Explicit bounds for the stability of maximum likelihood trees

M. Mariadassou A. Bar-Hen

Department MMIP
AgroParisTech

September 07
Phylogeny Workshop, Isaac Newton Institute

Motivation: the stability issue in phylogeny

- Inferred topology might not be the "true" topology;
- Problems to draw conclusions from the topology;
- For example, how confident are we in a clade?
- One cause of variability is data sampling;
- Focus on ML methods and their statistical properties;

Motivation: the stability issue in phylogeny

- Inferred topology might not be the "true" topology;
- Problems to draw conclusions from the topology;
- For example, how confident are we in a clade?
- One cause of variability is data sampling;
- Focus on ML methods and their statistical properties;

Notations I

Data structure

- Data matrix $\mathcal{X} = (X_{ij})$ of size $s \times n$;
- X_{ij} nucleotide j in species i valued in $\mathcal{A} = \{A, C, G, T\}$;
- $\mathbf{X}_i$ i-th column of $\mathcal{X}$, vector of size s;
- $\mathbf{X}_i$ nucleotide pattern of site i, *e.g.* $(AAATTT)'$.

Phylogenetic model T

- An evolution model: substitution model and associated parameters;
- A tree topology: branching pattern and branch lengths.

Notations I

Data structure

- Data matrix $\mathcal{X} = (X_{ij})$ of size $s \times n$;
- X_{ij} nucleotide j in species i valued in $\mathcal{A} = \{A, C, G, T\}$;
- $\mathbf{X}_i$ i-th column of $\mathcal{X}$, vector of size s;
- $\mathbf{X}_i$ nucleotide pattern of site i, *e.g.* $(AAATTT)'$.

Phylogenetic model T

- An evolution model: substitution model and associated parameters;
- A tree topology: branching pattern and branch lengths.

Notations II

- $\mathbf{X}_i$ $i.i.d.$ with shared distribution Q;
- Empirical distribution $Q_n = \sum_i \delta_{\mathbf{X}_i}$ of the nucleotides;
- Support of Q made of all patterns with positive probability:

$$\mathcal{N}_s \subset \mathcal{A}^s \quad \text{Card}(\mathcal{N}_s) \leq 4^s$$

- True and empirical mean log-likelihood of T:

$$\ell^T = \mathbb{E}_Q[\log \mathbb{P}(\mathbf{X}; T)] = \sum_{x \in \mathcal{N}_s} Q(x) \log \mathbb{P}(x; T)$$

$$\ell_n^T = \mathbb{E}_{Q_n}[\log \mathbb{P}(\mathbf{X}; T)] = \frac{1}{n} \sum_i \log \mathbb{P}(\mathbf{X}_i; T)$$

where $\mathbb{P}(x; T)$ is the probability of pattern x under model T;

Notations II

- $\mathbf{X}_i$ $i.i.d.$ with shared distribution Q;
- Empirical distribution $Q_n = \sum_i \delta_{\mathbf{X}_i}$ of the nucleotides;
- Support of Q made of all patterns with positive probability:

$$\mathcal{N}_s \subset \mathcal{A}^s \quad \text{Card}(\mathcal{N}_s) \leq 4^s$$

- True and empirical mean log-likelihood of T:

$$\begin{aligned} \ell^T &= \mathbb{E}_Q[\log \mathbb{P}(\mathbf{X}; T)] = \sum_{x \in \mathcal{N}_s} Q(x) \log \mathbb{P}(x; T) \\ \ell_n^T &= \mathbb{E}_{Q_n}[\log \mathbb{P}(\mathbf{X}; T)] = \frac{1}{n} \sum_i \log \mathbb{P}(\mathbf{X}_i; T) \end{aligned}$$

where $\mathbb{P}(x; T)$ is the probability of pattern x under model T;

ℓ^T as a scalar product

- Replace Q and Q_n, true and empirical pattern distribution, with:

$$\begin{aligned} \theta^x &= \mathbb{P}_Q(\mathbf{X} = x) \\ \theta_n^x &= \mathbb{P}_{Q_n}(\mathbf{X} = x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\mathbf{X}_i = x\}} \end{aligned}$$

  $\boldsymbol{\theta} = (\theta^x)_{x \in \mathcal{N}_s}$ and $\boldsymbol{\theta}_n = (\theta_n^x)_{x \in \mathcal{N}_s}$;

- Then, with $\log P^T = (\log \mathbb{P}(x, T))_{x \in \mathcal{N}_s}$.

$$\begin{aligned} \ell^T &= \mathbb{E}_Q[\log \mathbb{P}(\mathbf{X}; T)] = \boldsymbol{\theta}.\log P^T \\ \ell_n^T &= \mathbb{E}_{Q_n}[\log \mathbb{P}(\mathbf{X}; T)] = \boldsymbol{\theta}_n.\log P^T \end{aligned}$$

- $\ell^T - \ell_n^T = (\boldsymbol{\theta} - \boldsymbol{\theta}_n).\log P^T$

ℓ^T as a scalar product

- Replace Q and Q_n, true and empirical pattern distribution, with:

$$\begin{aligned} \theta^x &= \mathbb{P}_Q(\mathbf{X} = x) \\ \theta_n^x &= \mathbb{P}_{Q_n}(\mathbf{X} = x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\mathbf{X}_i = x\}} \end{aligned}$$

  $\boldsymbol{\theta} = (\theta^x)_{x \in \mathcal{N}_s}$ and $\boldsymbol{\theta}_n = (\theta_n^x)_{x \in \mathcal{N}_s}$;

- Then, with $\log P^T = (\log \mathbb{P}(x, T))_{x \in \mathcal{N}_s}$.

$$\begin{aligned} \ell^T &= \mathbb{E}_Q[\log \mathbb{P}(\mathbf{X}; T)] = \boldsymbol{\theta}.\log P^T \\ \ell_n^T &= \mathbb{E}_{Q_n}[\log \mathbb{P}(\mathbf{X}; T)] = \boldsymbol{\theta}_n.\log P^T \end{aligned}$$

- $\ell^T - \ell_n^T = (\boldsymbol{\theta} - \boldsymbol{\theta}_n).\log P^T$

Concentration Inequalities I

- $\ell^T - \ell_n^T = (\boldsymbol{\theta} - \boldsymbol{\theta}_n).\log P^T$
- To control $\ell^T - \ell_n^T$, we need to control $\boldsymbol{\theta} - \boldsymbol{\theta}_n$, the difference between the true and the empirical pattern distribution;
- Probability of $\{\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| > \epsilon\}$ decreases exponentially towards 0;
- At what rate?

Using measure concentration tools, we obtain:

$$\frac{\log \mathbb{P}(\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| > \varepsilon)}{n} \leq \frac{\log |\mathcal{N}_s|}{n} + \frac{\log 2}{n} + \max_{x \in \mathcal{N}_s} \frac{-\varepsilon^2}{\theta^x(1 - \theta^x + \varepsilon)}$$

Concentration Inequalities II

This leads to:

$$\frac{\log \mathbb{P}\left(|\ell^T - \ell_n^T| \geq \varepsilon\right)}{n} \leq \frac{\log |\mathcal{N}_s|}{n} + \frac{\log 2}{n} + \max_{x \in \mathcal{N}_s} \frac{-\tilde{\varepsilon}^2}{\theta^x(1-\theta^x+\tilde{\varepsilon})}$$

Where $\tilde{\varepsilon} = \dfrac{\varepsilon}{|\mathcal{N}_s| \| \log P^T \|}$.

Remarks:

- For a given confidence level, we know how n evolves with s;
- Sharp bound for small $\mathcal{N}_s \Rightarrow$ accurate estimation of $|\mathcal{N}_s|$ is crucial;
- For simple models (JC69,K2P), patterns (*e.g.* $YYRR$) can be merged $\Rightarrow$ smaller $\mathcal{N}_s$.

Concentration Inequalities II

This leads to:

$$\frac{\log \mathbb{P}\left(|\ell^T - \ell_n^T| \geq \varepsilon\right)}{n} \leq \frac{\log |\mathcal{N}_s|}{n} + \frac{\log 2}{n} + \max_{x \in \mathcal{N}_s} \frac{-\tilde{\varepsilon}^2}{\theta^x(1-\theta^x+\tilde{\varepsilon})}$$

Where $\tilde{\varepsilon} = \dfrac{\varepsilon}{|\mathcal{N}_s|\|\log P^T\|}$.

Remarks:

- For a given confidence level, we know how n evolves with s;
- Sharp bound for small $\mathcal{N}_s \Rightarrow$ accurate estimation of $|\mathcal{N}_s|$ is crucial;
- For simple models (JC69,K2P), patterns (*e.g.* $YYRR$) can be merged $\Rightarrow$ smaller $\mathcal{N}_s$.

Inversions events

- ML methods based on the model ranking induced by their likelihood score;
- But inference done on ranking induced by empirical likelihood score;
- Inversion events between models T and T' can happen;
- When comparing two models T and T', the true ranking may be different from the empirical one;
- How often does such an event happens?
- How does its probability $\mathbb{P}(\ell_n^T - \ell_n^{T'} < 0 | \ell^T - \ell^{T'} > 0)$ decreases when available information increases?

Inversions events

- ML methods based on the model ranking induced by their likelihood score;
- But inference done on ranking induced by empirical likelihood score;
- Inversion events between models T and T' can happen;
- When comparing two models T and T', the true ranking may be different from the empirical one;
- How often does such an event happens?
- How does its probability $\mathbb{P}(\ell_n^T - \ell_n^{T'} < 0 | \ell^T - \ell^{T'} > 0)$ decreases when available information increases?

Inversions events

- ML methods based on the model ranking induced by their likelihood score;
- But inference done on ranking induced by empirical likelihood score;
- Inversion events between models T and T' can happen;
- When comparing two models T and T', the true ranking may be different from the empirical one;
- How often does such an event happens?
- How does its probability $\mathbb{P}(\ell_n^T - \ell_n^{T'} < 0 | \ell^T - \ell^{T'} > 0)$ decreases when available information increases?

Concentration results

Still using concentration tools, we obtain:

Proposition

Assume that model T is better than model T' ($\ell^T > \ell^{T'}$), then the probability that T' is better than T for our sample is such that:

$$\frac{\log \mathbb{P}(\ell_n^T - \ell_n^{T'} < 0)}{n} \leq \frac{\log |\mathcal{N}_s|}{n} + \max_{x \in \mathcal{N}_s} \frac{-\varepsilon^2}{\theta^x(1 - \theta^x + \varepsilon)}$$

where $\varepsilon = \frac{\ell^T - \ell^{T'}}{|\mathcal{N}_s| \| \log P^T - \log P^{T'} \|}$ *and* $\boldsymbol{\theta} = (\mathbb{P}_Q(\mathbf{X} = x))_{x \in \mathcal{N}_s}$.

Remarks:

- Expected result: inversion probability decreases with $\ell^T - \ell^{T'}$;
- Patterns with same likelihood under T and T' can be removed from $\mathcal{N}_s$.

Comparison with bootstrap

- Bootstrap aim: to evaluate the probability of the inferred phylogeny not being the true one;
- The only variability is the observed variability (non-parametric bootstrap);
- Significance threshold decided *ex-ante* (66% or 95%) with no justification;
- No connection between n and s for threshold;
- Bootstrap relies heavily on simulations.

Some aspects of the method

- Accounts for more variability than just the one observed in the data;
- Accounts for s and n when calculating a confidence level;
- For a given $\mathcal{N}_s$ and a given confidence level $1-\alpha$, calculate the necessary number of sites for threshold.
- Focus on the likelihood score (instead of the topology);

Perspectives

- Accurately estimate $|\mathcal{N}_s|$;
- Compare models chosen with respect to the data;
- Compare our method with bootstrap on data (real and/or simulated);
- Consider process bounds instead of pointwise bounds;
- Anything else I can think about.