Models selection procedures for random objects driven by context tree models

Aline Duarte

works in collaboration with R. Fraiman, A. Galves, N. Hernandez, F. Lopes, G. Ost and C. Vargas

Universidade de São Paulo
CEPID NeuroMat

UFRJ, November 2021
A stochastic chain produces a sequence of stimuli.

$X_1, X_2, \ldots, X_n$ to $Y_1, Y_2, \ldots, Y_n$
A stochastic chain produces a sequence of stimuli 

$X_1, X_2, \ldots, X_n$ → $Y_1, Y_2, \ldots, Y_n$
A stochastic chain produces a sequence of stimuli $X_1, X_2, \ldots, X_n$ that affect game performance $Y_1, Y_2, \ldots, Y_n$.

https://game.numec.prp.usp.br/
Example of stochastic stimuli

**Stimulus unit**

- 2 - strong beat
- 1 - weak beat
- 0 - silent unit

- 2 - right
- 1 - center
- 0 - left

Stimuli sequence

1. Start with a deterministic sequence

2. Replace each weak beat by a silent unit independently and with a probability \( \epsilon > 0 \)

One example of stochastic sequence would be:

```
2 1 1
2 1 1
2 1 1
2 1 1
2 1 0
2 1 1
2 0 0
2 1 1
2 0 1
```
Example of stochastic stimuli

Stimulus unit

- 2 - strong beat
- 1 - weak beat
- 0 - silent unit

- 2 - right
- 1 - center
- 0 - left

Stimuli sequence

1. Start with a deterministic sequence

   \[ 2 \ 1 \ 1 \ 2 \ 1 \ 1 \ 2 \ 1 \ 1 \ 2 \ 1 \ 1 \ 2 \ 1 \ 1 \ 2 \ \cdots \]

2. Replace each weak beat by a silent unit independently and with a probability \( \epsilon > 0 \)

One example of stochastic sequence would be

\[ 2 \ 1 \ 1 \ 2 \ 1 \ 1 \ 2 \ 1 \ 1 \ 2 \ 1 \ 1 \ 2 \ \cdots \]
\[ 2 \ 1 \ 0 \ 2 \ 1 \ 1 \ 2 \ 0 \ 0 \ 2 \ 1 \ 1 \ 2 \ 0 \ 1 \ 2 \ \cdots \]
Characterizing the stochastic sequence

\[ \tau = \{2, 21, 11, 01, 20, 10, 00\} : \]

\[
\begin{array}{cccc}
\text{context } w & p(0|w) & p(1|w) & p(2|w) \\
\hline
2 & \epsilon & 1 - \epsilon & 0 \\
21 & \epsilon & 1 - \epsilon & 0 \\
20 & \epsilon & 1 - \epsilon & 0 \\
11 & 0 & 0 & 1 \\
10 & 0 & 0 & 1 \\
01 & 0 & 0 & 1 \\
00 & 0 & 0 & 1 \\
\end{array}
\]
Characterizing the stochastic sequence

\[ \tau = \{2, 21, 11, 01, 20, 10, 00\} : \]

The two objects characterize the stochastic simulation algorithm.
Characterizing the stochastic sequence

\[ \tau = \{2, 21, 11, 01, 20, 10, 00\} : \]

\[ \tau = \{2, 21, 11, 01, 20, 10, 00\} : \]

\[ \tau = \{2, 21, 11, 01, 20, 10, 00\} : \]
Characterizing the stochastic sequence

\[ 2 \ 1 \ 1 \ 2 \ 1 \ 1 \ 2 \ 1 \ 0 \ 2 \ 1 \ 1 \ 2 \ 0 \ 1 \ 2 \ 1 \ 1 \ 2 \ 0 \ 0 \ 2 \ \cdots \]

(family of transition probabilities)

| context w | p(0|w) | p(1|w) | p(2|w) |
|-----------|-------|-------|--------|
| 2         | \(\epsilon\) | 1 - \(\epsilon\) | 0       |
| 21        | \(\epsilon\) | 1 - \(\epsilon\) | 0       |
| 20        | \(\epsilon\) | 1 - \(\epsilon\) | 0       |
| 11        | 0     | 0     | 1       |
| 10        | 0     | 0     | 1       |
| 01        | 0     | 0     | 1       |
| 00        | 0     | 0     | 1       |

\[ \tau = \{2, 21, 11, 01, 20, 10, 00\}: \text{(partition of the possible pasts)} \]
Characterizing the stochastic sequence

\[ \begin{align*}
2112112102112012112002\cdots
\end{align*}\]

(family of transition probabilities)

| context $w$ | $p(0|w)$ | $p(1|w)$ | $p(2|w)$ |
|-------------|-----------|-----------|-----------|
| 2           | $\epsilon$ | $1-\epsilon$ | 0         |
| 21          | $\epsilon$ | $1-\epsilon$ | 0         |
| 20          | $\epsilon$ | $1-\epsilon$ | 0         |
| 11          | 0         | 0         | 1         |
| 10          | 0         | 0         | 1         |
| 01          | 0         | 0         | 1         |
| 00          | 0         | 0         | 1         |

$\tau = \{2, 21, 11, 01, 20, 10, 00\}$: (partition of the possible pasts)

- This two objects characterize the stochastic simulation algorithm
Formal definition

- The string $u$ is said to be a **suffix** of $v$ ($u \preceq v$), if there exists a string $s$ satisfying $v = su$.

- A **context tree** is defined as any set $\tau \subset A^* = \bigcup_{m=1}^{\infty} A^m$ satisfying
  
  (i) **Suffix Property.** No string $w \in \tau$ is a proper suffix of another string $s \in \tau$.

  (ii) **Irreducibility.** No string belonging to $\tau$ can be replaced by a proper suffix without violating the suffix property.
The string \( u \) is said to be a **suffix** of \( v \) \( (u \preceq v) \), if there exists a string \( s \) satisfying \( v = su \).

**A context tree** is defined as any set \( \tau \subset A^* = \bigcup_{m=1}^{\infty} A^m \) satisfying

(i) **Suffix Property.** No string \( w \in \tau \) is a proper suffix of another string \( s \in \tau \).

(ii) **Irreducibility.** No string belonging to \( \tau \) can be replaced by a proper suffix without violating the suffix property.

We say \( (X_n)_n \) is a **context tree model** with parameters \( (\tau, p) \) if there exist a context function \( c_\tau : A^* \to \tau \) such that

(i) for any \( n \geq l(\tau) \) and any finite sequence \( x_{-n}^{n-1} \in A^n \) such that \( p(x_{-n}^{n-1}) > 0 \), it holds that

\[
P(X_{n+1} = a | X_1^n = x_{-n}^{n-1}) = p(X_{n+1} = a | c_\tau(x_{-n}^{n-1})) \quad \text{for all} \ a \in A.
\]

(ii) no proper suffix of \( c_\tau(x_{-n}^{n-1}) \) satisfies condition 2.
Elements of a context tree (the leaves) are called contexts.

Introduced by Rissanen as an universal system of data compression.

Also called:
- Stochastic chains with memory of variable length and
- Variable length Markov chains.
How we define the dependence of $Y_k$ in terms of $\ldots, X_{k-3}, X_{k-2}, X_{k-1}, X_k$?

The distribution of $Y_k$ should depend only on the context associated to this sequence of past stimulus.
Formally

The bivariate stochastic chain \((X_n, Y_n)_n\) taking values in \(A \times F\) is a sequence of random objects driven by context tree models with parameters \((\tau, p, q)\), where \(q = (q(\cdot \mid w) : w \in \tau)\) is a family of probability measures on \((F, \mathcal{F})\), if

(i) \((X_n)_n\) is a context tree model with parameters \((\tau, p)\);

(ii) Conditionally to the sequence \((X_n)_n\), \((Y_n)_n\) are independent random variables and, for any integers \(\ell(\tau) \leq m \leq n\), any string \(x_{m-\ell(\tau)+1}^n \in A^{n-m+\ell(\tau)}\) and any sequence \(J_m, \ldots, J_n\) of \(\mathcal{F}\)-measurable sets, it holds

\[
P(Y_m \in J_m, \ldots, Y_n \in J_n \mid X_{m-\ell(\tau)+1}^n = x_{m-\ell(\tau)+1}^n)
= \prod_{k=m}^{n} q(Y_k \in J_k \mid c_\tau(x_1^k))
\]
What about model selection?

- For CTM we have statistical procedures to $(\hat{\tau}, \hat{\rho})$ consistently.
  - Algorithm context (Rissanen 83)
  - BIC (Csiszár and Talata 06)
  - SMC (Galves, Galves, García, Garcia and Leonardi 12)

- How to check the time-dependency of the answer sequence?

- It depends on the type of answer data: functional or categorical
Model selection procedures

\[
\begin{align*}
\text{Functional} & \Rightarrow \text{Algorithm context + Projective method} \\
\text{Categorical} & \Rightarrow \begin{cases}
\text{Algorithm context + Conditional likelihood} \\
\text{Algorithm context + Offspring emp. distr.} \\
\text{BIC}
\end{cases} \\
& \Rightarrow SMC
\end{align*}
\]
Algorithm context

1. Set as candidate tree *the biggest* context tree possible
**Algorithm context**

1. Set as candidate tree *the biggest* context tree possible
2. Choose a final branch which has not been tested yet
Algorithm context

1. Set as candidate tree *the biggest* context tree possible
2. Choose a final branch which has not been tested yet
   2.1 Test the branch *statistically* following some criterion
   2.2 Use the criterion to prune or not the branch
3. Run step 2 until all branches are tested.
Algorithm context

1. Set as candidate tree *the biggest* context tree possible
2. Choose a final branch which has not been tested yet
   - 2.1 Test the branch *statistically* following some criterion
   - 2.2 Use the criterion to prune or not the branch
3. Run step 2 until all branches are tested.

- A subtree in $\tau$ induced by $s$ is defined as the set $\tau_s = \{w \in \tau : s \prec w\}$.
- A subtree is said *terminal* if for all $w \in \tau_s$ it holds that $w = as$ for some $a \in A$. 
The biggest context tree

Given a finite string \( u \in A^* \), denote the number of its occurrence by

\[
N^X_n(u) = \sum_{t=l(u)}^{n-1} 1\{X_{t-|u|+1} = u\}.
\]

An admissible context tree of maximum height \( L \) for the sample \((X_1, ..., X_n)\) is any context tree \( \tau \) satisfying

i) \( w \in \tau \) if and only if \( l(w) \leq L \) and \( N^X_n(w) \geq 1 \).

(any context appears at least once in the sample \( X \))

ii) Any string \( v \in A^* \) with \( N^X_n(v) \geq 1 \) is a suffix of some \( w \in \tau \) or has a suffix \( w \in \tau \).

(any string which appears in the sample \( X \) is represented in \( \tau \))

The set of all admissible context trees of maximal height \( L \): \( \Gamma^L(X_1^n) \).
Consider $F = L^2([0, T])$ and $\mathcal{F}$ the Borel $\sigma-$algebra on $L^2([0, T])$.

For any string $s \in A^*$ with $l(s) \leq L$, define

$$I_n(s) = \{l(s) \leq m \leq n : X^m_{m-l(s)+1} = s\}.$$  

By definition, the set $I_n(s)$ has $N^X_n(s)$ elements.

Set $Y^{(s)}_k$ the $k-$occurrence of $s$.

Thus $Y^{(s)} = \{Y^{(s)}_1, ..., Y^{(s)}_{N^X_n(s)}\}$ is the subsample of $(Y_1, ..., Y_n)$ induced by the string $s$. 

Consider $F = L^2([0, T])$ and $\mathcal{F}$ the Borel $\sigma$–algebra on $L^2([0, T])$.

For any string $s \in A^*$ with $l(s) \leq L$, define

$$I_n(s) = \{l(s) \leq m \leq n : X_{m-l(s)+1}^m = s\}. \quad (1)$$

By definition, the set $I_n(s)$ has $N_n^X(s)$ elements.

Set $Y_{k}^{(s)}$ the $k$–occurrence of $s$.

Thus $Y^{(s)} = \{Y_1^{(s)}, ..., Y_{N_n^X(s)}^{(s)}\}$ is the subsample of $(Y_1, ..., Y_n)$ induced by the string $s$. 

![Diagram showing occurrences of $Y_1^{112}$, $Y_2^{112}$, $Y_1^{012}$, $Y_2^{012}$ with annotations 450 ms]
Given a sample \((X_1, Y_1), \ldots, (X_n, Y_n)\).

1. Use \((X_1, \ldots, X_n)\) to construct a admissible context tree \(\tau\) of maximum height \(L\).
Given a sample \((X_1, Y_1), \ldots, (X_n, Y_n)\).

1. Use \((X_1, \ldots, X_n)\) to construct a admissible context tree \(\tau\) of maximum height \(L\).

2. Choose a terminal subtree \(\tau_u = \{w \in \tau : w \prec u\}\) which has not been tested yet.

3. Test null hypothesis.

\[
H_0^{(u)} : \mathcal{L}\left( Y_{1}^{(s)}, \ldots, Y_{N_n}^{(s)}(s) \right) = \mathcal{L}\left( Y_{1}^{(v)}, \ldots, Y_{N_n}^{(v)}(v) \right), \forall s, v \in \hat{\tau}_u,
\]

3.1 ▶ If the null hypothesis is not rejected, prune the branch.
   ▶ If the null hypothesis is rejected, keep the branch.

4. Run step 2 until all branches are tested.
Cuesta-Albertos (2006): If the law of $R_k^{(u)}$ and $R_k^{(v)}$ are different, so are the laws of $Y_k^{(u)}$ and $Y_k^{(v)}$. 
Step 3.

Given a string $u \in A^*$, we test

$$H_0^{(u)} : \mathcal{L}(Y^{(s)}) = \mathcal{L}(Y^{(v)}), \; \forall s, v \in \hat{\tau}_u,$$

using the test statistic

$$\Delta_n(u) = \Delta_n^W(u) = \max_{s, v \in \hat{\tau}_u} \sqrt{\frac{N_n^X(s)N_n^X(v)}{N_n^X(s) + N_n^X(v)}} \KS(\hat{Q}_n^{s, W}, \hat{Q}_n^{v, W}).$$

Here $W$ is a realization of a Brownian bridge in the interval $[0, T]$, $\KS(\hat{Q}_n^{s, W}, \hat{Q}_n^{v, W})$ is the KS distance between the empirical distributions $\hat{Q}_n^{s, W}$ and $\hat{Q}_n^{v, W}$ of the projected samples.
Given a string $u \in A^*$, we test

$$H_0^{(u)} : \mathcal{L}(Y^{(s)}) = \mathcal{L}(Y^{(v)}), \ \forall \ s, v \in \hat{\tau}_u,$$

using the test statistic

$$\Delta_n(u) = \Delta_n^W(u) = \max_{s, v \in \hat{\tau}_u} \sqrt{\frac{N_n^X(s)N_n^X(v)}{N_n^X(s) + N_n^X(v)}} \text{KS}(\hat{Q}_n^{s, W}, \hat{Q}_n^{v, W}).$$

Here $W$ is a realization of a Brownian bridge in the interval $[0, T]$, $\text{KS}(\hat{Q}_n^{s, W}, \hat{Q}_n^{v, W})$ is the KS distance between the empirical distributions $\hat{Q}_n^{s, W}$ and $\hat{Q}_n^{v, W}$ of the projected samples.

- If $\Delta_n(u) < c$, prune the brunch
- Otherwise, keep the brunch
Hypothesis

Let $P$ be a probability measure on $(F, \mathcal{F})$. Suppose that $P$ satisfies

**H1.** The *Carleman condition*: all the absolute moments
\[ m_k = \int \|h\|^k P(dh), \quad k \geq 1, \]
are finite and
\[ \sum_{k \geq 1} m_k^{-1/k} = +\infty. \]

**H2.** *Continuity*: for any $h \in F$, where $P^h$ is defined by
\[ P^h((\infty, t]) = P(x \in F : \langle x, h \rangle \leq t), \quad t \in \mathbb{R}. \]

**H3.** The family $q$ is *identifiable*: for any context $s \in \tau$ there exits a pair $u, v \in \tau_w$ such that and $q(\cdot | v) \neq q(\cdot | u)$. 
Theorem: DFGOV 19

Let \((X_0, Y_0), \ldots, (X_n, Y_n)\) be a sample produced by sequence of random objects driven by a context tree model compatible with \((\tau^*, p^*, q^*)\), and let \(\hat{\tau}_n\) be the context tree selected from the sample by Algorithm with \(L \geq \ell(\tau^*)\) and threshold \(c_{\alpha_n} = \sqrt{(1/2)\ln(2/\alpha_n)}\), where \(\alpha_n \in (0, 1)\). If \((\tau^*, p^*)\) is irreducible and \(q^*\) is continuous and satisfies Carleman condition, then for \(\alpha_n \to 0\) slowly enough as \(n \to \infty\),

\[
\lim_{n \to \infty} P(\hat{\tau}_n \neq \tau^*) = 0.
\]
Categorical case

▶ Straight adaptation from single to bivariate case.
Categorical case

- Straight adaptation from single to bivariate case.

- Number of occurrences of the string $u$ in the sample $(X_1, \ldots, X_n)$ followed by the occurrence of the symbol $a$ in the sample $(Y_1, \ldots, Y_n)$

$$N_{n}^{XY}(u, a) = \sum_{t=l(u)}^{n-1} 1\{X_{t-l(u)+1} = u; Y_{t+1} = a\}.$$ 

- Empirical conditional transition probability

$$\hat{q}(a|u) = \frac{N_{n}^{XY}(u, a)}{N_{n}^{X}(u)}.$$
Statistic for the algorithm context

- The conditional likelihood for a sample \((X_1, Y_1), \ldots, (X_n, Y_n)\)

\[
L_{(\tau, \hat{q})}(X^n, Y^n) = \prod_{u \in \tau} \prod_{a \in A} \hat{q}(a | u)^{N_n^{XY}(u, a)},
\]

- Define the statistics

\[
\bar{\Delta}_n(u) = \sum_{b \in A} \sum_{a \in A} N_n^{XY}(bu, a) \log \frac{\hat{q}(a | bu)}{\hat{q}(a | u)},
\]
The conditional likelihood for a sample \((X_1, Y_1), \ldots, (X_n, Y_n)\)

\[
L(\tau, \hat{q})(Y_1^n \mid X_1^n) = \prod_{u \in \tau} \prod_{a \in A} \hat{q}(a \mid u)^{N_{n}^{XY}(u, a)},
\]

Define the statistics

\[
\bar{\Delta}_n(u) = \sum_{b \in A} \sum_{a \in A} N_{n}^{XY}(bu, a) \log \frac{\hat{q}(a \mid bu)}{\hat{q}(a \mid u)},
\]

and

\[
\tilde{\Delta}_n(u) = \max_{b \in A} \left( \max_{a \in A} |\hat{q}(a \mid u) - \hat{q}(a \mid bu)| \right).
\]
The conditional likelihood for a sample \((X_1, Y_1), \ldots, (X_n, Y_n)\)

\[
L(\tau, \hat{q})(Y_1^n \mid X_1^n) = \prod_{u \in \tau} \prod_{a \in A} \hat{q}(a \mid u)^{N_n^{XY}(u,a)},
\]

Define the statistics

\[
\bar{\Delta}_n(u) = \sum_{b \in A} \sum_{a \in A} N_n^{XY}(bu, a) \log \frac{\hat{q}(a \mid bu)}{\hat{q}(a \mid u)},
\]

and

\[
\tilde{\Delta}_n(u) = \max_{b \in A} \left( \max_{a \in A} |\hat{q}(a \mid u) - \hat{q}(a \mid bu)| \right).
\]

Fix a threshold \(\delta > 0\) and denote by \(\Delta\) either \(\bar{\Delta}\) or \(\tilde{\Delta}\)

(i) If \(\Delta_n(u) < \delta\) we prune the subtree \(\hat{\tau}_u\).

(ii) If \(\Delta_n(u) \geq \delta\) we keep \(\hat{\tau}_u\).
Formally

Set \( C_w((X, Y)_1^n) = 0 \) for all \( w \in \mathcal{T}_n^L \), and, for any \( u < w \in \mathcal{T}_n^L \) define

\[
C_{u,n} = \max \left\{ 1_{\{\Delta_n(u) \geq \delta\}}, \max_{b \in A} C_{bu,n} \right\}.
\]

The context tree estimator \( \hat{\tau}_{\mathcal{C}}^n = \hat{\tau}_{\mathcal{C}}^n((X, Y)_1^n) \) is given by

\[
\hat{\tau}_{\mathcal{C},n} = \{ w \preceq v \in \mathcal{T}_n^L : C_{w,n} = 0 \text{ and } C_{u,n} = 1 \text{ for all } u < w \}.
\]
Formally

Set $C_w((X, Y)_1^n) = 0$ for all $w \in \mathcal{T}_n^L$, and, for any $u \prec w \in \mathcal{T}_n^L$ define

$$C_{u, n} = \max \left\{ 1_{\{\Delta_n(u) \geq \delta\}}, \max_{b \in A} C_{bu, n} \right\}.$$

The context tree estimator $\hat{\tau}_{C, n} = \hat{\tau}_{C}((X, Y)_1^n)$ is given by

$$\hat{\tau}_{C, n} = \{ w \preceq v \in \mathcal{T}_n^L : C_{w, n} = 0 \text{ and } C_{u, n} = 1 \text{ for all } u \prec w \}.$$

Consistency

Let $(X_0, Y_0), \ldots, (X_n, Y_n)$ be a sample produced by sequence of random objects driven by a context tree model compatible with $(\tau^*, p^*, q^*)$. The context tree $\hat{\tau}^\delta_{C, n}$ selected from the sample by Algorithm, with $\ell(\tau^*) < L < \log(n)$, eventually almost surely as $n \to \infty$, satisfies

$$\hat{\tau}^\delta_{C, n} = \tau^*$$
Bayesian Information Criterion (BIC)

- The BIC estimator is defined as

\[
\hat{\tau}_{BIC,n}^{c} = \arg\max_{\tau \in \Gamma_n} \left\{ \log L(\tau, \hat{q})(Y_1^n \mid X_1^n) - c \, df(\tau) \log(n) \right\}.
\]

where \(df\) stands for the degree of freedom of the model.
The BIC estimator is defined as

\[
\hat{\tau}_{\text{BIC},n} = \arg\max_{\tau \in \Gamma_n} \left\{ \log L_{(\tau, \hat{q})}(Y_1^n \mid X_1^n) - c \ df(\tau) \log(n) \right\}.
\]

where \( df \) stands for the degree of freedom of the model.

For any \( w \in \mathcal{T}_n^L \), define \( V_{w,n} = n^{-c \cdot df(w)} L_{(w, \hat{q})}(Y_1^n \mid X_1^n) \) and the indicator \( \chi_{w,n} = 0 \)

For any \( w \prec u \in \mathcal{T}_n^L \) define recursively the quantity

\[
V_{w,n} = \max \left\{ n^{-c \cdot df(w)} L_{(w, \hat{q})}(Y_1^n \mid X_1^n), \prod_{b \in A} V_{bw,n} \right\}
\]

and the indicator

\[
\chi_{w,n} = 1 \left\{ \prod_{b \in A} V_{bw,n} > n^{-c \cdot df(w)} L_{(w, \hat{q})}(Y_1^n \mid X_1^n) \right\}.
\]
Bayesian Information Criterion (BIC)

- The BIC estimator is defined as

\[
\hat{\tau}_\text{BIC}, n = \arg\max_{\tau \in \Gamma_n^L} \left\{ \log L(\tau, \hat{q}) (Y_1^n | X_1^n) - c \, df(\tau) \log(n) \right\}.
\]

where \(df\) stands for the degree of freedom of the model.

- For any \(w \in \mathcal{T}_n^L\), define \(V_{w,n} = n^{-c \cdot df(w)} L(w, \hat{q}) (Y_1^n | X_1^n)\) and the indicator \(\chi_{w,n} = 0\)

- For any \(w \prec u \in \mathcal{T}_n^L\) define recursively the quantity

\[
V_{w,n} = \max \left\{ n^{-c \cdot df(w)} L(w, \hat{q}) (Y_1^n | X_1^n), \prod_{b \in A} V_{bw,n} \right\}
\]

and the indicator

\[
\chi_{w,n} = 1\left\{ \prod_{b \in A} V_{bw,n} > n^{-c \cdot df(w)} L(w, \hat{q}) (Y_1^n | X_1^n) \right\}.
\]

- The estimated context tree can be written as

\[
\hat{\tau}_\text{BIC}, n = \{ w \preceq s \in \mathcal{T}_n^L : \chi_{w,n} = 0 \text{ and } \chi_{u,n} = 1 \text{ for all } u \prec w \}.
\]
Consistency

Let \((X_0, Y_0), \ldots, (X_n, Y_n)\) be a sample produced by sequence of random objects driven by a context tree model compatible with \((\tau^*, p^*, q^*)\). The context tree \(\hat{\tau}^c_{BIC,n}\) selected from the sample by BIC procedure, with \(\ell(\tau^*) < L < \log(n)\), eventually almost surely as \(n \to \infty\), satisfies

\[
\hat{\tau}^c_{BIC,n} = \tau^*
\]
The tree estimation procedure presented are constant dependent

\[
\bar{\Delta}_n(u) = \sum_{b \in A} \sum_{a \in A} N^{XY}_n(bu, a) \log \frac{\hat{q}(a|bu)}{\hat{q}(a|u)} < \delta
\]

\[
\tilde{\Delta}_n(u) = \max_{b \in A} \left( \max_{a \in A} |\hat{q}(a|u) - \hat{q}(a|bu)| \right) < \delta
\]

\[
\arg\max_{\tau \in \Gamma_n} \left\{ \log L(\tau, \hat{q})(Y_1^n \mid X_1^n) - c \cdot df(\tau) \log(n) \right\}.
\]
(i) Construct a set of candidate models: champion trees.
(ii) Choose an optimal model within the set of champion trees.

The champion trees obtained will depend on the model selection procedure.
Smallest Maximizer Criterion

(i) Construct a set of candidate models: champion trees.
(ii) Choose a optimal model within the set of champion trees.

The champion trees obtained will depend on the model selection procedure

Compute the champion trees:
Denote by $\ell$ the constant either in the algorithm context or BIC. Start with $\ell = 0$ and successively increase the value of $\ell$
Theorem: General SMC

Assume \((X_1, Y_1), \ldots , (X_n, Y_n)\) is a sample of a sequence of random objects driven by a context tree model with parameters \((\tau^*, p^*, q^*)\). Given a set \(C_n \subset \Gamma^L(X_1^n)\), with \(|\tau^*| \leq L < \log(n)\), satisfying

(i) \(C_n\) is totally ordered with respect to \(\succ\) and
(ii) eventually almost surely \(\tau^* \in C_n\) as \(n \to \infty\).

The following holds:

1. For any \(\tau \in C_n\), with \(\tau \prec \tau^*\), there exists a constant \(c(\tau^*, \tau) > 0\) such that
   \[
   \log L(\tau^*, \hat{q}) - \log L(\tau, \hat{q}) \geq c(\tau^*, \tau)n
   \] (2)

2. For any \(\tau \prec \tau' \in C_n\), with \(\tau^* \preceq \tau\), there exists a constant \(c(\tau', \tau) > 0\) such that
   \[
   \log L(\tau', \hat{q}) - \log L(\tau, \hat{q}) \leq c(\tau', \tau) \log n
   \] (3)
The tree model selection procedures satisfy the Thm

Assume \((X_1, Y_1), \cdots, (X_n, Y_n)\) is a sample of a sequence of random objects driven by context tree model with parameters \((\tau^*, p^*, q^*)\).

Consider the map \(\ell \in [0, +\infty] \mapsto \hat{\tau}_n^\ell \in \Gamma^L(X_1^n)\) with \(|\tau^*| \leq L < \log(n)\)
and \(\hat{\tau}_n^\ell\) denoting either \(\hat{\tau}_C^\ell((X, Y)_1^n)\), \(\hat{\tau}_{\tilde{C}}^\ell((X, Y)_1^n)\) or \(\hat{\tau}_{BIC}^\ell((X, Y)_1^n)\)

Denote by

\[ C_n = \{\hat{\tau}_n^\ell : \ell \in [0, +\infty]\}. \]

The set \(C_n\) is totally ordered with respect to \(\succ\) and eventually almost surely \(\tau^* \in C_n\) as \(n \to \infty\).
Model selection procedures future work

\[
\begin{aligned}
\text{Functional} \Rightarrow & \quad \{ \\
\text{Algorithm context} + \text{Projective method} \\
\text{Algorithm context} + \text{Other statistics} \\
\} \\
\text{Categorical} \Rightarrow & \quad \{ \\
\text{Algorithm context} + \text{Conditional likelihood} \\
\text{Algorithm context} + \text{Offspring emp. distr.} \\
\text{BIC} \\
\text{Other Statistics} \\
\} \Rightarrow \text{SMC}
\end{aligned}
\]

Application in neurobiological data:
