Hidden context tree modeling of EEG data

Antonio Galves
joint work with A. Duarte, R. Fraiman, G. Ost and C. Vargas

Universidade de S.Paulo and NeuroMat

MathStatNeuro 2015
Looking for experimental evidence that the brain is a statistician

- Is the brain a statistician?
- Stanislas Dehaene claims that the idea that the brain is a Bayesian statistician is already sketched in von Helmholtz work!
- See for instance the two lessons by Dehaene available on the web:
Looking for experimental evidence that the brain is a statistician

▶ Is the brain a statistician?
▶ Stanislas Dehaene claims that the idea that the brain is a Bayesian statistician is already sketched in von Helmholtz work!
▶ See for instance the two lessons by Dehaene available on the web:
  ▶ Le cerveau statisticien
Looking for experimental evidence that the brain is a statistician

- Is the brain a statistician?
- Stanislas Dehaene claims that the idea that the brain is a Bayesian statistician is already sketched in von Helmholtz work!
- See for instance the two lessons by Dehaene available on the web:
  - Le cerveau statisticien
  - Le bébé statisticien
Is the brain a statistician?

- How to obtain experimental evidence supporting this conjecture?
- Dehaene presents experimental evidence that unexpected occurrences in regular sequences produce characteristic markers in EEG data.
- But we need more than evidences of mismatch negativity to support this conjecture.
- To discuss this issue we need to do statistical model selection in a new class of stochastic processes:
Is the brain a statistician?

- How to obtain experimental evidence supporting this conjecture?
- Dehaene presents experimental evidence that unexpected occurrences in regular sequences produce characteristic markers in EEG data.
- But we need more than evidences of mismatch negativity to support this conjecture.
- To discuss this issue we need to do statistical model selection in a new class of stochastic processes:
  
  Hidden context tree models.
Neurobiological problem

A random source produces sequences of auditory stimuli.

How to retrieve the structure of the source from EEG data?
Example of a random source: samba

▶ Auditory segments:
  ▶ 2 - strong beat
  ▶ 1 - weak beat
  ▶ 0 - silent event

▶ Chain generation:
  ▶ start with a deterministic sequence
    \[ \cdots 2 \ 1 \ 0 \ 1 \ 2 \ 1 \ 0 \ 1 \ 2 \ \cdots \]
  ▶ replace in a iid way each symbol 1 by 0 with probability $\epsilon$. 
A typical sample would be

\[
\cdots 2 \ 1 \ 0 \ 1 \ 2 \ 1 \ 0 \ 1 \ 2 \ 1 \ 0 \ 1 \ 2 \cdots \\
\cdots 2 \ 1 \ 0 \ 0 \ 2 \ 1 \ 0 \ 1 \ 2 \ 0 \ 0 \ 0 \ 2 \cdots 
\]
A typical sample would be

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

How to define the structure of this source?
A typical sample would be

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

How to define the structure of this source?

- By describing the algorithm producing each next symbol, given the \textit{shortest relevant} sequence of past symbols.
The structure of the random source
The structure of the random source

\[
\begin{array}{cccc}
\text{Contexts (w)} & p(0|w) & p(1|w) & p(2|w) \\
2 & \epsilon & 1 - \epsilon & 0 \\
21 & \epsilon & 1 - \epsilon & 0 \\
01 & 0 & 0 & 1 \\
20 & \epsilon & 1 - \epsilon & 0 \\
10 & \epsilon & 1 - \epsilon & 0 \\
000 & 0 & 0 & 1 \\
100 & 0 & 0 & 1 \\
200 & \epsilon & 1 - \epsilon & 1 \\
\end{array}
\]
The structure of the random source

\[
\begin{array}{c|ccc|c}
\text{Contexts}(w) & p(0|w) & p(1|w) & p(2|w) \\
\hline
2 & \epsilon & 1 - \epsilon & 0 \\
21 & \epsilon & 1 - \epsilon & 0 \\
01 & 0 & 0 & 1 \\
20 & \epsilon & 1 - \epsilon & 0 \\
10 & \epsilon & 1 - \epsilon & 0 \\
000 & 0 & 0 & 1 \\
100 & 0 & 0 & 1 \\
200 & \epsilon & 1 - \epsilon & 1 \\
\end{array}
\]
The stochastic chain generated by the source samba

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

\[ X_{-5} \, X_{-4} \, X_{-3} \, X_{-2} \, X_{-1} \, X_0 \, X_1 \, X_2 \]
$X_n \in A = \{0, 1, 2\}$
\( X_n \in A = \{0, 1, 2\} \)

\((X_n)_{n \in \mathbb{Z}}\) is stochastic chain
\( X_n \in A = \{0, 1, 2\} \)

\((X_n)_{n \in \mathbb{Z}}\) is stochastic chain

with memory of variable length
- $X_n \in A = \{0, 1, 2\}$
- $(X_n)_{n \in \mathbb{Z}}$ is stochastic chain
- with memory of variable length
- generated by the probabilistic context tree
\( X_n \in A = \{0, 1, 2\} \)

\((X_n)_{n \in \mathbb{Z}}\) is stochastic chain

with memory of variable length

generated by the probabilistic context tree

| Contexts(w) | \( p(0|w) \) | \( p(1|w) \) | \( p(2|w) \) |
|------------|-------------|-------------|-------------|
| 2          | \( \epsilon \) | \( 1 - \epsilon \) | 0           |
| 21         | \( \epsilon \) | \( 1 - \epsilon \) | 0           |
| 01         | 0           | 0           | 1           |
| 20         | \( \epsilon \) | \( 1 - \epsilon \) | 0           |
| 10         | \( \epsilon \) | \( 1 - \epsilon \) | 0           |
| 000        | 0           | 0           | 1           |
| 100        | 0           | 0           | 1           |
| 200        | \( \epsilon \) | \( 1 - \epsilon \) | 1           |
Context tree models

- Introduced by Rissanen
Context tree models

- Introduced by Rissanen
Context tree models

- Introduced by Rissanen
- stochastic chains with memory of variable length
Context tree models

- Introduced by Rissanen
- stochastic chains with memory of variable length
- generated by a probabilistic context tree

![Diagram of a context tree model]

000 100 200
10 20 01 21
2
The neurobiological question

Is it possible
The neurobiological question

Is it possible
to retrieve the samba context tree
The neurobiological question

Is it possible

to retrieve the samba context tree

from the EEG data recorded during the exposure to
the sequence of auditory stimuli generated by the
samba source?
EEG data
How to address the identification problem?

We have

\[ Y_{en} = \left( Y_{en}(t), t \in [0, T] \right) \]

the EEG signal recorded at electrode \( e \) during the exposure to the auditory stimulus \( X_n \).

\[ Y_{en} \in L^2([0, T]), \quad T = 450 \text{ ms} \]

is the time distance between the onsets of two consecutive auditory stimuli.
How to address the identification problem?

We have

- EEG data recorded with 18 electrodes
How to address the identification problem?

We have

- EEG data recorded with 18 electrodes
- for each electrode $e$ and each step $n$
How to address the identification problem?

We have

- EEG data recorded with 18 electrodes
- for each electrode $e$ and each step $n$
- call $Y^e_n = (Y^e_n(t), t \in [0, T])$
How to address the identification problem?

We have

- EEG data recorded with 18 electrodes
- for each electrode \( e \) and each step \( n \)
- call \( Y^e_n = (Y^e_n(t), t \in [0, T]) \) the EEG signal recorded at electrode \( e \) during the exposure to the auditory stimulus \( X_n \)
- \( Y^e_n \in L^2([0, T]) \), where \( T = 450 \text{ms} \) is the time distance between the onsets of two consecutive auditory stimuli
Hidden context tree model (HCTM)

Ingredients:

- finite alphabet $A$

In our example $A = \{0, 1, 2\}$

- measurable space $(F, \mathcal{F})$

  In our example $\mathcal{F} = L^2([0,T])$ and $\mathcal{F}$ is the Borel $\sigma$-algebra on $F$.

- probabilistic context tree $(\tau, p)$

- family $\{Q_w: w \in \tau\}$ of probabilities on $(F, \mathcal{F})$

- stochastic chain $(X_n, Y_n) \in A \times F$.

Antonio Galves joint work with A. Duarte, R. Fraiman, G. Ost and C. Vargas

Hidden context tree modeling of EEG data
Hidden context tree model (HCTM)

Ingredients:

- finite alphabet $A$
  
  In our example $A = \{0, 1, 2\}$
Hidden context tree model (HCTM)

Ingredients:

- finite alphabet $A$
  
  In our example $A = \{0, 1, 2\}$

- measurable space $(F, \mathcal{F})$
Hidden context tree model (HCTM)

Ingredients:

- finite alphabet $A$
  
  In our example $A = \{0, 1, 2\}$

- measurable space $(F, \mathcal{F})$
  
  In our example $F = L^2([0, T])$ and $\mathcal{F}$ is the Borel $\sigma$-algebra on $F$. 

Antonio Galves joint work with A. Duarte, R. Fraiman, G. Ost and C. Vargas

Hidden context tree modeling of EEG data
Hidden context tree model (HCTM)

Ingredients:

- finite alphabet $A$
  - In our example $A = \{0, 1, 2\}$
- measurable space $(F, \mathcal{F})$
  - In our example $F = L^2([0, T])$ and $\mathcal{F}$ is the Borel $\sigma$-algebra on $F$.
- probabilistic context tree $(\tau, p)$
Hidden context tree model (HCTM)

Ingredients:

- finite alphabet $A$
  
  In our example $A = \{0, 1, 2\}$

- measurable space $(F, \mathcal{F})$
  
  In our example $F = L^2([0, T])$ and $\mathcal{F}$ is the Borel $\sigma$-algebra on $F$.

- probabilistic context tree $(\tau, p)$

- family $\{Q_w : w \in \tau\}$ of probabilities on $(F, \mathcal{F})$
Hidden context tree model (HCTM)

Ingredients:

- finite alphabet $A$
  
  In our example $A = \{0, 1, 2\}$

- measurable space $(F, \mathcal{F})$
  
  In our example $F = L^2([0, T])$ and $\mathcal{F}$ is the Borel $\sigma$-algebra on $F$.

- probabilistic context tree $(\tau, p)$

- family $\{Q_w : w \in \tau\}$ of probabilities on $(F, \mathcal{F})$

- stochastic chain $(X_n, Y_n) \in A \times F$. 
Hidden context tree model

\[(X_n, Y_n)_{n \in \mathbb{Z}}\] HCTM compatible with \((\tau, p)\) and \((Q_w : w \in \tau)\) if

\[\text{any string} \quad x_{n-m}^m \in \mathcal{A}^{n-m+\ell(\tau)} \quad \text{and any sequence} \quad I_{n-m} = (I_m, ..., I_n) \quad \text{of} \quad \mathcal{F} - \text{measurable sets,}

\[P(Y_{n-m} \in I_{n-m} | X_{n-m}^{m-\ell(\tau)} = x_{n-m}^{m-\ ell(\tau)}) = \prod_{k=m}^{n} Q_{c_{\tau}}(x_k^{k-\ell(\tau)})^{I_k(\tau)}\]

\[\ell(\tau) = \text{height of} \quad \tau \quad \text{and} \quad c_{\tau}(x_{k}^{k-\ell(\tau)}) = \text{context assigned to} \quad x_{k}^{k-\ell(\tau)} \quad \text{by} \quad \tau\]
Hidden context tree model

\((X_n, Y_n)_{n \in \mathbb{Z}}\) HCTM compatible with \((\tau, p)\) and \((Q_w : w \in \tau)\) if

- \((X_n)_{n \in \mathbb{Z}}\) is generated by \((\tau, p)\)
Hidden context tree model

\((X_n, Y_n)_{n \in \mathbb{Z}}\) HCTM compatible with \((\tau, p)\) and \((Q_w : w \in \tau)\) if

- \((X_n)_{n \in \mathbb{Z}}\) is generated by \((\tau, p)\)
- for any \(m, n \in \mathbb{Z}\) with \(m \leq n\)
Hidden context tree model

\((X_n, Y_n)_{n \in \mathbb{Z}}\) HCTM compatible with \((\tau, p)\) and \((Q_w : w \in \tau)\) if

- \((X_n)_{n \in \mathbb{Z}}\) is generated by \((\tau, p)\)
- for any \(m, n \in \mathbb{Z}\) with \(m \leq n\)
- any string \(x_{m-\ell(\tau)+1}^n \in A^{n-m+\ell(\tau)}\)
Hidden context tree model

\((X_n, Y_n)_{n \in \mathbb{Z}}\) HCTM compatible with \((\tau, p)\) and \((Q_w : w \in \tau)\) if

- \((X_n)_{n \in \mathbb{Z}}\) is generated by \((\tau, p)\)
- for any \(m, n \in \mathbb{Z}\) with \(m \leq n\)
- any string \(x_{m-\ell(\tau)+1}^n \in A^{n-m+\ell(\tau)}\)
- and any sequence \(I_m^n = (I_m, \ldots, I_n)\) of \(\mathcal{F}\)-measurable sets,
Hidden context tree model

\((X_n, Y_n)_{n \in \mathbb{Z}}\) HCTM compatible with \((\tau, p)\) and \((Q_w : w \in \tau)\) if

1. (\(X_n)_{n \in \mathbb{Z}\) is generated by \((\tau, p)\)

2. for any \(m, n \in \mathbb{Z}\) with \(m \leq n\)

3. any string \(x_{m-\ell(\tau)+1}^n \in A_{n-m+\ell(\tau)}\)

4. and any sequence \(I^n_m = (I_m, \ldots, I_n)\) of \(\mathcal{F}\)-measurable sets,

\[
P\left(Y^n_m \in I^n_m \mid X^n_{m-\ell(\tau)+1} = x_{m-\ell(\tau)+1}^n\right) = \prod_{k=m}^n Q_{c_{\tau}(x_{k-\ell(\tau)+1}^k)}(I_k)
\]

5. \(\ell(\tau) = \) height of \(\tau\)

6. \(c_{\tau}(x_{k-\ell(\tau)+1}^k) = \) context assigned to \(x_{k-\ell(\tau)+1}^k\) by \(\tau\)
Rephrasing our problem

- A sequence of auditory stimuli produced by the samba source:
  \( (X_n) \in \mathbb{Z} \)

- Successive chunks of EEG signals:
  \( (Y_n) \in \mathbb{Z} \)

Question: Is \((X_n, Y_n) \in \mathbb{Z} \) a HCTM compatible with \( \tau \)?
Rephrasing our problem

Taking
Rephrasing our problem

Taking

- \((X_n)_{n \in \mathbb{Z}}\) sequence of auditory stimuli produced by the samba source.
Rephrasing our problem

Taking

- \((X_n)_{n \in \mathbb{Z}}\) sequence of auditory stimuli produced by the samba source
- \((Y^n_e)_{n \in \mathbb{Z}}\) successive chunks of EEG signals
Rephrasing our problem

Taking

- \((X_n)_{n \in \mathbb{Z}}\) sequence of auditory stimuli produced by the samba source
- \((Y^e_n)_{n \in \mathbb{Z}}\) successive chunks of EEG signals

Question: Is \((X_n, Y^e_n)_{n \in \mathbb{Z}}\) a HCTM compatible with \(\tau\)?
is $(X_n, Y_n^e)_{n \in \mathbb{Z}}$ a HCTM compatible with $\tau$?

In other terms, for any $w \in \tau$, is it true that

$$\mathcal{L}(Y_n^e | X_{n-\ell(w)+1} = w, X_{-\infty}^{-\ell(\tau)} = u) = \mathcal{L}(Y_n^e | X_{n-\ell(w)+1} = w, X_{-\infty}^{-\ell(\tau)} = v)$$

for any pair of strings $u$ and $v$?
Pruning the tree

- A version of Rissanen’s algorithm Context will be applied
Pruning the tree

▶ A version of Rissanen’s algorithm Context will be applied
▶ Start with a maximal admissible candidate tree
Pruning the tree

- A version of Rissanen’s algorithm Context will be applied
- Start with a maximal admissible candidate tree
- For any string $w$ and pair of symbols $a, b \in A$ with $aw$ and $bw$ belonging to the candidate tree
- test the equality

$$\mathcal{L}(Y^n_{n|X^{n-\ell(w)}_n} = aw) = \mathcal{L}(Y^n_{n|X^{n-\ell(w)}_n} = bw)$$
Pruning the tree

- If for all pairs of symbols \((a, b)\) the equality is rejected then prune all the leaves \(aw\)

- Repeat the pruning procedure until no more pruning is required
How to test the equality

\[ \mathcal{L}(Y_n|X_{n-\ell(w)}^n = aw) = \mathcal{L}(Y_n|X_{n-\ell(w)}^n = bw) \]?

Apply the projective method introduced by Cuestas-Albertos, Fraiman and Ransford (2006).
Experimental results

- Context tree selection procedure for the EEG data recorded during the exposure to the sequence of auditory stimuli generated by the samba source
- Sample composed by 20 subjects
- For each subject EEG data from 18 electrodes was recorded
Experimental results

Hidden context tree modeling of EEG data
Experimental results

Hidden context tree modeling of EEG data
White nodes indicate the number of subjects which correctly identify the node as not being a context. Black nodes indicate the number of subjects which correctly identify the node as a context. For instance, 18 subjects correctly identify that the symbol 0 alone is not enough to predict the next symbol. And 15 subjects correctly identify the symbol 2 as a context.