MIXTURE SOURCE IDENTIFICATION IN NON-STATIONARY DATA STREAMS WITH APPLICATIONS IN COMPRESSION

Afshin Abdi and Faramarz Fekri

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA
{abdi,fekri}@ece.gatech.edu

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_1^T, has been generated by multiple sources.
- Neither the sources nor the switching times are known.
- Let y_1^T be the (unknown) indexes of the sources that generated the sequence x_1^T
- Probability of x_t (sample at time t) depends on
  - the past samples x_1^{t-1}
  - the source that generated it, y_t,

$$P(x_1^T|y_1^T,\boldsymbol{\Delta}) = \prod_{t=1}^{T} P_{y_t}(x_t|x_1^{t-1})$$

$\boldsymbol{\Delta}$ is the set of mixture parameters.

Switching between Sources

- At time t, y_t, the index of the active source, depends on the previous ones, y_1^{t-1}.
- For simplicity, independent from x_1^{t-1}.
- Assume it is governed by a hidden Markov(1) source

$$P(y_1^T|\boldsymbol{\Delta}) = w_{y_1}\prod_{t=2}^{T} P_h(y_t|y_{t-1})$$

w_k: the initial probability of source k
$P_h(k|l)$: the probability of switching from source l to k

$$P(x_1^T|\boldsymbol{\Delta}) = \sum_{y_1^T} P(x_1^T|y_1^T,\boldsymbol{\Delta})P(y_1^T|\boldsymbol{\Delta})$$

Tree Source Modeling

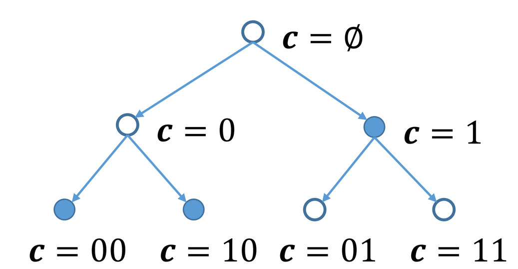
Each source, S_i, can be described by

1. A set $\mathcal{M}_i$ consisting of all sequences, *contexts*, such that
   - No sequence in $\mathcal{M}_i$ is a suffix of another one
   - For any arbitrary sequences $x_{-\infty}^{-1}$ there is a unique context $\mathbf{c} \in \mathcal{M}_i$ such that $\mathbf{c}$ is a suffix of $x_{-\infty}^{-1}$ and $P(X_0 = a|X_{-\infty}^{-1} = x_{-\infty}^{-1}) = P\left(X_0 = a|X_{-l(c)}^{-1} = \mathbf{c}\right) =: \theta(\mathbf{c},a)$
2. Conditional probability distributions on contexts

$$\boldsymbol{\Theta}_i = \{\theta(\mathbf{c},a):\ a \in \mathcal{A}, \boldsymbol{c} \in \mathcal{M}_i\}$$

Set $\mathcal{M}_i$ can be represented as a tree.
For example:

$\mathcal{M} = \{00, 10, 1\}$
$P(011000) = P(01)\ P(1|1)P(0|1)\ P(0|10)$ P(0|00)



BIC Context Tree Estimation for Single Source

For a hypothetical tree, $\mathcal{M}$, of maximum depth D

- Number of free parameters $d = (|\mathcal{A}| - 1)\ |\mathcal{M}|$
- Histogram over tree model, $\forall \boldsymbol{c} \in \mathcal{M}$ and $a \in \mathcal{A}$,

$$n_{\boldsymbol{x}}(\boldsymbol{c},a) = \left|\{i: D < i \le n\,, x_{i-l(c)}^{i-1} = \boldsymbol{c}, x_i = a\}\right|$$

$$n_{\boldsymbol{x}}(\boldsymbol{c}) = \sum_{a\in\mathcal{A}} n_{\boldsymbol{x}}(\boldsymbol{c},a)$$

- Maximum Log-Likelihood:

$$\mathcal{L}_{\mathcal{M}}(\boldsymbol{x}) \approx \sum_{\boldsymbol{c}\in\mathcal{M}, a\in\mathcal{A}} n_{\boldsymbol{x}}(\boldsymbol{c},a)\log\left(\frac{n_{\boldsymbol{x}}(\boldsymbol{c},a)}{n_{\boldsymbol{x}}(\boldsymbol{c})}\right)$$

- Bayesian Information Criteria (BIC) of model w.r.t. $\boldsymbol{x}$ of length T:

$$BIC_{\mathcal{M}}(\boldsymbol{x}) = -\mathcal{L}_{\mathcal{M}}(\boldsymbol{x}) + \frac{d}{2}\log T$$

- BIC model estimator

$$\widehat{\mathcal{M}}_{BIC}(\boldsymbol{x}) = \underset{\mathcal{M}}{\operatorname{argmin}}\ BIC_{\mathcal{M}}(\boldsymbol{x})$$

Theorem [Talata and Duncan, 2011]

For a stationary ergodic source, S, with context tree $\mathcal{M}_S$, for any constant D, $\widehat{\mathcal{M}}_{BIC}(\boldsymbol{x})|_D \to \mathcal{M}_S|_D$ almost surely as length of observed signal increases. Also, the maximum likelihood estimates $\hat{\theta}(\boldsymbol{c},a) = \frac{n_x(\boldsymbol{c},a)}{n_x(\boldsymbol{c})}$ converges to $P_S(a|\boldsymbol{c})$.

Model Estimation for Multiple Sources

- Observed a non-stationary sequence $\boldsymbol{x}$ of length T
- Effective histogram for the k^{th} source

$$\bar{n}_k(\boldsymbol{c},a;\boldsymbol{x}) = \sum_{t=1}^{T} P(Y_t = k|\boldsymbol{x})\ \mathbb{I}_{c,a}(t;\boldsymbol{x})$$

where

$$\mathbb{I}_{c,a}(t;\boldsymbol{x}) = \begin{cases} 1 & x_t = a,\ \ x_{t-l(c)}^{t-1} = \boldsymbol{c} \\ 0 & \text{otherwise} \end{cases}$$

- For multiple sequences, $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(N)}\}$

$$\bar{n}_k(\boldsymbol{c},a) = \sum_{n=1}^{N} \bar{n}_k(\boldsymbol{c},a;\boldsymbol{x}^{(n)})\ ,$$

$$\bar{n}_k(\boldsymbol{c}) = \sum_{a\in\mathcal{A}} \bar{n}_k(\boldsymbol{c},a)$$

- BIC model estimator for the k-th source

$$\mathcal{L}_{\mathcal{M}}(\mathcal{X};k) \approx \sum_{\boldsymbol{c}\in\mathcal{M}, a\in\mathcal{A}} \bar{n}_k(\boldsymbol{c},a)\log\frac{\bar{n}_k(\boldsymbol{c},a)}{\bar{n}_k(\boldsymbol{c})}$$

$$BIC_{\mathcal{M}}(\mathcal{X};k) = -\mathcal{L}_{\mathcal{M}}(\mathcal{X};k) + \frac{d}{2}\log\bar{n}_k$$

$$\widehat{\mathcal{M}}_k(\mathcal{X}) = \underset{\mathcal{M}}{\operatorname{argmin}}\ BIC_{\mathcal{M}}(\mathcal{X};k)$$

Lemma

Assuming ergodicity and stationarity of sources, $\frac{\bar{n}_k(\boldsymbol{c},a)}{\bar{n}_k(\boldsymbol{c})} \to P_k(a|\boldsymbol{c})$ almost surely as $N \to \infty$ provided that $P\left(Y_t = k|\boldsymbol{x}^{(n)}\right) \neq 0$ and $l\left(\boldsymbol{x}^{(n)}\right) > l(\boldsymbol{c})$ infinitely often.

Theorem

For a constant D, assume that $l\left(\boldsymbol{x}^{(n)}\right) > D$. Then, $\widehat{\mathcal{M}}_k(\mathcal{X})|_D \to \mathcal{M}_k|_D$ almost surely as $N \to \infty$.

EM-BIC Model Estimator

- Exact values of $P\left(Y_t = k|\boldsymbol{x}^{(n)}\right)$ are unknown.
- Iteratively estimate the probabilities

Outline of the Algorithm

- Start from an initial estimate for sources.
- Repeat until convergence:
  1. Use Baum-Welch algorithm and current estimates of sources to compute $P\left(Y_{t-1} = l\,, Y_t = k|\boldsymbol{x}^{(n)}\right)$ and $P\left(Y_t = k|\boldsymbol{x}^{(n)}\right)$
  2. Using updated posteriors, for the k^{th} source, estimate the model $$\widehat{\mathcal{M}}_k(\mathcal{X}) = \underset{\mathcal{M}}{\operatorname{argmin}}\ BIC_{\mathcal{M}}(\mathcal{X};k)$$
  3. Update sources' parameters and the switching probabilities

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 *1*
- Transition probability between source

$$A = \begin{bmatrix} 0.857 & 0.052 & 0.091 \\ 0.041 & 0.879 & 0.08 \\ 0.142 & 0.038 & 0.82 \end{bmatrix}$$

- Optimum compression rate: 1.4648 bits per symbol
- 100 sequence of length 1000 created for training and another set for test.
- KL divergence between true sources and estimated ones: less than 0.001
- Compression performance compared to PAQ8, Zip, using different number of estimated sources in the mixture model.

Algorithm	Zip	PAQ8	$\hat{K} = 1$	$\hat{K} = 2$	$\hat{K} = 3$	$\hat{K} = 4$	$\hat{K} = 5$
Redundancy (bits/1K symbols)	277.6	19.2	68.8	20.0	2.4	3.2	4.8