Notes 9 : Consistency Results

MATH 833 - Fall 2012 *Lecturer: Sebastien Roch*

References: [SS03, Chapter 8].

1 Consistency

A natural property of statistical estimators is the following:

DEF 9.1 (Consistency) *Let $\Lambda = \{\lambda_\theta\}_{\theta\in\Theta}$ be a family of distributions parametrized by $\theta \in \Theta$. We say that a sequence of estimators $\{\hat\theta_n\}_{n\geq 0}$ of θ, where $\hat\theta_n$ is based on n i.i.d. samples from λ_θ, is* consistent *if, under λ_θ, $\hat\theta_n$ converges in probability to θ.*

Clearly, identifiability is necessary for consistency to hold.

We fix $X = [n]$. Although the consistency results we discuss here hold more generally, we illustrate them on the CFN model for simplicity. Further, since MCTs (under the assumptions of positive root distribution and non-zero determinant transition matrices) are identifiable up to the root placement, we root the tree arbitrarily at leaf 1.

DEF 9.2 (CFN Model) *A CFN model is an MCT $(\mathcal{T}, \mathcal{P}, \mu_\rho)$ on $C = \{0, 1\}$ with symmetric transition matrices with positive determinant and uniform μ_ρ. In particular, each transition matrix $P^e = \bar{P}^e$ is characterized by a single parameter $0 < p_e < 1/2$, the mutation probability along edge e.*

Applying the log-det formula, we define the branch length of an edge as

$$w_e = -\log(1 - 2p_e).$$

Maximum parsimony. A classical result of Felsenstein [Fel78] implies that parsimony is not consistent.

THM 9.3 (MP is Not Consistent) *Maximum parsimony is not consistent.*

Proof: Take $n = 4$ and consider the quartet tree $q = 12|34$. Denoting by e_m the middle edge and by e_x the edge incident to $\phi(x)$, define

$$p_{e_1} = p_{e_3} = \frac{1}{2} - \varepsilon, \quad p_{e_2} = p_{e_4} = p_{e_m} = \varepsilon,$$

with $\varepsilon > 0$ (small). For $A \subseteq X\backslash\{4\}$, let $\bar{p}_A$ be the probability that the character observed under the CFN model on q corresponds to the split $A|X\backslash A$. The expected parsimony score of q is

$$1 - \bar{p}_\emptyset + \bar{p}_{\{1,3\}} + \bar{p}_{\{2,3\}},$$

since all characters can be explained with a single mutation except for $\{1,3\}|\{2,4\}$ and $\{2,3\}|\{1,4\}$. Similarly, the expected parsimony score of $q' = 13|24$ (for a sample generated under q with the parameters above) is

$$1 - \bar{p}_\emptyset + \bar{p}_{\{1,2\}} + \bar{p}_{\{2,3\}}.$$

But note that as $\varepsilon \to 0$

$$\bar{p}_{\{1,3\}} = \frac{1}{4} + O(\varepsilon),$$

and

$$\bar{p}_{\{1,2\}} = O(\varepsilon),$$

so that q' has a smaller expected score. By the law of large numbers, with probability one the wrong tree will eventually be chosen. The phenomenon underlying this example is known as *long-branch attraction.* ■

Maximum likelihood. In the maximum likelihood (ML) problem, we are looking for a tree $\hat{\mathcal{T}} = (\hat{T}, \hat{\phi})$ with $\hat{T} = (\hat{V}, \hat{E})$ rooted at leaf $\hat{\phi}(1)$ and edge mutation parameters $\hat{\mathcal{P}} = \{\hat{p}_e\}_{e\in\hat{E}}$, so that the log-likelihood, that is, the logarithm of the probability of observing the samples $\boldsymbol{\Xi} = \{\Xi^1_X, \ldots, \Xi^k_X\}$ under $(\hat{\mathcal{T}}, \hat{\mathcal{P}})$,

$$\mathcal{L}(\boldsymbol{\Xi}\,|\,\hat{\mathcal{T}}, \hat{\mathcal{P}}) = \sum_{i=1}^{k} \mathcal{L}(\Xi^i_X\,|\,\hat{\mathcal{T}}, \hat{\mathcal{P}})$$

is maximized, where each term in the sum is the log-likelihood of a single sample.

THM 9.4 (Consistency of ML. See [Cha96] for details.) *Under the CFN model, ML is consistent.*

Proof:(Sketch) The consistency of ML follows from a standard argument appealing to the identifiability of the model, the law of large numbers applied to the log-likelihood, the continuity of the log-likelihood, and the non-negativity of the Kullback-Leibler divergence (which follows from Jensen's Inequality).

LEM 9.5 (Information Inequality. See [CT91].) *Let μ, ν be (stricly positive) probability distributions on a finite set S. The* Kullback-Leibler divergence *between μ and ν is defined as*

$$D(\mu||\nu) = \sum_{\alpha \in S} \mu(\alpha) \log \frac{\mu(\alpha)}{\nu(\alpha)}.$$

Then, $D(\mu||\nu) \geq 0$ with equality if $\mu \equiv \nu$.

■

Distance methods. Most reasonable distance methods are consistent. We give one example. Recall that the log-det distance is given by:

DEF 9.6 (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*

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

In the case of the CFN model, we have

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

where p^{ab} is probability that the states at a and b differ—in particular, $p^{ab} = p^{ba}$. Hence

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

For $q = ab|cd$, define

$$\delta(q) = \frac{1}{2}[\delta(a, c) + \delta(b, d) - \delta(a, b) - \delta(c, d)].$$

Recall that, if δ is a tree metric (as is the case for the log-det distance), then among all 4-tuples over $X' = \{a, b, c, d\}$ $\delta(q)$ takes three possible values

$$\delta(q) \in \{w_{e_0}, 0, -w_{e_0}\}, \tag{1}$$

where e_0 is the middle edge of $\mathcal{T}|X'$.

Consider the following algorithm.

- For all $a, b \in X$ distinct, let

  $$\hat{p}^{ab} = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\{\Xi_a^i \neq \Xi_b^i\}.$$

  and

  $$\hat{\delta}(a, b) = -\log[1 - 2\hat{p}^{ab}].$$

  (Make the last quantity $+\infty$ if the term inside the log is negative.)

- Set $\mathcal{Q} = \emptyset$.

- For all $a, b, c, d \in X$ distinct,

  - Setting $X' = \{a, b, c, d\}$, let

    $$xy|wz = \arg\max\{\hat{\delta}(xy|wz) \ : \ x, y, w, z \in X' \text{ distinct}\},$$

    where $\hat{\delta}(q)$ is defined similarly to $\delta(q)$ above.

  - Add $xy|wz$ to $\mathcal{Q}$.

- Apply the Strong Quartet Evidence algorithm to $\mathcal{Q}$ to recover $\mathcal{T}$.

Clearly, by (1), if we were to run the algorithm above with δ rather than $\hat{\delta}$, the correct tree $\mathcal{T}$ would be reconstructed. However, $\hat{\delta}$ is only an approximation of δ. By the strong law of large numbers, this approximation gets arbitrarily better as $k \to \infty$ with probability 1. The consistency of the algorithm then follows from the following inequality:

$$\max_q |\delta(q) - \hat{\delta}(q)| \leq 2 \max_{a,b} |\delta(a, b) - \hat{\delta}(a, b)| < \frac{1}{2} \min_e w_e \equiv \frac{1}{2} w_*,$$

with probability 1 for all sufficiently large k.

General Models. The results we discussed here extend beyond the CFN model (under approrpiate assumptions). See for instance [Cha96, DMR09].

References

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

[CT91] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley Series in Telecommunications. John Wiley & Sons Inc., New York, 1991. A Wiley-Interscience Publication.

[DMR09] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Phylogenies without branch bounds: Contracting the short, pruning the deep. In *RECOMB*, pages 451–465, 2009.

[Fel78] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Biol.*, pages 401–410, 1978.

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