Problem Statement

- Most source identification algorithms are based on stationarity of data.
- Consider a non-stationary data stream in which data statistics may change abruptly from sample to sample.
- Identify sources (the models and parameters) from multiple observations.

Data Generation Model

- A single stream (block) of length \( T \), \( x_t^T \), has been generated by multiple sources.
- Neither the sources nor the switching times are known.
- Let \( y_t^T \) be the (unknown) indexes of the sources that generated the sequence \( x_t^T \).
- Probability of \( x_t \) (sample at time \( t \)) depends on the past samples \( x_{t-1}^{t-1} \).

Tree Source Modeling

Each source, \( S_j \), can be described by

1. A set \( M_j \) consisting of all sequences, contexts, such that
   - No sequence in \( M_j \) is a suffix of another one
   - For any arbitrary sequences \( x_{t-1}^L \), there is a unique context \( c \in M_j \) such that \( c \) is a suffix of \( x_{t-1}^L \) and \( P(x_t = c|X_{t-1}^L) = P(x_t = a|X_{t-1}^L) = \theta(c, a) \).
2. Conditional probability distributions on contexts \( \Theta_j = (\theta(c,a) : a \in \mathcal{A}, c \in M_j) \).

Set \( M_j \) can be represented as a tree.

For example:
\[ M = \{00, 10, 11\} \]
\[ P(0|00) = 0.142, P(1|00) = 0.038, P(1|11) = 0.82 \]

BIC Context Tree Estimation for Single Source

For a hypothetical tree, \( \mathcal{M}_k \), of maximum depth \( D \):

- Number of free parameters: \( d = (|\mathcal{A}|-1) |\mathcal{M}_k| \)
- Histogram over tree model, \( \forall c \in \mathcal{M}_k \) and \( a \in \mathcal{A}, n_k(c,a) = \left| \{ x_t : 1 \leq t \leq N, x_{t-1}^{t-1} = c, x_t = a \} \right| \)
- Maximum Log-Likelihood:
  \[ L_{\mathcal{M}_k}(X;k) = \sum_{c \in \mathcal{M}_k, a \in \mathcal{A}} n_k(c,a) \log \left( \frac{n_k(c,a)}{n_k(a)} \right) \]
- Bayesian Information Criteria (BIC) of model w.r.t. \( x \) of length \( T \):
  \[ BIC_{\mathcal{M}_k}(X;k) = -L_{\mathcal{M}_k}(X;k) + \frac{d}{2} \log T \]
- BIC model estimator
  \[ \hat{\mathcal{M}}_k(X;k) = \arg\min_{\mathcal{M}_k} BIC_{\mathcal{M}_k}(X;k) \]

Theorem [Talata and Duncan, 2011]

For a stationary ergodic source, \( S_k \), with context tree \( \mathcal{M}_k \), for any constant \( D \), \( \hat{\mathcal{M}}_k(X;k) \rightarrow \mathcal{M}_k \) almost surely as length of observed signal increases. Also, the maximum likelihood estimates \( \hat{\theta}(c,a) = n_k(c,a)/n_k(a) \) converges to \( P_k(a|c) \).

Simulation Results

- 3 random sources over alphabet \( \mathcal{A} = \{0,1,2,3\} \)
- Different trees with depths 2, 2 and 3 for each source
- Entropy of each source is \( I \)
- Transition probability between source

EM-BIC Model Estimator

- Exact values of \( P(Y_t = k|x_0^T) \) are unknown.
- Iteratively estimate the probabilities.

Outline of the Algorithm

1. Start from an initial estimate for sources.
2. Repeat until convergence:
   1. Use Baum-Welch algorithm and current estimates of sources to compute \( P(Y_{t-1} = l, Y_t = k|x_0^T) \) and \( P(Y_t = k|x_0^T) \).
   2. Using updated posteriors, for the \( k \)th source, estimate the model \( \hat{\mathcal{M}}_k(X;k) = \arg \min \text{BIC}_{\mathcal{M}_k}(X;k) \).
   3. Update sources' parameters and the switching probabilities.

Algorithm

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Zip</th>
<th>PAQ6</th>
<th>K = 1</th>
<th>K = 2</th>
<th>K = 3</th>
<th>K = 4</th>
<th>K = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation</td>
<td>277.6</td>
<td>19.2</td>
<td>68.8</td>
<td>20.0</td>
<td>2.4</td>
<td>3.2</td>
<td>4.8</td>
</tr>
</tbody>
</table>