Data-Driven Learning of the Number of States in Multi-State Autoregressive Models

Jie Ding, Mohammad Noshad, and Vahid Tarokh

Abstract

In this work, we consider the class of multi-state autoregressive processes that can be used to model non-stationary time-series of interest. In order to capture different autoregressive (AR) states underlying an observed time series, it is crucial to select the appropriate number of states. We propose a new model selection technique based on the Gap statistics, which uses a null reference distribution on the stable AR filters to check whether adding a new AR state significantly improves the performance of the model. To that end, we define a new distance measure between AR filters based on mean squared prediction error (MSPE), and propose an efficient method to generate random stable filters that are uniformly distributed in the coefficient space. Numerical results are provided to evaluate the performance of the proposed approach.

I. INTRODUCTION

The data that many of the practical areas deal with are in the form of time-series, and therefore, modeling and forecasting time-series has fundamental importance. There may be occasional changes in the behavior of time series. Some examples are the changes in the stock market due to the financial crisis, or the variations of the EEG signal caused by the mode change in the brain. In the econometrics literature, this kind of time series is referred to as "regime-switching" model [1, 2]. In regime switching models, there are $M$ states in the time-series, and the set of data points in state $m$ is denoted by $S_m$, $m = 1, 2, \cdots, M$. Given a time-series $\{x^{(n)}\}_{n=1}^N$ and assuming that $x^{(n)}$ belongs to state $m$, the probability density function (pdf) of $x^{(n)}$ conditioned on its past is given by $P_m(x^{(n)}|x^{(n-1)}, \cdots, x^{(1)})$. A more detailed survey on this model can be found in [3].

The autoregressive (AR) model, one of the commonly used techniques to model stationary time-series [2], is usually used to model each state in regime-switching model. In Gaussian form of this model, the value at time-instance $n$ is assumed to be normally distributed with mean $\gamma^T_m x^{(n)}$ and variance $\sigma^2_m$, i.e., $x^{(n)} \sim \mathcal{N}(\gamma^T_m x^{(n)}, \sigma^2_m)$. Here $x^{(n)} = [1, x^{(n-1)}, \ldots, x^{(n-L)}]$. $\gamma_m = [\gamma_{m0}, \gamma_{m1}, \gamma_{m2}, \cdots, \gamma_{mL}]^T$ is a real-valued vector of length $L$ that characterizes state $m$, and $\sigma^2_m$ is the noise variance associated with state $m$. We refer to this model as multi-state AR model and to $\gamma_m$ as the AR filter or AR coefficients of state $m$. The above model with $\gamma_m = 1$ was first analyzed by Lindgren [4] and Baum et al. [5]. The model with general $\gamma_m$ is widely studied in the speech recognition literature. We note that the multi-state AR model is a general statistical model that can be used to fit data in many real-world applications. It was shown that the model is capable of modeling non-linear and non-stationary time series with multimodal conditional distributions and with heterogeneous (6). There are two basic underlying assumptions in this model: 1. AR dependencies assumption, which is reasonable if the observations are obtained sequentially in time; 2. Multi-state assumption, which is reasonable if the stochastic process exhibits different behaviors in different time intervals. For example, in the case of business cycles or financial crises, stock prices may have dramatic while not permanent changes, and those dynamics can be described by stochastic transitions among different states.

Despite the wide applications of the multi-state AR model, there are few results on how to estimate the number of states $M$ in a time-series. Obviously, different values of $M$ will produce a nested family of models and a model with larger $M$ will fit the observed data better. The drawback of using complex models with large $M$ is the over-fitting problem which decreases the predictive power of the model. Hence, a proper model selection procedure that identifies the appropriate number of states is vital. It is natural to test the null hypothesis that there are $M$ states against the alternative of $M + 1$. Unfortunately, the likelihood ratio test of this hypothesis fails to satisfy the usual regularity conditions since some parameters of the model are unidentified under the null hypothesis. An alternative is to apply Akaike information criterion (AIC) [7] or Bayesian information criterion (BIC) [8] to introduce a penalty on the complexity of the model and estimate the number of AR states. However, in general AIC and BIC are shown to be inaccurate in estimating the number of states [9].

In this paper, we propose a new model selection criterion. Specifically we propose a technique to determine the appropriate number of states $M$ by extending the Gap statistics [10] to time-series. The idea is to identify $M$ by comparing the goodness of fit for the observed data with its expected value under a null reference distribution. To that end, we first draw a reference curve which plots the "goodness of fit" versus $M$ and describes how much adding new AR states improves the goodness of fit, based on the most non-informative distributed data. We then draw a similar curve based on the observed data. In this work we choose the "goodness of fit" measure to be the mean squared prediction error (MSPE). Finally, the point at which the gap...
between the two curves is maximized is chosen as the estimated $M$. Besides the simplicity and effectiveness, another benefit of the proposed model selection criterion is that it is adaptive to the underlying characteristics of AR processes. The criterion for the processes that evolves smoothly (i.e., the roots of whose characteristic polynomial are small) is different from the criterion for those that evolve quickly. In this sense, it takes into account the characteristics behind the observed data in an unsupervised manner, even though no domain knowledge or prior information is given.

The remainder of the paper is organized as follows. In Section II we propose the Gap statistics for estimating the number of AR states in a time-series. Section III formulates a specific class of the multi-state AR model, where the transitions between states are assumed to be a Markov process. We emphasize that this parametric model is considered primarily for simplicity and the proposed Gap statistics can be applied to general multi-state AR process. A new initialization approach is also proposed that can effectively reduce the impact of a bad initialization on the performance of expectation-maximization (EM) algorithm. Section IV presents some numerical results to evaluate the performance of the proposed approach. Experiments show that the accuracy of the proposed approach in estimating the number of AR states surpasses those of AIC and BIC.

II. GAP STATISTICS

Computing each point on the reference curve using the new distance measure turns out to be a clustering problem in the space of AR coefficients with a fixed length, which is solved using Monte Carlo approach. To that end, we introduce an approach to generate AR filters that are randomly distributed in a bounded space, and apply the $k$-medoids algorithm to approximate the optimal solution for the clustering problem. We first outline our proposed model selection criterion in Subsection II-A and then elaborate on the distance measure in Subsections II-B and the generation of random AR filters in Subsections II-C.

A. The Model Selection Criterion

We use superscript $(n)$ to represent the data at $n$th time-instance, and $\mathcal{N}(\mu, \sigma^2)$ to denote the normal distributions of mean $\mu$ and variance $\sigma^2$. Symbols in bold face represent vectors or matrices. We start from a simple scenario where the data $X = x^{(n)}$, $n = 1, \ldots$ is generated using a single stable AR filter $\psi_A$:

$$x^{(n)} = \psi_A^T x^{(n)} + \varepsilon^{(n)}, \quad \varepsilon^{(n)} \sim \mathcal{N}(0, \sigma_A^2).$$

Consider that we are at time-instance $n - 1$ and want to predict the value at time $n$. If $\hat{x}^{(n)} = \psi_A^T x^{(n)}$ is used for prediction, the MSPE is $E\{(x^{(n)} - \psi_A^T x^{(n)})^2\} = \sigma_A^2$. If another AR filter is used for prediction instead of $\psi_A$, i.e., $\hat{x}^{(n)} = \psi_B^T x^{(n)}$, the MSPE becomes $E\{(x^{(n)} - (\psi_B)^T x^{(n)})^2\}$. We define the distance between $\psi_A$ and $\psi_B$ based on their MSPE as follows

$$D(\psi_A, \psi_B) = E\{(x^{(n)} - \psi_B^T x^{(n)})^2\} - \sigma_A^2 = E\left\{\left(\psi_A - \psi_B\right)^T x^{(n)}\right\}^2.$$ (2)

It is easy to observe that $D(\psi_A, \psi_B)$ is always nonnegative, which means using mismatch filter for prediction increases the MSPE. We provide more detailed discussions on the distance in the next subsection.

As has been mentioned in Section I our model selection criterion is based on a reference curve that describes how much adding a new state increases the goodness of fit in “the worst case”. To that end, we consider an $M$-state AR process where at each time-instance $n$, random mismatch coefficients is applied to be the linear predictor. We need to find such $M$ filters that the average mismatch error is minimized. It turns out to be the following clustering problem in the space of stable AR filters $R_L(r) \subset \mathbb{R}^L$, where

$$R_L(r) = \{(\lambda_1, \ldots, \lambda_L) | z^L + \sum_{\ell=1}^L \lambda_\ell z^{L-\ell} = \prod_{\ell=1}^L (z - a_\ell), \lambda_\ell \in \mathbb{R}, |a_\ell| \leq r, 1 > r > 0, \ell = 1, \ldots, L\}.

Clustering of Stable Filters: For a fixed $M$, let $\mathcal{F} = \{\psi_1, \psi_2, \ldots, \psi_F\}$ be a set of uniformly generated stable filters of a given length $L$. We cluster $\mathcal{F}$ into $M$ disjoint clusters $C_1, C_2, \ldots, C_M$, and define the within cluster sum of distances to be

$$W_M = \min_{\gamma_1, \ldots, \gamma_M} \left\{\frac{1}{F} \sum_{m=1}^M \sum_{\psi \in C_m} D(\gamma_m, \psi)\right\} + 1,$$ (3)

where $D(\gamma_m, \psi)$ is defined in (2) and will be further simplified in (6) and (7). By computing $\log(W_M)$ for $M = 1, \ldots, M_{\text{max}}$, we obtain the reference curve. The optimization problem (3) is solved by the $k$-medoids algorithm [11].

The model selection criterion is outlined in Table I We note that the bound for the roots $0 < r < 1$ is determined by the estimated filters, and thus the reference is data-dependent. Intuitively, if the process is very “smooth”, i.e., the roots of the characteristic polynomials of each AR process are close to zero, the MSPE curve would have smaller values. Thus, the random filters from which the reference curve is calculated should also be drawn from a smaller bounded space.
Algorithm 1 Model Selection Based on Gap Statistics

Input: \( X = \{ x^{(n)} \}_{n=1}^{N} \), \( M_{\text{max}} \) (which is assumed to contain the “correct” number of states)

Output: \( M_{\text{opt}} \).

1. \( \text{for } M = 1 \rightarrow M_{\text{max}} \) \( \text{do} \)
2. \( \text{Fit a multi-state AR model to the data (for instance using the EM algorithm described in Algorithm 3)} \)
3. \( \text{Compute the MSPE } W_M \text{ based on the estimated model.} \)
4. \( \text{end for} \)
5. \( \text{Plot } \log(W_M), M = 1, \ldots, M_{\text{max}}, \text{ referred to as the “observed MSPE curve”} \)
6. \( \text{Compute the largest absolute value of the roots of each estimated AR filter for the case } M = M_{\text{max}}, \text{ denoted by } r_1, \ldots, r_{M_{\text{max}}}. \)
7. \( \text{Let } r = \min\{\max\{r_1, \ldots, r_{M_{\text{max}}}\}, 1\}. \)
8. \( \text{for } \ell = 1 \rightarrow \text{Iter} \text{ (number of instances to draw a set of random filters)} \) \( \text{do} \)
9. \( \text{Run Algorithm 3 (to be introduced in Subsection II-C) to randomly generate } F \text{ independent stable filters } \hat{\Psi} = \{ \psi_1, \psi_2, \ldots, \psi_F \} \)
10. \( \text{for } M = 1 \rightarrow M_{\text{max}} \) \( \text{do} \)
11. \( \text{Run Algorithm 2 to approximate the optimum of } \hat{\Psi}, \text{ and obtain } \log(W_{M\ell}), M = 1, \ldots, M_{\text{max}}. \)
12. \( \text{end for} \)
13. \( \text{end for} \)
14. \( \text{Select the number of AR states } M_{\text{opt}} \text{ to be the first } M \text{ that } \log(W_{M\ell}) - \log(W_M) \geq \log(W_{M\ell+1}) - \log(W_{M\ell+1}), M = 1, \ldots, M_{\text{max}}. \)

Algorithm 2 Clustering Stable AR filters via “k-medoids” Algorithm

Input: A set of stable filters \( \hat{\Psi} = \{ \psi_1, \ldots, \psi_F \} \), the number of desired clusters \( M \), a small positive number \( \delta \) (used for the stopping criterion).

Output: The minimum within-cluster sum of distances (WCSD) \( w_{\ell} \) and \( \{ c_{\ell1}, \ldots, c_{\ellM} \} \subset \hat{\Psi} \) that approximate the \( C \) centers.

1. \( \text{Generate a matrix } D_{F \times F} \text{ whose elements are pairwise distances between filters: } D_{uv} = D(\psi_u, \psi_v). \)
2. \( \text{Initialize } M \text{ clusters, which are characterized by } M \text{ centers } c_1, \ldots, c_{ML} \text{ and associated } M \text{ sets of indices } I_1, \ldots, I_M \text{ that form a partition of } \{1, \ldots, F\}. \)
3. \( \text{Compute } w_{\ell1} = \sum_{m=1}^{M} \sum_{u \in I_m} D(\psi_{c_m}, \psi_u). \text{ Let } w_0 = \text{Inf, } \ell = 1. \)
4. \( \text{while } w_{\ell-1} > \delta w_{\ell-1} \) \( \text{do} \)
5. \( \ell = \ell + 1 \)
6. \( \text{while } \hat{w}_{\ell-1} \geq w_{\ell-1} \text{ and there exists a set of } M \text{ different indices } \{c_m\}_{m=1}^{M} \text{ not examined yet} \) \( \text{do} \)
7. \( \text{Change the existing centers } \{c_m\}_{m=1}^{M} \text{ to } \{c_{m'}\}_{m=1}^{M} \neq \{c_{m'}\}_{m=1}^{M}. \)
8. \( \text{For each } u = 1, \ldots, F, \text{ let the } \ell \text{th filter } \psi_u \text{ belong to the cluster whose center is the closest to it, i.e., } u \in I_m \text{ if } D(\psi_{c_m}, \psi_u) \leq D(\psi_{c_{m'}}, \psi_u), \text{ for each } m, m' = 1, \ldots, M. \text{ Update the WCSD given the new clusters: } \hat{w}_{\ell-1} = \sum_{m=1}^{M} \sum_{u \in I_m} D(\psi_{c_m}, \psi_u). \)
9. \( \text{end while} \)
10. \( \text{end while} \)
11. \( \text{if } \hat{w}_{\ell-1} < w_{\ell-1} \text{ then} \)
12. \( w_{\ell} = \hat{w}_{\ell-1}. \)
13. \( \text{end if} \)
14. \( \text{end while} \)

B. Distance Measure for Autoregressive Processes

In this subsection, we provide the explicit formula for the distance in Equation (2). Assume that the data is generated by a stable filter \( \psi_A \) of length \( L \). Let \( \Psi_A(z) = \sum_{\ell=1}^{L} \psi_Az^{-\ell} \) be the characteristic polynomial of \( \psi_A \), and \( a_1, \ldots, a_L \) denote the roots of \( 1 - \Psi_A(z) \), i.e., \( 1 - \Psi_A(z) = \prod_{\ell=1}^{L} (1 - a_{\ell}z^{-1}) \), where \( a_1, \ldots, a_L \) lie inside the unit circle (\( |a_{\ell}| < 1 \)). Similarly define \( \Psi_B(z), b_1, \ldots, b_L \) for \( \psi_B \). The value in (2) can be computed using the power spectral density and Cauchy’s integral theorem as:

\[
D(\psi_A, \psi_B) = D_0(\psi_A, \psi_B) + \left( \frac{1}{1 - \sum_{\ell=1}^{L} \psi_{B\ell}} \right) ^{2},
\]

(4)

where \( D_0(\psi_A, \psi_B) = \frac{\sigma^2}{2\pi} \int_{-\pi}^{\pi} \frac{\left| \Psi_A(e^{j\omega}) - \Psi_B(e^{j\omega}) \right|^2}{\left| 1 - \Psi_A(e^{j\omega}) \right|^2} d\omega = \sigma^2 \sum_{k=1}^{L} \prod_{\ell=1}^{L} (a_k - b_{\ell}) \left( \frac{\prod_{\ell=1}^{L} (1 - a_kb_{\ell}^*)}{\prod_{\ell=1}^{L} (1 - a_ka_{\ell}^*)} - 1 \right), \)

(5)

where \( a^* \) denotes the complex conjugate of \( a \).
Remark 1. For now we assume that \( \varphi(x) \) has zero mean and use \( D_{0}(\cdot) \) in Equation (6) instead of \( D(\cdot) \) in Equation (4) to compute the reference curve. The derived reference curve can be applied to the general case. The reason is that it is more difficult to detect two AR states with the same mean than those that have different means. Therefore, the reference curve for the zero mean case (the “worst” case) is expected to work well in general.

The distance measure defined in Equation (6) is proportional to \( \sigma_{A}^{2} \). We consider \( \sigma_{A}^{2} = \sigma^{2} \) which results in a constant \( \log \sigma^{2} \) in the computation of \( \log W_{M} \) in (4). Since it is the same for different \( M \), without loss of generality we further set \( \sigma^{2} = 1 \).

The distance between two AR filters can be explicitly expressed in terms of the coefficients. This is computationally desirable if the filters are random samples generated in the coefficient domain, as will be discussed in Subsection II-C.

Notations: Consider two polynomials of nonnegative powers \( p(z) \) and \( q(z) \) respectively of degrees \( u \geq 0 \) and \( v \geq 0 \). Let \( \overline{q}(z), pq(z) \) respectively denote the reciprocal polynomial of \( q(z) \), and the multiplication of \( p(z) \) and \( q(z) \), i.e., \( \overline{q}(z) = z^{v}q(z^{-1}), pq(z) = p(z)q(z) \). Let \( \text{Res}(p(z), q(z)) \) be the resultant of \( p(z) \) and \( q(z) \). Define \( \text{Po}(p(z), q(z)) = \sum_{k=1}^{u} q(\alpha_{k}) \) and \( \text{Po}(p(z), 0) = 0 \), where \( \alpha_{1}, \cdots, \alpha_{u} \) are the roots of \( p(z) \).

Lemma 1. The value of \( \text{Res}(p(z), q(z)) \) and \( \text{Po}(p(z), q(z)) \) can be computed as polynomials in terms of the coefficients of \( p(z) \) and \( q(z) \).

The proof follows from the fact that the resultant of \( p(z) \) and \( q(z) \) is given by the determinant of their associated Sylvester matrix \([12]\), and that for any \( n \in \mathbb{N} \), \( \sum_{k=1}^{n} a_{k} \) can be computed as polynomials in the coefficients of \( p(z) \) via Newton’s identities. We further provide the following result.

Lemma 2. Let \( p_{A}(z) = z^{L}(1-\Psi_{A}(z)) = \prod_{l=1}^{L}(z-a_{l}), p_{B}(z) = z^{L}(1-\Psi_{B}(z)) = \prod_{l=1}^{L}(z-b_{l}), p'_{A}(z) = \partial(zp_{A}(z))/\partial z \). The value of \( D_{0}(\psi_{A}, \psi_{B}) \) in Equation (6) (with \( \sigma_{A} = 1 \)) can be computed in terms of the coefficients of \( \psi_{A}, \psi_{B} \) as \( D_{0}(\psi_{A}, \psi_{B}) = \frac{\text{Po}(p_{A}(z), p_{B}(z)\overline{p'_{A}(z)})S([u_{1}, \cdots, u_{L-1}], 0) - \text{Po}(p_{A}(z), p_{B}p'_{A}(z)S([u_{1}, \cdots, u_{L-2}], p'_{A}(z)))}{\text{Res}(p_{A}(z), p'_{A}(z))} \)

where \( u_{i} = \text{Po}(p_{A}(z), (p'_{A}(z)^{i})^{j}) \), \( v_{i} = \text{Po}(p_{A}(z), (p'_{A}(z)^{i})^{j}) \), \( i = 1, \cdots, L-1 \), and the function \( S(\cdot, \cdot) \) is defined as (where \( \text{det}(\cdot) \) denotes the determinant of a square matrix)

\[
S([s_{1}, \cdots, s_{h}], t) = \frac{1}{h!} \text{det} \left( \begin{array}{cccc}
1 & 0 & \cdots & 0 \\
-s_{2} - t^{2} & s_{1} - t & 2 & \cdots \\
& \vdots & \ddots & \vdots \\
& & \ddots & \ddots & h-1 \\
-s_{h-1} - t^{h-2} & -t^{h-1} & \cdots & s_{2} - t^{2} & s_{1} - t
\end{array} \right)
\]

for \( h > 0 \), \( S(\cdot, \cdot) = 1 \) for \( h = 0 \), and \( S(\cdot, \cdot) = 0 \) for \( h < 0 \).

C. Generation of Random Uniform Filters with Bounded Roots

As mentioned before, Gap statistics requires a reference curve that is calculated by clustering the filters randomly chosen from a reference distribution. In some scenarios we need to generate sample filters from \( R_{L}(r) \), where \( r \) is calculated from the observed data. Inspired by the work of Beadle and Djurić [13], we provide the following result on how to generate a random point in \( R_{L}(r) \) with uniform distribution.

Lemma 3. Generation of an independent uniform sample of \( (\lambda_{1, L}, \cdots, \lambda_{L, L}) \in R_{L}(r) \) can be achieved by the following procedure:
1. Draw \( \lambda_{1,1} \) uniformly within the interval \([-r, r] \);
2. For \( k = 2, \cdots, L \), assume that we have obtained a random uniform sample. Draw \( \lambda_{k,k} \) from a pdf proportional to the following term within the interval \([-r^{k}, r^{k}] \)

\[
\left(1 + \frac{\lambda_{k,k}}{y^{k}} \right) \left(1 - \frac{\lambda_{k,k}}{y^{k}} \right)
\]

where \( \lambda_{i,k} = \lambda_{i,k-1} + \frac{\lambda_{k,k} y^{k} - y^{k-1}}{y^{k} - y^{k-1}} \), \( i = 1, \cdots, k-1 \).
Remark 2. We assume that the reference distribution is uniform with radius determined by the data and is no larger than 1, as shown in Step 6 of Algorithm [2] with the intuition that the extent to which a time series changes may be revealed by the observed data in an unsupervised manner. Besides this, the distance in Equation [9] is derived only for stable filters. If we are to generate a uniform random point in $R_L(r)$, the technique presented in Lemma [9] can be equivalently formulated in a specially simple way which is summarized in the following lemma. The procedure is also described in Algorithm [3]. The proof of the lemmas will be included in a longer version of this paper.

Lemma 4. A sample of $(\lambda_1, L, \cdots, \lambda_{L,L})$ that is uniformly distributed in $R_L(r)$ can be generated by recursion $\Lambda_0(z) = 1, \Lambda_k(z) = z\Lambda_{k-1}(z) + r^k\alpha_k\Lambda_{k-1}(z/r^2)$, where $\alpha_k = 2\beta_k - 1$ and $\beta_k \sim \text{Beta}([k/2 + 1], [(k + 1)/2]), k = 1, \cdots, L$ are independently generated.

Algorithm 3 Generating a uniform sample $(\lambda_1, L, \cdots, \lambda_{L,L})$ within $R_L(r)$

Input: $L$.
Output: $\Lambda_L(z) = z^L + \sum_{\ell=1}^L \lambda_{\ell,L}z^{L-\ell}$.
1. for $k = 1 \to L$
2. randomly draw $\beta_k$ from the beta distribution $\beta_k \sim \text{Beta}([k/2 + 1], [(k + 1)/2])$
3. Let $\alpha_k = 2\beta_k - 1$ and $\Lambda_k(z) = z\Lambda_{k-1}(z) + r^k\alpha_k\Lambda_{k-1}(z/r^2)$.
4. end for

III. Model

A popular model for describing these switching behaviors is hidden Markov model (HMM) [14], in which the states are switched between each other according to a Markov process. In the HMM model, the transition probability from state $m$ to state $m'$ is independent of $n$ and is denoted by $T_{mm'}$, i.e., $P(x^{(n)} \in S_m, x^{(n+1)} \in S_{m'}) = T_{mm'}$. Although the method proposed in Section II for estimating the number of AR states is a general technique that is applicable to any multi-state AR model, for simplicity we use HMM to describe the transition between states.

A. Notations and Formulations

Let $X = \{x^{(n)}\}_{n=1}^N$ be a correlated time-series and $x^{(n)} = [1, x^{(n-1)}, \cdots, x^{(n-L)}]^T$. Then $\gamma_m^T x^{(n)}$, $m = 1, \cdots, M$, is a linear function that describes the auto-regressive model of the system at state $m$, where $\gamma_m = [\gamma_{m0}, \gamma_{m1}, \gamma_{m2}, \cdots, \gamma_{mL}]^T$ and $\gamma_{m\ell} \in \mathbb{R}$, for $\ell = 0, \cdots, L$. The error term at time instance $n$, $x^{(n)} - \gamma_m^T x^{(n)}$, is modeled by normal distribution $\varepsilon^{(n)} \sim \mathcal{N}(0, \sigma_m^2)$. Throughout the paper, $x^{(-L+1)}, \cdots, x^{(0)}$ are assumed to be fixed and known for brevity. Let $Z = \{z^{(n)}\}_{n=1}^N$ and $Y = \{y^{(n)}\}_{n=1}^N$ be a sequence of missing (unobserved) indicators, where $z^{(n)}$ is a $M \times 1$ vector, and

$$z^{(n)}_m = \begin{cases} 1 & \text{if } x^{(n-1)} \in S_m, x^{(n)} \in S_{m'}, \\ 0 & \text{otherwise,} \end{cases}$$

$$y^{(n)}_m = \begin{cases} 1 & \text{if } x^{(n)} \in S_m, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, $z^{(n)} = y^{(n-1)}(y^{(n)})^T$. We note that $y^{(n)}$ is a binary vector of length $M$ containing a unique “1”, while with a slight abuse of notation $y^{(n)}$ is the location of that “1”. We assume that $\{y^{(n)}\}_{n=1}^N$ is a Markov chain with probability transition matrix $T$, and $y^{(1)}$ is drawn from $\mathcal{M}(\alpha_1, \cdots, \alpha_M)$, where $\mathcal{M}$ denotes the multinomial distributions. In other words, the assumed data generating process (given a fixed $M$) is:

$$y^{(n)} \sim \begin{cases} \mathcal{M}(\alpha_1, \cdots, \alpha_M) & \text{if } n = 1, \\ \mathcal{M}(T_{y^{(n-1)}}, \cdots, T_{y^{(n)}}, M) & \text{otherwise,} \end{cases} \quad \text{(8)}$$

$$X^{(n)} \sim \mathcal{N}(\gamma_{y^{(n)}}^T x^{(n)}, \sigma_{y^{(n)}}^2), \quad n = 2, \cdots, N. \quad \text{(9)}$$

Let $\Theta = \{T, \gamma_m, \sigma_m^2 \mid m = 1, \cdots, M \}$ be the set of unknown parameters to be estimated. Though computing the maximum-likelihood estimation (MLE) of the above probabilistic model [9] is not tractable, it can be approximated by a local maxima via EM algorithm [15]. The EM algorithm produces a sequence of estimates by recursive application of E-step and M-step to the complete log-likelihood till a predefined convergence criterion is achieved. The complete log-likelihood can be written as

$$\sum_{n=1}^N \log p(x^{(n)} \mid x^{(n)}) = \sum_{n=1}^N \sum_{m=m', m''}^{N} z_{m,m''}^{(n)} \left( \log \frac{T_{mm'}}{\sqrt{2\pi}\sigma_{m''}} + \frac{(x^{(n)} - \gamma_{m'}^T x^{(n)})^2}{2\sigma_{m''}^2} \right). \quad \text{(10)}$$
For brevity, we provide the EM formulas below without derivation. In the E-step, we obtain a function of unknown parameters by taking the expectation of \( Q(\Theta \mid X, \Theta^{\text{old}}) \) w.r.t missing data \( Y \) and \( Z \) given the most updated parameters,

\[
Q(\Theta \mid X, \Theta^{\text{old}}) = \sum_{n=1}^{N} \sum_{m,m'=1}^{N} w^{(n)}_{mm'} \left( \log \left( \frac{T_{mm'}}{\sqrt{2\pi \sigma_{m'}}} \right) + \frac{(x^{(n)} - \gamma_{m}^T x^{(n)})^2}{2\sigma_{m'}^2} \right),
\]

where

\[
w^{(n)}_{mm'} = E(z^{(n)}_{mm'} \mid \Theta^{\text{old}}) = P(y^{(n)-1} = m, y^{(n)} = m' \mid X)
\]
can be computed recursively. We note that the parameters involved in the right-hand side of (12) take values from the last update. The “old” superscriptions are omitted for brevity.

\[
\gamma_{m} = \left( \sum_{n=1}^{N} \sum_{m'=1}^{M} w^{(n)}_{mm'} x^{(n)} (x^{(n)})^T \right)^{-1} \left( \sum_{n=1}^{N} \sum_{m'=1}^{M} w^{(n)}_{mm'} x^{(n)} x^{(n)} \right) \quad \sigma_{m}^2 = \frac{\sum_{n=1}^{N} \sum_{m'=1}^{M} w^{(n)}_{mm'} (x^{(n)} - (\gamma_{m})^T x^{(n)})^2}{\sum_{m'=1}^{M} \sum_{n=1}^{N} w^{(n)}_{mm'}}.
\]

**B. Initialization of EM**

The final results of EM strongly depend on the initialization, and an inappropriate initialization can cause the EM algorithm to converge to a local maximum that is far away from the global optimum. A common solution for this problem is to use multiple random initializations and choose the output with the highest likelihood \([16]\), which can be significantly time consuming. Here, we use a new initialization technique to get a reliable output from EM. This technique is based on the fact that in the time-series in economic or other fields, the self-transition probability of the states is usually close to one, i.e., \( T_{mm} \approx 1 \), and hence, by adopting this assumption we can use the initialization method described in Table 4 which is empirically observed to produce more reliable and efficient EM results.

**Algorithm 4 EM algorithm with the proposed initialization approach**

Input: \( X = \{x^{(n)}\}_{n=1}^{N} \).
Output: \( \hat{\Theta} \).

1: for \( M = 1 \to M_{\text{Max}} \) do
2: \hspace{1em} if \( M > 1 \) then
3: \hspace{2em} Retrieve “good” estimate from the previous EM (for \( M = 1 \)): \( \hat{\gamma}_{1}, \ldots, \hat{\gamma}_{M-1}, \hat{\sigma}_{1}^2, \ldots, \hat{\sigma}_{M-1}^2 \)
4: \hspace{1em} end if
5: \hspace{1em} for \( n = 1 \to N - N_0 + 1 \) do
6: \hspace{2em} Estimate the AR coefficients \( \hat{\xi}_i \) and the noise variance \( \hat{\sigma}_i^2 \) from the sequence \( \{x^{(n)}, \ldots, x^{(n+N-0-1)}\} \) via linear least squares method.
7: \hspace{1em} end for
8: \hspace{1em} Cluster \( \hat{\xi}_1, \ldots, \hat{\xi}_N,N-N_0+1 \) into \( M \) cluster using k-means algorithm and obtain the centers \( \hat{\theta}_1^2, \ldots, \hat{\theta}_M^2 \) with the corresponding noise variances \( \hat{\sigma}_1^2, \ldots, \hat{\sigma}_M^2 \). Pick up such \( \hat{\theta}_m \), that maximize the sum of Euclidean distances to \( \gamma_1, \ldots, \gamma_{M-1} \).
9: \hspace{1em} Set the initial parameters to be \( \{\gamma_1, \ldots, \gamma_{M-1}, \theta_m\} \) and \( \{\hat{\sigma}_1^2, \ldots, \hat{\sigma}_M^2, \hat{\xi}_m\} \).
10: Run EM updates described in (12)-(14) till certain stopping criterion is achieved.
11: end for

Clearly, if \( L << N/N_0 << 1/T_{mm} \), \( \hat{\xi}_i \)'s are normally distributed around the true AR coefficients, \( \gamma_m \)'s. This “split” style rule is used elsewhere (e.g. Ueda et al. \([17]\)).

**IV. Numerical Experiments**

We have generated synthetic data under three different scenarios to evaluate the performance of the proposed model selection technique: Scenario 1: \( (L, M, r) = (4, 3, 1) \); Scenario 2: \( (L, M, r) = (1, 4, 0.8) \); Scenario 3: \( (L, M, r) = (2, 2, 0.6) \). For each scenario, 100 instances of multi-state AR time series of length 1000 are independently generated, each of which consists of \( M \) ARs which are uniformly drawn from the space \( R_L(r) \). For each AR the mean is uniformly generated from \([-4, 4]\) and variance is 1. The transition matrix is \( T_{mm} = 0.98, T_{mm'} = 0.02/(M - 1), m, m' = 1, \ldots, M, m \neq m' \). For each of the 100 instances, the model parameters for each fixed \( M = 1, \ldots, M_{\text{Max}} = 6 \) are estimated using EM algorithm. Table 1 shows the estimated number of AR filters using BIC, AIC and Gap statistics. Two types of Gap statistics are used to estimate the number of states. In the first one that is represented by Gap (U), the reference curve is generated from sample AR filters that have roots inside the unit circle, and therefore, is independent of the data. In the second form of the Gap statistics, represented by Gap (B), the sample filters are restricted to have roots inside a circle with radius \( r \), where \( r \) is calculated from the data according to Algorithm 1. As it can be observed, Gap statistics outperforms AIC and BIC in all three scenarios. In the third scenario, the Gap (B) gives a better estimate of the number of states compared to Gap (U) since it is data dependent.
Fig. 1: 10000 independent uniform random stable filters of $L = 2$ and the centers for 2 of clusters, for $r = 0.6, 0.8, 1$

Fig. 2: The reference curves for $r = 0.6, 0.8, 1, L = 4$, which are obtained based on $Iter = 32, F = 1000$ (see Algorithm 1).

Fig. 3: A random instance of multi-state AR time-series: $L = 4, M = 3, T_{mm} = 0.98, \mu_m = 0, \sigma_m = 1, T_{mm'} = 0.01, m, m' = 1, \cdots, 3, m \neq m'$.

Fig. 4: The reference curves and the observed MSPE curve for the time series shown in Fig. 3. The gap between the two curves is maximized at $M = 3$.

V. CONCLUSIONS

In this paper we propose a model selection technique to estimate the number of states in a time-series. The proposed approach, which is based on the Gap statistics, uses a reference curve to check whether adding a new state significantly decreases the prediction error. The reference curve is calculated by clustering uniformly generated stable AR filters. The numerical results show that the performance of the proposed model selection criterion surpasses that of AIC and BIC.

| Method | Scenario 1 | Scenario 2 | Scenario 3 |
|--------|------------|------------|------------|
|        | AIC | BIC | Gap (U) | Gap (B) | AIC | BIC | Gap (U) | Gap (B) | AIC | BIC | Gap (U) | Gap (B) |
| Estimated Number of AR Filters | 1 | 0 | %20 | 0 | 0 | %10 | 0 | 0 | %30 | %70 | %40 | 0 |
| 2 | %40 | %80 | %20 | 0 | %60 | %80 | %10 | 0 | %50 | %30 | %60 | %40 |
| 3 | %60 | 0 | %80 | %100 | %40 | %10 | %60 | %10 | %20 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | %30 | %90 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

TABLE I: The estimated number of AR filters for three different scenarios using AIC, BIC and Gap statistics (with the true number of filters for each scenario highlighted)
References:
[1] S.M. Goldfeld and R.E. Quandt. A Markov model for switching regressions. *J. Econometrics*, 1(1):3–15, 1973.
[2] J.D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.
[3] J.D. Hamilton. Regime-switching models. *The New Palgrave Dictionary of Economics*, 2, 2008.
[4] G. Lindgren. Markov regime models for mixed distributions and switching regressions. *Scand. J. Stat.*, pages 81–91, 1978.
[5] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, pages 164–171, 1970.
[6] C. S. Wong and W. K. Li. On a mixture autoregressive model. *J. Roy. Statist. Soc. Ser. B*, 62(1):95–115, Sep 2000.
[7] H. Akaike. Information theory and an extension of the maximum likelihood principle. *2nd Int. Sym. Info. Theory*, pages 267–281, Sep 1973.
[8] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, Mar 1978.
[9] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *J. Classification*, 13(2):195–212, 1996.
[10] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *J. Roy. Statist. Soc. Ser. B*, 63(2):411–423, 2001.
[11] L. Kaufman and P. Rousseeuw. *Clustering by means of medoids*. North-Holland, 1987.
[12] B. Sturmfels. Introduction to resultants. In *Proceedings of Symposia in Applied Mathematics*, volume 53, pages 25–40, 1998.
[13] E.R. Beadle and P.M. Djuric. Uniform random parameter generation of stable minimum-phase real ARMA (p,q) processes. *IEEE Signal Process. Lett.*, 4(9):259–261, Sep 1997.
[14] A Portitz. Linear predictive hidden markov models and the speech signal. In *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, volume 7, pages 1291–1294. IEEE, 1982.
[15] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, pages 1–38, 1977.
[16] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.
[17] N. Ueda, R. Nakano, Z. Ghahramani, and G.E. Hinton. SMEM algorithm for mixture models. *Neural Comput.*, 12(9):2109–2128, 2000.