Motivation

〜 Inference for discrete biological time series is often hard
Inference for discrete biological time series is often hard

Difficulty: “Memory” modelling
E.g. for a binary time series with memory length of only 20 bits $2^{20}$ parameters must be estimated before even getting started

Need A LOT of data
Motivation

~ Motivation for discrete biological time series is often hard

~ Difficulty: “Memory” modelling
  E.g. for a binary time series with memory length of only 20 bits $2^{20}$ parameters must be estimated before even getting started

~ Need A LOT of data

~ Difficulty: Big Data
  Most existing methods do not realistically scale with large data
  Even “Big Data” are not enough for classical estimation

~ Need smarter, parsimonious models
Earlier Work

∼ Rissanen’s 1983-1986 fundamental work on the Minimum Description Length (MDL) principle and the introduction of tree/FSMX sources

∼ The basic results of Willems et al 1995-2000 on data compression via Context Tree Weighting (CTW) and related algorithms

∼ Classical inferential procedures and asymptotics of Bühlmann et al’s 1999-2004 Variable-Memory Markov chains (VLMC)
Markov chain \{\ldots, X_0, X_1, \ldots\} with alphabet \( A = \{0, 1, \ldots, m - 1\} \) of size \( m \)
Markov chain \{\ldots, X_0, X_1, \ldots\} with alphabet \( A = \{0, 1, \ldots, m - 1\} \)
of size \( m \)

Memory length \( d \)
\[
P(X_n|X_{n-1}, X_{n-2}, \ldots) = P(X_n|X_{n-1}, X_{n-2}, \ldots, X_{n-d})
\]
Fixed- and Variable-Memory Markov Chain Models

<table>
<thead>
<tr>
<th>Markov chain</th>
<th>{\ldots, X_0, X_1, \ldots} with alphabet ( A = {0, 1, \ldots, m - 1} ) of size ( m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory length ( d )</td>
<td>( P(X_n</td>
</tr>
<tr>
<td>Distribution</td>
<td>To fully describe it, we need to specify ( m^d ) conditional distributions ( P(X_n</td>
</tr>
</tbody>
</table>
# Fixed- and Variable-Memory Markov Chain Models

**Markov chain** \(\{\ldots, X_0, X_1, \ldots\}\) with **alphabet** \(A = \{0, 1, \ldots, m - 1\}\) of size \(m\)

**Memory length** \(d\) \(P(X_n|X_{n-1}, X_{n-2}, \ldots) = P(X_n|X_{n-1}, X_{n-2}, \ldots, X_{n-d})\)

**Distribution** To fully describe it, we need to specify \(m^d\) conditional distributions \(P(X_n|X_{n-1}, \ldots, X_{n-d})\) one for each context \((X_{n-1}, \ldots, X_{n-d})\)

**Problem** \(m^d\) grows very fast, e.g., with \(m = 8\) symbols and memory length \(d = 10\), we need \(\approx 10^9\) distributions
Fixed- and Variable-Memory Markov Chain Models

**Markov chain** \( \{ \ldots, X_0, X_1, \ldots \} \) with alphabet \( A = \{0, 1, \ldots, m - 1\} \) of size \( m \)

**Memory length** \( d \) \[ P(X_n | X_{n-1}, X_{n-2}, \ldots) = P(X_n | X_{n-1}, X_{n-2}, \ldots, X_{n-d}) \]

**Distribution** To fully describe it, we need to specify \( m^d \) conditional distributions \( P(X_n | X_{n-1}, \ldots, X_{n-d}) \) one for each context \( (X_{n-1}, \ldots, X_{n-d}) \)

**Problem** \( m^d \) grows very fast, e.g., with \( m = 8 \) symbols and memory length \( d = 10 \), we need \( \approx 10^9 \) distributions

**Idea** Use *variable length contexts* described by a *context tree* \( T \)
Variable-Memory Markov Chains: An Example

Alphabet $m = 3$ symbols
Memory length $d = 5$

Each past string $X_{n-1}, X_{n-2}, \ldots$ corresponds to a unique context on a leaf of the tree

Model: context tree $T$
Variable-Memory Markov Chains: An Example

Alphabet \( m = 3 \) symbols
Memory length \( d = 5 \)

Each past string \( X_{n-1}, X_{n-2}, \ldots \)
corresponds to a unique context
on a leaf of the tree

**Parameters:** \( \theta = \{ \theta_s ; s \in T \} \)
The distr of \( X_n \) given the past
is given by the distr on that leaf

E.g. \( P(X_n = 1 | X_{n-1} = 0, X_{n-2} = 2, X_{n-3} = 1, \ldots) = \theta_{022}(1) \)
Variable-Memory Representation: Advantages

\[ \sim \textbf{Parsimony} \quad \text{E.g. above with memory length 5} \]
\[ \text{instead of } 3^5 = 243 \text{ conditional distributions, only need to specify 13} \]

\[ \sim \text{For an alphabet of size } m \text{ and memory depth } d \text{ there are } m^d \text{ contexts} \]
\[ \Rightarrow \text{potentially huge savings} \]
Variable-Memory Representation: Advantages

~\rightarrow \textbf{Parsimony} \; \text{E.g.} \; \text{above with memory length 5}
\text{instead of } 3^5 = 243 \text{ conditional distributions, only need to specify 13}

~\rightarrow \text{For an alphabet of size } m \text{ and memory depth } d \text{ there are } m^d \text{ contexts}
\Rightarrow \text{potentially huge savings}

~\rightarrow \text{Determining the underlying context tree of an empirical time series is of great scientific and engineering interest}
Variable-Memory Representation: Advantages

〜 Parsimony  E.g. above with memory length 5 instead of $3^5 = 243$ conditional distributions, only need to specify 13

〜 For an alphabet of size $m$ and memory depth $d$ there are $m^d$ contexts ⇒ potentially huge savings

〜 Determining the underlying context tree of an empirical time series is of great scientific and engineering interest

Applications

<table>
<thead>
<tr>
<th>Model selection</th>
<th>Estimation</th>
<th>Change-point detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmentation</td>
<td>Anomaly detection</td>
<td>Markov order estimation</td>
</tr>
<tr>
<td>Filtering</td>
<td>Prediction</td>
<td>Entropy estimation</td>
</tr>
<tr>
<td>Causality testing</td>
<td>Compression</td>
<td>Content recognition</td>
</tr>
</tbody>
</table>
Bayesian Modelling of VMMCs

Notation. 1. Models $\equiv$ Trees

2. $X^j_i$ denotes the block $(X_i, X_{i+1}, \ldots, X_j)$

3. $\theta = \{\theta_s; s \in T\}$ for all the parameters (given $T$)

4. $X = X_{-d+1}, \ldots, X_0, X_1, \ldots, X_n$ all the observed data
Bayesian Modelling of VMMCs

Notation. 1. Models $\equiv$ Trees

2. $X^j_i$ denotes the block $(X_i, X_{i+1}, \ldots, X_j)$

3. $\theta = \{\theta_s; s \in T\}$ for all the parameters (given $T$)

4. $X = X_{-d+1}, \ldots, X_0, X_1, \ldots, X_n$ all the observed data

Prior on models Indexed family of priors on trees $T$ of depth $\leq D$

$$\pi(T) = \pi_D(T; \beta) = \alpha^{|T|-1}\beta^{|T|-L_D(T)}$$

with $\alpha = (1 - \beta)^{1/(m-1)}$; $|T| = \#$ leaves of $T$; $L_D(T) = \#$ leaves at $D$

[Lemma: This is OK]
Bayesian Modelling of VMMCs

**Notation.**

1. Models $\equiv$ Trees
2. $X^j_i$ denotes the block $(X_i, X_{i+1}, \ldots, X_j)$
3. $\theta = \{\theta_s; s \in T\}$ for all the parameters (given $T$)
4. $X = X_{-d+1}, \ldots X_0, X_1, \ldots, X_n$ all the observed data

**Prior on models** Indexed family of priors on trees $T$ of depth $\leq D$

$$\pi(T) = \pi_D(T; \beta) = \alpha^{|T| - 1} \beta^{|T| - L_D(T)}$$

with $\alpha = (1 - \beta)^{1/(m-1)}$; $|T| = \# \text{ leaves of } T$; $L_D(T) = \# \text{ leaves at } D$

**[Lemma: This is OK]**

**Prior on parameters** Given a context tree $T$, the parameters $\theta = \{\theta_s; s \in T\}$ are taken to be independent

with each $\pi(\theta_s|T) \sim \text{Dirichlet}(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2})$
Bayesian Inference of VMMCs

**Likelihood**  Given a model $T$ and parameters $\theta = \{\theta_s; s \in T\}$

$$f(X|\theta, T) = \prod_{i=1}^{n} P(X_i|X_{i-1}^{i-D}) = \prod_{s \in T} \prod_{j \in A} \theta_s(j)^{a_s(j)}$$

where the **count vectors** $a_s$ are defined by:

$$a_s(j) = \# \text{ times letter } j \text{ follows context } s \text{ in } X_1^n$$
Bayesian Inference of VMMCs

Likelihood Given a model $T$ and parameters $\theta = \{\theta_s; s \in T\}$

$$f(X|\theta, T) = \prod_{i=1}^{n} P(X_i|X_{i-1}^{i-D}) = \prod_{s \in T} \prod_{j \in A} \theta_s(j)^{a_s(j)}$$

where the count vectors $a_s$ are defined by:

$$a_s(j) = \# \text{ times letter } j \text{ follows context } s \text{ in } X_1^n$$

Model selection goal: The posterior distribution

$$\pi(T|X) = \frac{\int_{\theta} f(X|\theta, T) \pi(\theta|T) \, d\theta \pi(T)}{f(X)}$$
Bayesian Inference of VMMCs

**Likelihood** Given a model $T$ and parameters $\theta = \{\theta_s; s \in T\}$

$$f(X|\theta, T) = \prod_{i=1}^{n} P(X_i|X_{i-1}^{i-D}) = \prod_{s \in T} \prod_{j \in A} \theta_s(j)^{a_s(j)}$$

where the **count vectors** $a_s$ are defined by:

$$a_s(j) = \# \text{ times letter } j \text{ follows context } s \text{ in } X^n$$

**Model selection goal:** The **posterior distribution**

$$\pi(T|X) = \frac{\int_{\theta} f(X|\theta, T)\pi(\theta|T) \, d\theta \, \pi(T)}{f(X)}$$

**Main obstacle:** The **mean marginal likelihood**

$$f(X) = \sum_T \pi(T) \int_{\theta} f(X|\theta, T)\pi(\theta|T) \, d\theta$$

$\sim$ the number of models in the sum grows *doubly exponentially* in $D$
△ 1. [Tree.] Construct a tree with nodes corresponding to all contexts of length 1, 2, \ldots, D contained in $X$
Maximum A Posteriori Probability Tree Algorithm (MAPT)

1. [Tree. ] Construct a tree with nodes corresponding to all contexts of length 1, 2, . . . , D contained in X

2. [Estimated probabilities. ] At each node s compute the count vectors $a_s$ and the probabilities

$$P_{e,s} = \frac{\prod_{j=0}^{m-1}[(1/2)(3/2)\cdots(a_s(j) - 1/2)]}{(m/2)(m/2 + 1)\cdots(m/2 + M_s - 1)}$$

where $M_s = a_s(0) + \cdots + a_s(m - 1)$
Maximum A Posteriori Probability Tree Algorithm (MAPT)

△ 1. [Tree. ] Construct a tree with nodes corresponding to all contexts of length $1, 2, \ldots, D$ contained in $X$

△ 2. [Estimated probabilities. ] At each node $s$ compute the count vectors $a_s$ and the probabilities

$$P_{e,s} = \frac{\prod_{j=0}^{m-1}[(1/2)(3/2)\cdots(a_s(j) - 1/2)]}{(m/2)(m/2 + 1)\cdots(m/2 + M_s - 1)}$$

where $M_s = a_s(0) + \cdots + a_s(m - 1)$

△ 3. [Maximal probabilities. ] At each node $s$ compute

$$P_{m,s} = \begin{cases} P_{e,s}, & \text{if } s \text{ is a leaf} \\ \max\{\beta P_{e,s}, (1 - \beta)\prod_{j \in A} P_{m,s_j}\}, & o/w \end{cases}$$
Maximum A Posteriori Probability Tree Algorithm (MAPT)

△ 1. [Tree.] Construct a tree with nodes corresponding to all contexts of length 1, 2, . . . , D contained in $X$

△ 2. [Estimated probabilities.] At each node $s$ compute the count vectors $a_s$ and the probabilities

$$P_{e,s} = \frac{\prod_{j=0}^{m-1}[(1/2)(3/2) \cdots (a_s(j) - 1/2)]}{(m/2)(m/2 + 1) \cdots (m/2 + M_s - 1)}$$

where $M_s = a_s(0) + \cdots + a_s(m - 1)$

△ 3. [Maximal probabilities.] At each node $s$ compute

$$P_{m,s} = \begin{cases} P_{e,s}, & \text{if } s \text{ is a leaf} \\ \max\{\beta P_{e,s}, (1 - \beta) \prod_{j \in A} P_{m,sj}\}, & \text{o/w} \end{cases}$$

△ 4. [Pruning.] For each node $s$, if the above max is achieved by the first term, then prune all its descendants
Theorem: The MAPT Computes the MAP Tree

Theorem

The (pruned) tree $T_1^*$ resulting from the MAPT procedure has maximal \textit{a posteriori} probability among all trees:

$$\pi(T_1^*|X) = \max_T \pi(T|X) = \max_T \left\{ \int_\theta \frac{f(X|\theta, T)\pi(\theta|T)}{f(X)} d\theta \pi(T) \right\}$$
Theorem: The MAPT Computes the MAP Tree

Theorem

The (pruned) tree $T_1^*$ resulting from the MAPT procedure has maximal *a posteriori* probability among all trees:

$$
\pi(T_1^*|X) = \max_T \pi(T|X) = \max_T \left\{ \int_\theta f(X|\theta,T) \pi(\theta|T) \, d\theta \, \pi(T) \right\} 
$$

Note

The MAPT computes a doubly exponentially hard quantity in $O(n \cdot D^2)$ time

One of the very few examples of nontrivial Bayesian models for which the mode of the posterior is explicitly computable probably the most complex/interesting one
Additional Results

(i) \textit{Top k MAP models} \quad T_1^*, T_2^*, \ldots, T_k^*
Additional Results

(i) Top $k$ MAP models $T_1^*, T_2^*, \ldots, T_k^*$

(ii) Mean marginal likelihood $f(X)$ computed like the MAP but with averages instead of maxima
### Additional Results

(i) **Top $k$ MAP models**  \( T_1^*, T_2^*, \ldots, T_k^* \)

(ii) **Mean marginal likelihood**  \( f(X) \) computed like the MAP but with averages instead of maxima

(iii) **Model posterior probabilities**  \[
    \pi(T|X) = \frac{\pi(T) \prod_{s \in T} P_{e,s}}{f(X)}
\]
Additional Results

(i) **Top k MAP models** 
\[ T_1^*, T_2^*, \ldots, T_k^* \]

(ii) **Mean marginal likelihood** 
\[ f(X) \text{ computed like the MAP but with averages instead of maxima} \]

(iii) **Model posterior probabilities** 
\[ \pi(T|X) = \frac{\pi(T) \prod_{s \in T} P_{e,s}}{f(X)} \]

(iv) **Posterior odds** 
\[ \frac{\pi(T|X)}{\pi(T'|X)} = \frac{\pi(T) \prod_{s \in T, s \notin T'} P_e(a_s)}{\pi(T') \prod_{s \in T', s \notin T} P_e(a_s)} \]
Additional Results

(i) **Top k MAP models** $T_1^*, T_2^*, \ldots, T_k^*$

(ii) **Mean marginal likelihood** $f(X)$ computed like the MAP but with averages instead of maxima

(iii) **Model posterior probabilities**

\[
\pi(T|X) = \frac{\pi(T) \prod_{s \in T} P_{e,s}}{f(X)}
\]

(iv) **Posterior odds**

\[
\frac{\pi(T|X)}{\pi(T'|X)} = \frac{\pi(T) \prod_{s \in T, s \notin T'} P_e(a_s)}{\pi(T') \prod_{s \in T', s \notin T} P_e(a_s)}
\]

(v) **Full conditional density of $\theta$**

\[
\pi(\theta|T, X) \sim \prod_{s \in T} \text{Dirichlet}(a_s(0) + 1/2, a_s(1) + 1/2, \ldots, a_s(m - 1) + 1/2)
\]
### Additional Results

(i) **Top k MAP models**  
\[ T_1^*, T_2^*, \ldots, T_k^* \]

(ii) **Mean marginal likelihood**  
\[ f(X) \text{ computed like the MAP but with averages instead of maxima} \]

(iii) **Model posterior probabilities**  
\[ \pi(T | X) = \frac{\pi(T) \prod_{s \in T} P_e,s}{f(X)} \]

(iv) **Posterior odds**  
\[ \frac{\pi(T | X)}{\pi(T' | X)} = \frac{\pi(T) \prod_{s \in T, s \notin T'} P_e(a_s)}{\pi(T') \prod_{s \in T', s \notin T} P_e(a_s)} \]

(v) **Full conditional density of \( \theta \)**  
\[ \pi(\theta | T, X) \sim \prod_{s \in T} \text{Dirichlet}(a_s(0) + 1/2, a_s(1) + 1/2, \ldots, a_s(m - 1) + 1/2) \]

(vi) **MCMC exploration of the posterior**  
Metropolis-within-Gibbs sampling from \( \pi(\theta, T | X) \) using (iv) and (v)
A Large Data Set: Spike Trains

Data  Single neuron spike train in frontal eye fields (FEF) area located in the frontal cortex (Brodmann area 8) of the primate (monkey) brain

Study  FEF-V4 coupling during attention
        FEF is responsible for saccadic and voluntary eye movement
        Important role in the control of visual attention

MAPT  With \( n \approx 10^8 \) data points (ms resolution)
       \( m = 2, \beta = 1/2 \) and depth \( D = 130 \)

[MIT-NIH data: Gregoriou-Gotts-Zhou-Desimone Science (2012)]
A Large Data Set: Spike Trains

**Data**  Single neuron spike train in frontal eye fields (FEF) area

**Study**  FEF-V4 coupling during attention

**MAPT**  With $n \approx 10^8$ data points (ms resolution)

\[ m = 2, \ \beta = 1/2 \ \text{and depth} \ D = 130 \]

**Resulting MAPT model**

- Number of leaves: $|T| = 1054$
- Max depth: $D = 130$
- Max number of 1s/context: 3 (and two contexts with 4)
- Max number consecutive 1s: 2 (chemistry)
- Departure from simple renewal at 30ms

$\sim$ 1st/2nd order Markov renewal structure
A Fun Data Set: Wood Peewee Bird Song

**Data**  Recorded bird song data, transcribed as a sequence of (mono-)phthongs

Goal: Understand structure, complexity, variation and function

[Craig (1943) “The song of the wood pewee”]
A Fun Data Set: Wood Peewee Bird Song

**Data**  Recorded bird song data, transcribed as a sequence of (mono-)phthongs
Goal: Understand structure, complexity, variation and function

**MAPT**  With $n = 1327$ samples
$m = 3$, $\beta = 3/4$ and depth $D = 10$

posterior: $\pi(T^*_1|x) \approx 12.4\%$

prior: $\pi(T^*_1) \approx 3 \times 10^{-4}$
Wood Peewee Bird Song: Next 4 Models

\[ \pi(T_3^* | x) \approx 1.7\% \]

\[ \pi(T_2^* | x) \approx 2.2\% \]

\[ \pi(T_4^* | x) \approx 1.7\% \]

\[ \pi(T_5^* | x) \approx 1.7\% \]
## Bird Song Models: Comparison with Other Methods

<table>
<thead>
<tr>
<th>Result</th>
<th>MAPT</th>
<th>VLMC</th>
<th>MTD</th>
<th>gMTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1^*$, $d = 5$</td>
<td>complex tree, $d = 18$</td>
<td>complete, $d = 10$</td>
<td>complete, $d = 2$</td>
<td></td>
</tr>
<tr>
<td><strong>AIC</strong></td>
<td>687.4</td>
<td>796.8</td>
<td>1102.1</td>
<td>966.8</td>
</tr>
<tr>
<td><strong>BIC</strong></td>
<td>801.4</td>
<td>1273.6</td>
<td>1143.5</td>
<td>1003.0</td>
</tr>
</tbody>
</table>
Bird Song Models: Comparison with Other Methods

<table>
<thead>
<tr>
<th></th>
<th>MAPT</th>
<th>VLMC</th>
<th>MTD</th>
<th>gMTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>result</td>
<td>$T_1^*$, $d = 5$</td>
<td>complex tree, $d = 18$</td>
<td>complete, $d = 10$</td>
<td>complete, $d = 2$</td>
</tr>
<tr>
<td>AIC</td>
<td>687.4</td>
<td>796.8</td>
<td>1102.1</td>
<td>966.8</td>
</tr>
<tr>
<td>BIC</td>
<td>801.4</td>
<td>1273.6</td>
<td>1143.5</td>
<td>1003.0</td>
</tr>
</tbody>
</table>

Our Bayesian framework gives

- interesting and *interpretable* results
- good models by any metric
- a quantitative measure of accuracy
- allows for more applications
Bird Song Models: Comparison with Other Methods

<table>
<thead>
<tr>
<th></th>
<th>MAPT</th>
<th>VLMC</th>
<th>MTD</th>
<th>gMTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>result</td>
<td>$T_1^*, d = 5$</td>
<td>complex tree, $d = 18$</td>
<td>complete, $d = 10$</td>
<td>complete, $d = 2$</td>
</tr>
<tr>
<td>AIC</td>
<td>687.4</td>
<td>796.8</td>
<td>1102.1</td>
<td>966.8</td>
</tr>
<tr>
<td>BIC</td>
<td>801.4</td>
<td>1273.6</td>
<td>1143.5</td>
<td>1003.0</td>
</tr>
</tbody>
</table>

Our Bayesian framework gives

- interesting and *interpretable* results
- good models by any metric
- a quantitative measure of accuracy
- allows for more applications
- rich model-selection information via $k$-MAPT and MCMC

E.g., in $10^6$ steps, with an acceptance rate of $\approx 0.575$

we visit 269562 different models

The 100 most visited trees have 9-17 leaves and depths $4 \leq d \leq 6$
Results on empirical (including some “big”) data

- Genetics (DNA/RNA)
- Proteins and cross-omics data
- Neuroscience
- Whale/dolphin/bird song data

Applications

- Model selection
- Segmentation
- Filtering
- Causality testing
- Estimation
- Anomaly detection
- Prediction
- Compression
- Change-point detection
- Markov order estimation
- Entropy estimation
- Content recognition