P^{mc}_{\cdot\gamma}) P^{mb}.$$

Notice that:

- The L.H.S. depends only on μ_X.
- The R.H.S. is an eigenvector decomposition.

We want to extract P^{mb} from the decomposition above. There are two issues:

1. If the entries of $P^{mc}_{\cdot\gamma}$ are not distinct, the eigenvector decomposition is not unique.
2. The eigenvectors are not ordered.

The second point is taken care of by the reconstructibility from rows assumption. The solution to the first issue is to "force" the entries to be distinct. For instance, pick a random vector $(G_\gamma)_{\gamma\in C}$ where each entry is an independent $N(0,1)$. Define

$$\tilde{P}^{ab,G} = \sum_{\gamma\in C} P^{ab,\gamma} G_\gamma,$$

and

$$\tilde{P}^{mc}_{\cdot} = \sum_{\gamma\in C} P^{mc}_{\cdot\gamma} G_\gamma.$$

Then, with probability 1, the entries of $\tilde{P}^{mc}_{\cdot}$ are distinct and

$$(P^{ab})^{-1}\tilde{P}^{ab,G} = (P^{mb})^{-1}\mathrm{Diag}(\tilde{P}^{mc}_{\cdot})P^{mb}.$$

■

Further reading

The definitions and results discussed here were taken from Chapter 8 of [SS03]. Much more on the subject can be found in that excellent monograph. See also [SS03] for the relevant bibliographic references. The identifiability proofs are from [Ste94, Cha96].

References

[Cha96] Joseph T. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.*, 137(1):51–73, 1996.

[SS03] Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[Ste94] M. Steel. Recovering a tree from the leaf colourations it generates under a Markov model. *Appl. Math. Lett.*, 7(2):19–23, 1994.