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

Afshin Abdi, Faramarz Fekri

Georgia Institute of Technology

ABSTRACT

We consider a non-stationary data stream in which the data statistics may change abruptly from one sample to another, i.e. each sample might be generated from a different (unknown) source in a mixture of K sources. The problem of identifying the models and parameters of K sources, as well as the source switching model is investigated. We proposed an algorithm based on Bayesian Information Criterion and Expectation Maximization to determine the models and estimate the mixture parameters. The estimated data generation model can be used in memory-assisted universal compression to decrease the coding rate further. Simulation results confirmed that using the proposed algorithm for source identification and universal compression can significantly decrease the compression redundancy.

Index Terms— Source Identification, Non-Stationary Data Modeling, Memory-Assisted Universal Compression

1. INTRODUCTION

Modeling data generation and source identification is a fundamental problem encountered in various applications from pattern recognition to data compression. Most identification problems are developed based on the assumption that the data is stationary. However, in many applications, this is not the case. For example, in single source scenarios, the properties of the source may vary over time, or there might be numerous (hidden) sources that at each time, the data are generated from one of them. In this paper, we assume that the changes in the data statistics are happening abruptly. In our setup, neither change times nor the probabilistic models are known a priori.

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_T)$ be a finite-length stochastic process over alphabet $\mathcal{A}$, and $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$, $x_t \in \mathcal{A}$, a realization of $\boldsymbol{X}$. Assume that data is being generated by a mixture of K sources $\mathcal{S} = \{S_1, \ldots, S_K\}$ where at time t, an unknown source S_k emits a symbol according to the probability $P_k(x|x_1^{t-1})$, where $x_1^{t-1} = (x_1, x_2, \ldots, x_{t-1})$ is the memory of the past samples. Note that there is no assumption on the structure of $P_k(.)$'s, i.e. whether if the sources are i.i.d.or a Markov of certain order.

When the exact source model is unknown a priori, directly applying maximum likelihood estimator (MLE) to the most general model would be problematic; First, usually the most complex model would be adapted as it often gives higher likelihoods, even if the source was characterized by a simpler model. Second, there might not exist enough data samples to reliably estimate parameters of a complex model. Further, in applications like compression, when it is required to store or transmit the model parameters, the overhead due to the complex model representation may be comparable to data itself. Thus, we need an algorithm to find the simplest but reliable model that describes the data generation accurately.

Model identification of the source model has been investigated by many authors when a long sequence from an ergodic stationary source is observed, and Minimum Description Length principle (MDL) [3] or Bayesian Information Criterion (BIC) have been successfully applied. For a Markov source, with a known upper bound on its order, it is shown that BIC and MDL are strongly consistent for order estimation. Without such a bound, the consistency of BIC order estimator is shown in [4]. The consistency of BIC model estimator for finite memory sources and an arbitrary ergodic stationary source are also shown in [5, 6].

In [7], the authors investigated the problem when multiple short sequences are observed from a mixture of sources, such that each sequence is entirely generated by an unknown source in the mixture. We developed an iterative application of the EM and BIC to estimate the models and parameters of the mixture and showed that it can recover true sources' models when sufficient number of sequences are available.

In this paper, we apply BIC to determine the sources' models that generate non-stationary data sequences and estimate their parameters. Section 2 discusses the proposed algorithm. In Section 3, we investigate a possible application of the proposed method in universal compression and as to how it can be used to reduce the redundancy with respect to the minimum achievable rate. Finally, simulations to verify the proposed algorithm are provided in Section 4.

1.1. Notations

For an arbitrary sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$, its length is denoted by $l(\boldsymbol{x})$ and the subsequence $(x_m, x_{m+1}, \ldots, x_n)$ by x_m^n. For $n < m$, x_m^n is the empty sequence, $\varnothing$. For a set $\mathcal{A}$, $|\mathcal{A}|$ is the number of elements in $\mathcal{A}$.

For $a \in \mathcal{A}$ and a sequence $\boldsymbol{x}$, the indicator function of a at time t, $1 \leq t \leq l(\boldsymbol{x})$, is defined as $\mathbb{1}_a(t; \boldsymbol{x}) = 1$ if $x_t = a$,

otherwise it is zero. Similarly, for $a \in \mathcal{A}$ and sequence $\boldsymbol{c}$,

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

$P(x;\boldsymbol{\theta})$ denotes the probability distribution of x determined by the parameters $\boldsymbol{\theta}$. And $\mathbb{E}_p(.)$ is used to denote the expectation taken with respect to the distribution given by p.

2. MIXTURE CHARACTERIZATION

Assume that there are K ergodic stationary sources $\mathcal{S} = \{S_1, \ldots, S_K\}$ where each source S_k generates data according to the model $P_k(\cdot)$. Let $\boldsymbol{S}$ be the set of all models and the corresponding parameters for these K sources. At each time instant t, $1 \le t \le T$, one of the sources becomes *active* and generates x_t based on the past samples, x_1^{t-1}. In other words, if we denote the index of the active source at time t by y_t, then the probability of observing x_t is $P_{y_t}(x_t|x_1^{t-1})$ and knowing the entire sequence of sources, $\boldsymbol{y}$, the probability of observing $\boldsymbol{x}$ is

$$P(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{S}) = \prod_{t=1}^{T} P_{y_t}(x_t|x_1^{t-1}) \tag{1}$$

In practice, the indexes of the sources which generate $\boldsymbol{x}$ are not known a priori. Hence, to fully characterize data generation model, we need to make some assumptions on the switching among the sources, or equivalently, the generation of sequence $\boldsymbol{y}$. For simplicity, we assume that $\boldsymbol{y}$ is independent of $\boldsymbol{x}$, i.e., the switching among sources is independent from the generated sequence. This leads us to assume that the sequence $\boldsymbol{y}$ is also generated from an unknown *hidden source*. Although, finite state machines are more general and suitable to model $\boldsymbol{y}$, in this paper, for simplicity, we assume that this hidden source is a Markov of order one[1]. Therefore, the probability of a specific source order, $\boldsymbol{y}$, is

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

where $\boldsymbol{w} = (w_1, \ldots, w_K)$ is the initial probability distributions and $P_h(Y_t = j|Y_{t-1} = i)$ is the probability of switching from source i at time $t-1$ to source j at time t. We assume that this transition probability is fixed and denot it by $a_{i,j}$.

Let $\boldsymbol{\Theta} = (\boldsymbol{w}, P_h, \boldsymbol{S})$ be the set of all parameters of the model. The probability of observing a sequence $\boldsymbol{x}$ is

$$P(\boldsymbol{x};\boldsymbol{\Theta}) = \sum_{\boldsymbol{y}} P_h(\boldsymbol{y}) P(\boldsymbol{x}|\boldsymbol{y}) \tag{3}$$

To fully identify the data generation model from a set of observed data $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$, we need to (1) estimate number of sources; K, (2) estimate the hidden source's parameters; $\boldsymbol{w}$ and P_h, (3) find models of each source, S_k, $1 \le k \le K$, and (4) finally, estimate parameters of each source's model. Note that as currently there is no straightforward method to find the number of sources, K, in the development of the algorithm we assume that it is known and fixed. By comparing the performance (e.g. compression rate or a measure on the fitness of model) for different values of K, the optimum number of sources is determined.

In the following, we explain our proposed method to iteratively estimate the models and parameters of the sources.

2.1. Parameters of the hidden source

Knowing the models and parameters from the previous iteration of the algorithm, estimating the parameters of the hidden source is straightforward and the same as the ordinary HMM, which is repeated here for the sake of completeness:

$$w_k^{(i+1)} = \frac{1}{N} \sum_{n=1}^{N} P(y_1 = k|\boldsymbol{x}^{(n)};\boldsymbol{\Theta}^i) \tag{4a}$$

$$a_{k,l}^{(i+1)} = \frac{\sum_{n=1}^{N} \sum_t P(y_{t-1} = k, y_t = l|\boldsymbol{x}^{(n)};\boldsymbol{\Theta}^i)}{\sum_{n=1}^{N} \sum_t P(y_{t-1} = k|\boldsymbol{x}^{(n)};\boldsymbol{\Theta}^i)} \tag{4b}$$

Note that $P(y_{t-1} = k, y_t = l|\boldsymbol{x};\boldsymbol{\Theta}^i)$ and $P(y_t = k|\boldsymbol{x};\boldsymbol{\Theta}^i)$ can be computed efficiently using Baum-Welch algorithm.

2.2. Models and parameters of sources

Here, we consider the class of all stationary ergodic sources that can be modeled by a tree. The context set of a source S, denoted by $\mathcal{T}_S$, is the set of all sequences, $\boldsymbol{c}$, such that none of them is a suffix of another one in $\mathcal{T}_S$. Additionally, for all sequences $x_{-\infty}^{-1}$, there exists a *unique* $\boldsymbol{c} \in \mathcal{T}_S$ such that $x_{-l(\boldsymbol{c})}^{-1} = \boldsymbol{c}$ and $\forall a \in \mathcal{A}$: $P(X_0 = a|x_{-\infty}^{-1}) = P(X_0 = a|\boldsymbol{c})$.

We are interested in finding a consistent estimator of the tree $\mathcal{T}_S$ and its parameters $\boldsymbol{\theta} = \{\theta(a|\boldsymbol{c}) : \forall \boldsymbol{c} \in \mathcal{T}_S, \forall a \in \mathcal{A}\}$ where $\theta(a|\boldsymbol{c}) := P_S(X_0 = a|X_{-l(\boldsymbol{c})}^{-1} = \boldsymbol{c})$.

Denote the number of occurrences of context $\boldsymbol{c}$ followed by letter a in $\boldsymbol{x} = x_1^T$ by $n_{\boldsymbol{x}}(\boldsymbol{c}, a) = \sum_t \mathbb{1}_{\boldsymbol{c},a}(t;\boldsymbol{x})$. Let $n_{\boldsymbol{x}}(\boldsymbol{c}) = \sum_{a \in \mathcal{A}} n_{\boldsymbol{x}}(\boldsymbol{c}, a)$. Hence, the maximum log-likelihood of $\boldsymbol{x}$ with respect to a context tree $\mathcal{T}$ is

$$\mathcal{L}_{\mathcal{T}}(\boldsymbol{x}) = \sum_{\boldsymbol{c} \in \mathcal{T}, a \in \mathcal{A}} n_{\boldsymbol{x}}(\boldsymbol{c}, a) \log \left(\frac{n_{\boldsymbol{x}}(\boldsymbol{c}, a)}{n_{\boldsymbol{x}}(\boldsymbol{c})} \right) \tag{5}$$

with the convention that $0 \log 0 = 0$.

In [8] and [6], Bayesian Information Criterion (BIC) is used to estimate the context tree of an ergodic stationary source when a sufficiently long sequence from that source is observed. For a hypothetical tree, $\mathcal{T}$, the BIC for a sequence $\boldsymbol{x}$ of length $l(\boldsymbol{x}) = n_{\boldsymbol{x}}$ is defined as

$$BIC_{\mathcal{T}}(\boldsymbol{x}) = -\mathcal{L}_{\mathcal{T}}(\boldsymbol{x}) + \frac{(|\mathcal{A}| - 1)|\mathcal{T}|}{2} \log n_{\boldsymbol{x}} \tag{6}$$

and the BIC context tree estimator is given by

$$\widehat{\mathcal{T}}_{BIC}(\boldsymbol{x}) = \operatorname*{argmin}_{\mathcal{T}} BIC_{\mathcal{T}}(\boldsymbol{x}). \tag{7}$$

[1] The analysis of more complex hidden source model is almost the same as any source with finite memory is equivalent to a Markov(1) source over an extended alphabet.

Theorem 1 (2.11 [6]). *For any stationary ergodic source with context tree $\mathcal{T}_S$, for any constant integer D, $\widehat{\mathcal{T}}_{BIC}(\boldsymbol{x})|_D \to \mathcal{T}_S|_D$ almost surely as $l(\boldsymbol{x}) \to \infty$.*

Further, the maximum likelihood estimates $\hat{\theta}(\boldsymbol{c}, a) = \frac{n_{\boldsymbol{x}}(\boldsymbol{c},a)}{n_{\boldsymbol{x}}(\boldsymbol{c})}$ converges to the source parameters $P_S(a|\boldsymbol{c})$.

where $\mathcal{T}|_D$ is the truncation of tree to depth D, defined as

$$\begin{aligned}\mathcal{T}|_D =&\{\boldsymbol{c} \in \mathcal{T},\ l(\boldsymbol{c}) \le D\} \cup \\ &\{\boldsymbol{c},\ l(\boldsymbol{c}) = D,\ \boldsymbol{c} \text{ is a suffix of some } \boldsymbol{c}' \in \mathcal{T}\}\end{aligned}$$

In [7], we have extended the above results when multiple (relatively short) sequences from one or several sources are observed such that each sequence is independently generated from an unknown source. However, here, each part of any sequence might be generated from a different source and therefore, the results of [7] are not directly applicable.

For a given set of observations $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)}\}$, assume that $P(Y_t = k|\boldsymbol{x}^{(n)})$ is known for all t (or by some means, we have an accurate estimate of the values). For an arbitrary tree $\mathcal{T}$, all $a \in \mathcal{A}$ and $\boldsymbol{c} \in \mathcal{T}$, let

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

Define $\overline{n}_k(\boldsymbol{c})$ and $\overline{n}_k$ similarly. Assuming ergodicity and stationarity of the individual and hidden sources [9], it is easy to show that

Lemma 2. *For all k, $1 \le k \le K$, $\frac{\overline{n}_k(\boldsymbol{c},a)}{\overline{n}_k(\boldsymbol{c})} \to P_k(a|\boldsymbol{c})$ almost surely as $N \to \infty$, provided that $P(Y_t = k|\boldsymbol{x}^{(n)}) \neq 0$ and $l(\boldsymbol{x}^{(n)}) > l(\boldsymbol{c}a)$ infinitely often.*

Now, for the k^{th} source, the maximum log-likelihood and BIC with respect to a tree $\mathcal{T}$ are defined as

$$\begin{aligned}&\hat{\theta}_k(\boldsymbol{c}, a) = \frac{\overline{n}_k(\boldsymbol{c}, a)}{\overline{n}_k(\boldsymbol{c})} \\ &\mathcal{L}_{\mathcal{T}}(\mathcal{X}; k) = \sum_{\boldsymbol{c} \in \mathcal{T}, a \in \mathcal{A}} \overline{n}_k(\boldsymbol{c}, a) \log \hat{\theta}_k(\boldsymbol{c}, a) \\ &BIC_{\mathcal{T}}(\mathcal{X}; k) = -\mathcal{L}_{\mathcal{T}}(\mathcal{X}; k) + \frac{(|\mathcal{A}| - 1)\,|\mathcal{T}|}{2} \log \overline{n}_k\end{aligned}$$

and the BIC tree estimator for source S_k is given by

$$\widehat{\mathcal{T}}_k(\mathcal{X}) = \operatorname*{argmin}_{\mathcal{T}} BIC_{\mathcal{T}}(\mathcal{X}; k) \tag{8}$$

Using Lemma 2, Lemmas 3.1 and 3.2 from [8], the following results can be shown;

Theorem 3. *For a constant D, assume that $l(\boldsymbol{x}_i) > D$. Then $\widehat{\mathcal{T}}_k(\mathcal{X})|_D = \mathcal{T}_k|_D$ almost surely as $N \to \infty$, where $\mathcal{T}_k$ is the true context tree of source S_k. Moreover, the maximum likelihood estimates of the parameters converge to the source parameters.*

As the values of $P(Y_t = k|\boldsymbol{x}^{(n)})$ are not known, using the above theorem to find the sources' models and parameters is not possible. As such, we propose using the estimated parameters from the previous iteration of the EM algorithm to approximate $P(Y_t = k|\boldsymbol{x}^{(n)})$ and compute $\overline{n}_k(\boldsymbol{c}, a)$ and $\overline{n}_k(\boldsymbol{c})$, which are then used to refine the estimations of $\mathcal{T}_k$ and $\boldsymbol{\theta}_k$, for $1 \le k \le K$, at each iteration of the EM algorithm.

Summarizing our proposed approach to identify underlying process generating non-stationary data sequences, knowing estimations at the i^{th} iteration of the algorithm, $\boldsymbol{\Theta}^i$,

1. Use modified Baum-Welch algorithm to compute $P(Y_{t-1} = l, Y_t = k|\boldsymbol{x}^{(n)}; \boldsymbol{\Theta}^i)$ and $P(Y_t = k|\boldsymbol{x}^{(n)}; \boldsymbol{\Theta}^i)$ for all n, t, k and l.
2. Use (4) to update transition model among sources.
3. Use $P(.|\boldsymbol{x}^{(n)}; \boldsymbol{\Theta}^i)$ to estimate $\overline{n}_k(\boldsymbol{c}, a)$, $\overline{n}_k(\boldsymbol{c})$ and $\overline{n}_k$.
4. Apply Thm. 3 to update sources' models and parameters.

3. APPLICATION IN UNIVERSAL COMPRESSION

In universal compression of data from a single stationary parametric source, it is known that the redundancy of compressing a sequence of length n is $\frac{d}{2}\log(n) + O(1)$ where d is the number of free (unknown) parameters, and using a common memory of length m between the encoder and the decoder, reduces the redundancy to $\frac{d}{2}\log(1 + \frac{n}{m}) + o(1)$ [10]. However, for a data sequence for which each part might be generated from a different unknown source, no such bound for the redundancy of universal compression exists. If the sources' statistics were unknown but their indexes, $\boldsymbol{y}$, were available for the data sequences, then the redundancy of universal compression would be similar to [10, 11]. However, without such an extra information, there is no known bound for the redundancy of universal compression for hidden Markov processes.

Memory-assisted compression was proposed to close the gap between universal compression and the optimum code-length in finite-length regimes [10]. One way of exploiting the memory of past data is to estimate the parameters of the mixture of sources and use them to approximate the probability distribution of the next symbol, i.e. if $\widehat{\boldsymbol{\Theta}}$ is the estimated parameter, having observed and compresses x_1^{t-1}, to compress the next symbol, $P(X_t|x_1^{t-1}, \widehat{\boldsymbol{\Theta}})$ is computed and an entropy coder like arithmetic codec uses that information to compress x_t. Hence, the the code-length would be approximately $-\log P(\boldsymbol{x}; \widehat{\boldsymbol{\Theta}})$ instead of the optimal $-\log P(\boldsymbol{x}; \boldsymbol{\Theta})$.

4. SIMULATION RESULTS

To verify our approach, we created 3 random tree sources of depths 2, 2 and 3 over an alphabet of size 4 (Fig. 1 shows

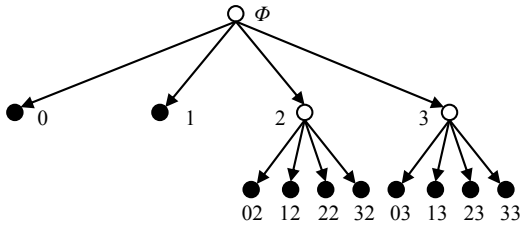

Fig. 1. The tree structure of the first source (black circle are contexts)

an example of a tree structure used in our experiments.). The sources are designed such that for each context $\boldsymbol{c}$, their conditional entropy is 1, i.e. $H_k\big(X_t|X_{t-l(\boldsymbol{c})}^{t-1}=\boldsymbol{c}\big)=1$. Hence, their entropy is 1 and given a specific sequence of sources $\boldsymbol{y}$, the minimum average code-length per symbol is

$$\frac{1}{T}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{y}}\left(-\log P(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\Theta})\right)=1$$

The initial and transition probabilities between sources are

$$\boldsymbol{w}=\begin{bmatrix}0.402\\0.274\\0.324\end{bmatrix}\text{ and }A=\begin{bmatrix}0.857&0.052&0.091\\0.041&0.879&0.08\\0.142&0.038&0.82\end{bmatrix}$$

hence, the asymptotic per-symbol entropy to encode and transmit $\boldsymbol{y}$ is $H(Y)=0.7325$.

For simulations, $N=100$ sequences of length 1000 are generated randomly and used as the memory. Further, we have generated another set of data to verify and compare the compression performance algorithm.

Source Identification: For source identification, we initialized the algorithm with K random i.i.d. sources and then ran the EM algorithm until the improvement in the likelihood of data is negligible (less that 10^{-5} per symbol). The normalized (negative of) log-likelihood of the whole sequences is considered as the cost function. Results of simulations for different values of K are given in table 1.

Table 1. Average cost for different values of K

K	1	2	3	4	5
cost	1.49	1.44	1.42	1.42	1.42

We noticed that the cost generally decreases by increasing number of hypothetical sources, K, but for $K\geq 3$ the improvement is negligible. Therefore, by comparing the cost functions, we conclude that $K=3$ is the optimum choice for the number of sources in the mixture.

The estimated transition matrix for the sources is

$$\widehat{A}\approx\begin{bmatrix}0.856&0.051&0.093\\0.042&0.874&0.086\\0.142&0.042&0.816\end{bmatrix}$$

Also, we found out that the models of the sources were found correctly (the algorithm decides on the same tree as the original source model). For example, the first source was identified as in Fig. 1 and the KL divergence between the found source and the true source is less than 0.001.

Memory-Assisted Compression: We used estimated parameters for the compression of test data and compared the redundancy of our method to various values of $\widehat{K}$ with those of Zip and PAQ8 [12] algorithms.To implement the memory-assisted PAQ8 algorithm, we used the common memory between the encoder and the decoder to train it and then used the trained algorithm to compress the test data. This gives slightly better compression performance than the ordinary PAQ. To compute the redundancy, the optimum code-length is computed using true sources' models which is $R^*=1.4648$ bits per symbol.

The redundancy values are given in table 2. To verify the effect of the number of detected sources on the compression performance, we used estimated sources for different values of $\widehat{K}$ in addition to our proposed algorithm which uses $\widehat{K}=3$ estimated sources. Note that the case $\widehat{K}=1$ approximately equals the universal compression using a single model. Comparing performances, we see that increasing the number of hypothetical sources from 3 to 5 makes the compression performance slightly worse due to the increased number of parameters to estimate. Also, the proposed algorithm is close to the optimum code rate (the models and parameters are known) and performs better than PAQ8. We expect to see higher compression gains when the alphabet size increases or data is being generated from a more complex models.

Table 2. Redundancy of memory-assisted universal compression for different methods

Compression Algorithm	$\widehat{K}=1$	$\widehat{K}=2$	Proposed ($\widehat{K}=3$)	$\widehat{K}=4$	$\widehat{K}=5$	PAQ8	Zip
Redundancy (bits/1K symbols)	68.8	20.0	2.40	3.20	4.80	19.2	277.6

5. CONCLUSION

In this paper, we have investigated the problem of universal compression of non-stationary data sequences. First, we proposed an extended hidden Markov process to model the data. For the memory-assisted universal compression, we proposed to use the memory to identify the sources in the mixture. We showed that under some conditions, the proposed Expectation Maximization algorithm in conjunction with Bayesian Information Criterion can successfully estimate the models of the sources, their parameters and the transition probabilities between sources. Then, we used these estimations in a memory-assisted compression to entropy-code the sequences. Our simulation results showed that the proposed approach outperforms the existing algorithms PAQ8 and Zip.

6. REFERENCES

[1] C. Mattern, "Combining non-stationary prediction, optimization and mixing for data compression," in *Data Compression, Communications and Processing (CCP), 2011 First International Conference on*, June 2011, pp. 29–37.

[2] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," International Computer Science Institute, Tech. Rep. TR-97-021, 1998.

[3] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *Information Theory, IEEE Transactions on*, vol. 44, no. 6, pp. 2743–2760, Oct 1998.

[4] I. Csiszar and P. Shields, "The consistency of the bic markov order estimator," in *Information Theory, 2000. Proceedings. IEEE International Symposium on*, 2000, pp. 26–.

[5] A. Garivier, "Consistency of the unlimited BIC context tree estimator," *Information Theory, IEEE Transactions on*, vol. 52, no. 10, pp. 4630–4635, Oct 2006.

[6] Z. Talata and T. Duncan, "BIC context tree estimation for stationary ergodic processes," *Information Theory, IEEE Transactions on*, vol. 57, no. 6, pp. 3877–3886, June 2011.

[7] A. Abdi and F. Fekri, "Source identification and compression of mixture data from finite observations," in *Information Theory Workshop - Fall (ITW), 2015 IEEE*, Oct 2015, pp. 29–33.

[8] I. Csiszar and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *Information Theory, IEEE Transactions on*, vol. 52, no. 3, pp. 1007–1016, March 2006.

[9] Y. Ephraim and N. Merhav, "Hidden markov processes," *Information Theory, IEEE Transactions on*, vol. 48, no. 6, pp. 1518–1569, Jun 2002.

[10] A. Beirami and F. Fekri, "Memory-assisted universal source coding," in *Data Compression Conference (DCC), 2012*, April 2012, pp. 392–392.

[11] A. Beirami, M. Sardari, and F. Fekri, "Results on the optimal memory-assisted universal compression performance for mixture sources," in *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, Oct 2013, pp. 890–895.

[12] M. Mahoney, "Adaptive weighing of context models for lossless data compression," Florida Tech., Technical Report , CS-2005-16, 2005, 2005.