Signal Processing 149 (2018) 68–81

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Symbolic analysis-based reduced order Markov modeling of time series data☆

Devesh K. Jha[a,b], Nurali Virani[a,c], Jan Reimann[d], Abhishek Srivastav[e], Asok Ray[a,d,*]

[a] Department of Mechanical and Nuclear Engineering, Pennsylvania State University, University Park, PA 16802, USA
[b] Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, USA
[c] Machine Learning Laboratory, GE Global Research Center, Niskayuna, NY, USA
[d] Department of Mathematics, Pennsylvania State University, University Park, PA 16802, USA
[e] Machine Learning Laboratory, GE Global Research Center, San Ramon, CA, USA

ARTICLE INFO

Article history:
Received 24 October 2017
Revised 17 February 2018
Accepted 11 March 2018
Available online 12 March 2018

Keywords:
Symbolic dynamics
Markov modeling
Model order reduction

ABSTRACT

This paper presents reduced-order modeling of time-series data for a special class of Markov models using symbolic dynamics. These models are constructed from the time-series signal by partitioning the data and then inferring a probabilistic finite state automaton (PFSA) from the resulting symbol sequence, capturing a finite history (or memory) of symbol strings. In the proposed approach, the size of the temporal memory of a symbol sequence is estimated from spectral properties of the resulting stochastic matrix corresponding to a first-order Markov model of the symbol sequence. Then, agglomerative hierarchical clustering is used to cluster states of the corresponding full-order Markov model to construct a reduced-order Markov model based on information-theoretic criteria with a non-deterministic algebraic structure; the parameters of the reduced-order model are identified from the original model by making use of a Bayesian inference rule. The model size is inferred using an information-theoretic inspired criteria; the Markov parameters of the reduced-order model are identified from the original model by making use of a Bayesian inference rule. The paper also identifies theoretical bounds on the error induced in the reduced-size model in terms of expected Hamming distance between the sequences generated by the original and final reduced-size models. The proposed concept is elucidated and validated by two examples on different data sets. The first example analyzes a set of time series of pressure oscillations in a swirl-stabilized combustor, where controlled protocols are used to induce flame instabilities. Variations in the complexity of the derived Markov model represent how the system operating condition changes from stable to an unstable combustion regime. The second example is built upon a public data set of NASA's repository for prognosis of rolling-element bearings. It is shown that: (i) even with a small state-space, the reduced-order models are able to achieve comparable performance, and (ii) the proposed approach provides flexibility in the selection of a reduced-order model for data representation and learning.

1. Motivation and introduction

Markov models are widely used as a statistical learning tool for uncertain dynamical systems [1], where, in general, a Markov chain with unobserved states is constructed from the associated temporal data; in this setting, the learning task is to infer the states and the corresponding parameters of the Markov chain. In addition to hidden Markov modeling (HMM), several other techniques have been proposed for Markov modeling of time-series data. For example, in symbolic time series analysis (STSA)-based Markov modeling [2,3], the states of a Markov chain are represented as a collection of words (i.e., symbol blocks, also referred to as memory words) of different lengths, which can be identified from the time-series data on a discrete space with finite cardinality [2–5]. The symbols are created from the continuously varying time-series by projection onto a set with finite cardinality. The learning involved for these models is inference of the hyperparameters of discretization and memory. A common ground among all these modeling

☆ This work is partially supported by the U.S. Air Force Office of Scientific Research under Grant No. FA9550-15-1-0400 dealing with dynamic data-driven application systems (DDDAS). Jan Reimann was partially supported by NSF Grant DMS-1201263. Any opinions, findings and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

* Corresponding author.
E-mail addresses: devesh.dkj@gmail.com (D.K. Jha), nurali.virani88@gmail.com (N. Virani), jan.reimann@psu.edu (J. Reimann), srivastav@ge.com (A. Srivastav), axr2@psu.edu (A. Ray).

https://doi.org/10.1016/j.sigpro.2018.03.004
0165-1684/

tools is that a Markov chain is induced by probabilistic representation of a deterministic finite state automaton (DFSA), called probabilistic finite state automaton (PFSA) [6]. While the PFSA-based inference provides a consistent and deterministic graph for learning the algebraic structure of the model, it is generally not a very compact representation and may lead to a large number of states in the resulting Markov model. To circumvent this problem, attempts have been made to reduce the state-space by merging statistically similar states of the model [3]. However, the problem may still persist because these models are constructed by partitioning the phase space of dynamical systems and the merging states that are statistically similar may lead to algebraic inconsistency. On the other hand, if the states are merged to preserve the algebraic consistency, it may lead to statistical impurity in the resulting models (i.e., states that have different statistics could be merged together). Other approaches for state aggregation in Markov chains could be found in [7–9]; however, these tools do not consider inference of the Markov model from the data, which may not be suitable for analysis of dynamic data-driven application systems (DDDAS) [10].

The state space for Markov models, created by symbolic analysis, may grow exponentially with increase in memory or order of the symbolic sequence. Estimating the right memory is critical for temporal modeling of patterns observed in the sequential data. However, some of the states may be statistically similar and thus merging them could reduce the size of state-space. Several researchers (e.g., [11,12]) have reported reduced-order Markov modeling of time-series data from temporal patterns, where the size of temporal memory of the symbolic data is estimated from the spectral properties of a PFSA and the constraint of deterministic algebraic structure is imposed by the end objective due to this choice of the data representation model.

The current paper proposes to merge the states by removing the constraint of deterministic algebraic properties associated with PFSA, where the states of the Markov chain are now collection of words from its alphabet of length estimated in the last step; this state aggregation induces a non-determinism in the finite state model. The parameters of the reduced-order Markov model are estimated by a Bayesian inference technique from the parameters associated with the higher-order Markov model. The reduced-order model for data representation is selected by using information-theoretic criteria, where a unique stopping point terminates the state-merging procedure. A bound on the distortion of the predictive capability of the models is identified for order reduction of the state-space. The final product is a generative model for the data; however, some of the predictive capabilities of a DFSA could be lost.

Contributions. Reduced-order Markov modeling of time series, presented in this paper, is constructed in a PFSA setting with a nondeterministic algebraic structure. Nondeterminism is induced by merging states of a PFSA with deterministic algebraic structure inferred from discrete sequential data, which in turn allows a compact representation of temporal data. In contrast to the approach reported by Mukherjee and Ray [3], the proposed method relies more strongly on information-theoretic concepts to arrive at a consistent stopping criterion for model selection. The resulting reduced-order model has fewer parameters to estimate, which leads to faster convergence and decision-making. These reduced-order Markov models can be used with streaming data for sequential hypothesis testing for early fault detection [13,14]. In addition, a bound is quantified on degradation in the model's predictive capability due to state-space reduction, based on the Hamming distance between the sequences generated by the original model and the reduced-order model. The algorithms presented in the paper are elucidated and validated on two different examples: (i) time series of pressure oscillations, collected from a swirl-stabilized combustor apparatus [15] to monitor thermo-acoustic instabilities, and (ii) a public data set for prognosis of rolling-element bearings [16]. In addition to the results on Markov modeling, some of the results on pressure instabilities could be of independent interest in the combustion community for machine learning and active control, as discussed below.

Excellent surveys on the current understanding of the mechanisms for the combustion instability phenomena could be found in [17–21]. Active combustion instability control (ACIC) with fuel modulation has proven to be an effective method for suppression of pressure oscillations in combustors [22,23]. Based on the work available in literature, one may conclude that the performance of ACIC is primarily limited by the large delay in the feedback loop and the limited actuator bandwidth [22,23]. Early detection of combustion instability can potentially alleviate the problems due to time delays in the ACIC feedback loop and thus possibly improve the combustion performance [13,24–27]. While the results in these papers are encouraging, there is no interpretation of the expected variations in the data-driven model that represent changes in the operating regimes of the underlying process. In contrast to the work reported in the existing literature, the current paper demonstrates the changes in the complexity of the underlying time-series model for the pressure fluctuations as the system moves to instability. In summary, this paper has presented a concept of structural changes in the underlying stochastic model and pertinent parameters due to combustion instabilities.

Organization. The paper is organized in six sections including the present section. Section 2 succinctly provides the background and mathematical preliminaries on symbolic dynamics and Markov modeling. Section 3 describes the details of the technical approach for inferring a reduced-order Markov model from time series data. Section 4 presents validation of the underlying theoretical concept on the experimental data of pressure oscillations from a laboratory-scale combustion apparatus [15]. Section 5 presents the results of validation of the underlying theoretical concept on a public data set [16,28]. Finally the paper is summarized and concluded in Section 6 along with recommendations for future research.

2. Background and mathematical preliminaries

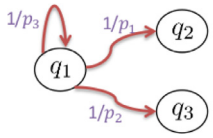
Symbolic analysis of time-series data is a relatively recent tool, where continuous sensor data are converted to symbol sequences via partitioning of the continuous domain [2,29]. The stationary dynamics of the symbols sequences are then modeled as a probabilistic finite state automaton (PFSA), which is defined as follows:

Definition 2.1 (PFSA)**.** A probabilistic finite state automaton (PFSA) is a tuple $\mathcal{M} = (\mathcal{Q}, \mathcal{A}, \delta, \boldsymbol{M})$ where

- $\mathcal{Q}$ is a finite set of states of the automaton having cardinality $|\mathcal{Q}|$;
- $\mathcal{A}$ is a finite alphabet set of symbols having cardinality $|\mathcal{A}|$;
- $\delta : \mathcal{Q} \times \mathcal{A} \to \mathcal{Q}$ is the state transition function;
- $\boldsymbol{M} : \mathcal{Q} \times \mathcal{A} \to [0, 1]$ is the $|\mathcal{Q}| \times |\mathcal{A}|$ emission matrix (also known morph matrix). The matrix $\boldsymbol{M} = [m_{ij}]$ is row stochastic such that m_{ij} is the probability of generating symbol a_j from state q_i.

Remark 2.1. An alternative representation of a PFSA is $\mathcal{M} = (\mathcal{Q}, \boldsymbol{\Pi})$ where $\boldsymbol{\Pi} : \mathcal{Q} \times \mathcal{Q} \to [0, 1]$ is called the $|\mathcal{Q}| \times |\mathcal{Q}|$ state-transition probability matrix. The matrix $\boldsymbol{\Pi} = [\pi_{ij}]$ is row stochastic and π_{ij} is the probability $\Pr(q_j|q_i)$ of visiting state q_j from state q_i. The stationary state probability $\boldsymbol{p}$ of an irreducible (also called ergodic) PFSA is the sum-normalized eigenvector of $\boldsymbol{\Pi}$ corresponding to its (unique) unity eigenvalue [30].

Remark 2.2. The PFSA defined above has a deterministic algebraic structure which is governed by the transition function δ; thus



$$q_1 = \{11, 22, 33\} \text{ (for example)}$$
$$p_1 + p_2 + p_3 = 1$$

Fig. 1. Graphical model showing non-determinism in a PFSA. The symbol 1 emitted from state q_1 leads to different states with fixed probabilities indicating non-deterministic behavior.

the emission of a given symbol from a particular state leads to a fixed state (including the original state, which forms a self-loop). However, the symbol emissions are probabilistic (represented by the emission matrix). On the other hand, the transition function for a non-deterministic finite state automaton is given by a map, $\delta : \mathcal{Q} \times \mathcal{A} \to 2^{\mathcal{Q}}$ where, $2^{\mathcal{Q}}$ denotes the power set of $\mathcal{Q}$ and includes all subsets of $\mathcal{Q}$. The idea is elucidated in Fig. 1, where the same symbol may lead to multiple states, albeit in a probabilistic fashion. This configuration allows flexibility in modeling (possibly) at the expense of some predictive accuracy.

For symbolic analysis of time-series data, a class of PFSA called the *D*-Markov machine [2,3] has been proposed as a sub-optimal but computationally efficient model to encode the dynamics of symbol sequences as a finite state machine.

Definition 2.2 D -Markov machine**.** A *D*-Markov machine is a statistically stationary stochastic process $S = \ldots a_{-1}a_0a_1 \ldots$ (modeled by a PFSA in which each state is represented by a finite history of at most D symbols), where the probability of occurrence of a new symbol depends only on at most the last D symbols, i.e.,

$$\Pr(s_n \mid \ldots s_{n-D} \ldots s_{n-1}) = \Pr(s_n \mid s_{n-D} \ldots s_{n-1})$$

where D is called the depth (or memory) of the Markov machine.

Remark 2.3. A *D*-Markov machine is thus a D^{th}-order Markov approximation of the discrete symbolic process. For many stable and controlled engineering systems that tend to forget their initial conditions, a finite length memory assumption is reasonable. The *D*-Markov machine is represented as a PFSA and states of this PFSA are words over alphabet $\mathcal{A}$ of length D (or less); the state transitions are described by a sliding block code of memory D and anticipation length of one [31]. The notion of depth D is approximately applicable to systems with fading memory, where it is expected that the predictive influence of a symbol progressively diminishes. An accurate estimation of depth for the symbolic dynamical process is required for the precise modeling of the underlying dynamics of the discrete sequence.

Remark 2.4. Given a symbol sequence and a structure of *D*-Markov machine, the emission matrix M and the transition matrix $\mathbf{\Pi}$ are estimated using frequency counting with uniform prior. In the frequency counting method, the number of occurrences of symbol s_j at state q_i is computed to obtain the count n_{ij}. This count is used to estimate the emission matrix parameters as follows:

$$m_{ij} = m(q_i, s_j) \triangleq \frac{1 + n_{ij}}{|\Sigma| + \sum_{\ell=1}^{|\Sigma|} n_{i\ell}}.$$

Here the initial value of each parameter is chosen to be $1/|\Sigma|$; this implies that, in the absence of any symbol emission at the end of an observation period, the probability of symbol emission from each state is uniformly $1/|\Sigma|$. This initialization ensures that each entry of the morph matrix is strictly positive; the details of this estimation procedure are given in [3,32].

Next an information-theoretic metric is introduced, which will be used for merging the states of the Markov model in Section 3.

Definition 2.3 (Kullback–Leibler Divergence [32])**.** The Kullback–Leibler (K–L) divergence of a discrete probability distribution P from another discrete probability distribution $\tilde{P}$ is defined as follows.

$$D_{\mathrm{KL}}(P\|\tilde{P}) = \sum_{x \in X} p_X(x) \log\left(\frac{p_X(x)}{\tilde{p}_X(x)}\right)$$

It is noted that K-L divergence is not a proper distance as it is not symmetric. However, to treat K-L divergence as a distance it is generally converted into symmetric divergence as follows, $d(P, \tilde{P}) = D_{\mathrm{KL}}(P\|\tilde{P}) + D_{\mathrm{KL}}(\tilde{P}\|P)$. This is defined as the K-L distance between the distributions P and $\tilde{P}$. This distance will be used to find out the structure in the set of the states of the PFSA-based Markov model whose states are words, over the alphabet of the PFSA, of length equal to the depth estimated for the discretized sequence. Later sections present some results, which are used to bound the distortion in model accuracy using K-L divergence.

3. Technical approach

This section presents the details of the proposed approach for inferring a Markov model from time series data. As discussed earlier, the first step is the discretization of time-series data to generate a discrete symbol sequence. While it is possible to optimize the symbolization of time-series by using some optimization criterion, such a technique is not presented here. In this paper, the data are discretized using the unbiased principle of entropy maximization of discrete sequences [33] via maximum entropy partitioning (MEP) [34]. The proposed approach for Markov modeling then consists of the following steps:

- Estimate the approximate size of temporal memory (or order) of the symbol sequence.
- Cluster the states of the high-order Markov model.
- Estimate the parameters of the reduced-order Markov model (e.g., the state transition matrix $\mathbf{\Pi}$).
- Select the final model using information theoretic scores, described later in Section 3.3.

Memory of the discrete sequence is estimated using a recently introduced method [11,12] based on the spectral analysis of the Markov model that is induced by a PFSA with unity depth (i.e., $D = 1$). It is noted that the above steps are followed during training to construct the reduced-order model from time series and, during test, the model parameters are estimated. The key ideas behind these steps are explained next.

3.1. Estimation of reduced-order Markov model

This section presents the estimation of the reduced-order model from observed symbol sequences. This is a two-step procedure – in the first step, the memory of the model is estimated; in the second step, the states of the full-order model are clustered to identify the reduced-order model. These steps are explained in the following subsections.

3.1.1. Depth estimation for Markov models

Srivastav [11] has interpreted the significance of depth D of a symbol sequence (see Definition 2.2) as the number n of time steps after which probability of current symbol is independent of any past symbol, i.e.,

$$\Pr(s_k|s_{k-n}) = \Pr(s_k) \ \forall n > D, \tag{1}$$

where the statistical dependence is evaluated on individual past symbols as $\Pr(s_k|s_{k-n})$ instead of assessing the dependence on words of length D as $\Pr(s_k|s_{k-1}, \ldots, s_{k-D})$. It is shown that if the observed process is *forward causal*, then observing any additional

intermediate symbols $s_{k-1}, \ldots, s_{k-n+1}$ does not induce a dependence between s_k and s_{k-n} if it did not exist on the individual level [11]. Since the equality in Eq. (1) may not strictly hold, the following approximation is made for estimation of depth D as follows:

$$\|\Pr(s_k|s_{k-n}) - \Pr(s_k)\| \leq \epsilon \ \forall n > D \tag{2}$$

where $\|\bullet\|$ is a distance metric (or norm) and ϵ is a user-specified tolerance or convergence condition. It is noted that $\Pr(s_k)$ is the stationary distribution for the one-step transition matrix. Then, Eq. (2) is further approximated using the eigenvalues of the one-step Markov matrix as explained below. Interested readers are referred to [11] for a detailed analysis of the same.

Let $\mathbf{\Pi} = [\pi_{ij}^{(1)}]$ be the one-step (i.e., $D = 1$) state-transition probability matrix of the PFSA $\mathcal{M}$ constructed from a symbol sequence, i.e.,

$$\mathbf{\Pi} = \Pr(s_k|s_{k-1}). \tag{3}$$

Then, the n-step transition probability $\pi_{ij}^n = \Pr(s_k = a_j|s_{k-n} = a_i)$ is obtained as:

$$\pi_{ij}^n = \sum_{a_r \in \mathcal{A}} \pi_{ir}^{\ell} \pi_{rj}^{(n-\ell)}, \tag{4}$$

which is the probability of observing a symbol a_r at an intermediate step ℓ and followed by summing over all possible choices $a_r \in \mathcal{A}$. By expanding Eq. (4) recursively, it is seen that π_{ij}^n is the ij^{th} element of the matrix $\mathbf{\Pi}^n$. The following relationship is obtained in a matrix form:

$$\Pr(s_k|s_{k-n}) = \mathbf{\Pi}^n. \tag{5}$$

If the one-step transition matrix $\mathbf{\Pi}$ is irreducible and aperiodic, then all (except the unity) eigenvalues of $\mathbf{\Pi}$ satisfy $|\lambda_j| < 1 \ \forall j \geq 2$ [2]. If the one-step transition matrix π is diagonalizable, then it follows by using the eigen-decomposition that $\mathbf{\Pi} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1}$. Then, the n-step transition matrix can be written as:

$$\mathbf{\Pi}^n = \boldsymbol{U}\boldsymbol{\Lambda}^n\boldsymbol{U}^{-1}. \tag{6}$$

Thus, the convergence rate of $\mathbf{\Pi}^n$ can be used to obtain the stationary distribution $\mathbf{\Pi}^\infty$ for computing the size of the temporal memory defined by Eq. (2). Using the distance of the state-transition probability matrix after steps from the stationary point, depth D is obtained such that the following condition is satisfied:

$$|\text{trace}(\mathbf{\Pi}^n) - \text{trace}(\mathbf{\Pi}^\infty)| \leq \sum_{j=2}^{J} \left|\lambda_j\right|^n < \epsilon \quad \forall n > D, \tag{7}$$

where n is the number of iterations of the PFSA; J is number of (non-zero) eigenvalues of $\mathbf{\Pi}$; and ϵ is the user-specified threshold as an appropriate convergence condition, i.e., the depth D of the symbol sequence is estimated for a choice of ϵ by estimating the stochastic matrix for the one-step PFSA. For real-life applications, an approximate value of ϵ can be estimated by the decay rate of the eigenvalues of $\mathbf{\Pi}$ by adjusting n in Eq. (7). In general, a value of n is selected such that with any further increment in n, there is no noticeable change in the value of Eq. (7). The physical explanation for the such an approximation is based on the fact that typical controlled dynamical systems have stable orbits and they tend to forget their initial conditions. Thus, it is reasonable to justify the approximation as systems with fading memory, which is represented by Eq. (7).

Next, another pass of data is executed to estimate the parameters of the PFSA (i.e., $\mathbf{\Pi} = \Pr(s_k|s_{k-1}, \ldots, s_{k-D})$) whose states are words of length D over the symbols in the alphabet over $\mathcal{A}$; this step is critical for modeling accuracy.

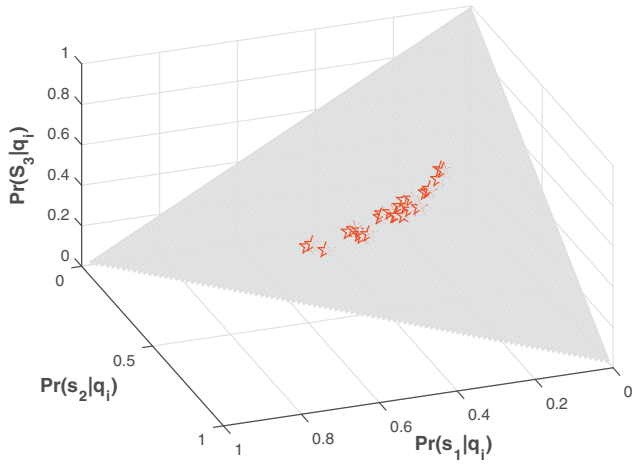


Fig. 2. The symbol emission probabilities for a Markov chain with 3 symbols shown on a simplex. Symmetric K–L distance is used to find the structure in the state-set in the information space and the states are clustered based on the revealed structure.

3.1.2. Model reduction using agglomerative hierarchical clustering

The states of the reduced-order Markov model are identified from the set of words of length D over the alphabet $\mathcal{A}$. This step is executed by agglomerative hierarchical clustering, which is a bottom-up approach [35] and generates a sparse network (e.g., a binary tree) of the state set $\mathcal{Q}$, where $|\mathcal{Q}| = |\mathcal{A}|^D$ by successive addition of edges between the elements of $\mathcal{Q}$. In general, the agglomerative hierarchical clustering algorithm is built upon a hierarchical structure and is capable of visualizing the structure of the set of the original states by using an appropriate metric. Initially, each of the states $q_1, q_2, \ldots, q_n$ of the model $\mathcal{M}$ is in its own cluster $C_i \in \mathcal{C}$, where $\mathcal{C}$ is the set of all clusters, $C_1, C_2, \ldots, C_n$, for the hierarchical cluster tree. The distance between any two states in $\mathcal{Q}$ is measured using the K-L distance between the symbol emission probabilities conditioned on them, i.e.,

$$d(q_i, q_j) = D_{\text{KL}}(\Pr(\mathcal{A}|q_i)\|\Pr(\mathcal{A}|q_j)) + D_{\text{KL}}(\Pr(\mathcal{A}|q_j)\|\Pr(\mathcal{A}|q_i)) \tag{8}$$

where the terms on the right have the following meaning.

$$D_{\text{KL}}(\Pr(\mathcal{A}|q_i)\|\Pr(\mathcal{A}|q_j)) = \sum_{s \in \mathcal{A}} \Pr(s|q_i) \log\left(\frac{\Pr(s|q_i)}{\Pr(s|q_j)}\right)$$

In terms of the distance measured by Eq. (8), the pair of clusters that are nearest to each other are merged and this step is repeated until only one cluster is left. A pseudo-code for the clustering algorithm has been provided in the Appendix as Algorithm 1, where the underlying tree structure displays the order of splits in the state set of the higher-order Markov model and is used to aggregate the states close to each other. For clarification of presentation, Fig. 2 shows an example of a Markov chain with 27 states and 3 symbols on a simplex plane, where each pentagon on the simplex plane represents one row of the symbol emission matrix. The hierarchical clustering is used to find the structure of the state set on the simplex plane using the K–L distance. The set of states clustered together could be obtained based on the number of final states required in the reduced-order Markov model.

3.2. Parameter estimation of the reduced-order Markov model

The parameters of the Markov model are identified after clustering the states of the original PFSA that has $|\mathcal{A}|^D$ states. This is accomplished by a Bayesian inference technique, where the state transition matrix $\mathbf{\Pi}$, the emission matrix $\boldsymbol{M}$, and the state proba-

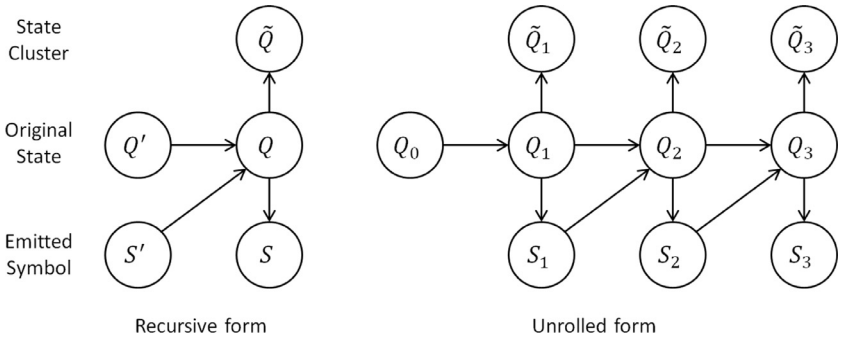


Fig. 3. Graphical models representing the dependencies between the random variables.

bility vector $\boldsymbol{p}$ of the original PFSA model $\mathcal{M}$ are available; in addition, the deterministic assignment map

$$f : \mathcal{Q} \to \widetilde{\mathcal{Q}}$$

is available, where $\mathcal{Q}$ is the state set of the original model, and $\widetilde{\mathcal{Q}}$ is the state set of the reduced-order model. Since the reduced order model is represented by the tuple $\widetilde{\mathcal{M}} = (\widetilde{\mathcal{Q}}, \widetilde{\boldsymbol{\Pi}})$, where $\widetilde{\boldsymbol{\Pi}} = [\tilde{\pi}_{ij}]$ is the state transition matrix, a Bayesian inference technique is applied to infer the individual values of transition probabilities $\tilde{\pi}_{ij} = \Pr(\tilde{q}_{k+1} = j \mid \tilde{q}_k = i)$ for all $i, j \in \widetilde{\mathcal{Q}}$.

Let the random variable Q_k be the state of the original model at the k^{th} time epoch and let S_k be the symbol emitted from that state. Then, this probabilistic emission process is governed by the emission matrix $\boldsymbol{M}$. The state of the reduced order model is obtained from a deterministic mapping of the state of the PFSA model; therefore the state of the reduced-order model is also a random variable, denoted as $\widetilde{Q}_k = f(Q_k)$. The Bayesian network representing the dependencies between these variables is shown in the recursive as well as unrolled form in Fig. 3.

The conditional density $\Pr(\widetilde{Q}_k = \tilde{q} \mid Q_k = q)$ can be evaluated by checking if state q belongs to the state cluster $\tilde{q}$ and assigning the value of 1 if true, else assign it the value of 0. Since it is known that $\widetilde{\mathcal{Q}}$ partitions the set $\mathcal{Q}$, the conditional density is well-defined. Thus, it can be written as

$$\Pr(\widetilde{Q}_k = \tilde{q} \mid Q_k = q) = \mathrm{I}_{\tilde{q}}(q), \tag{9}$$

where I is the indicator function with $\mathrm{I}_{\tilde{q}}(q) = 1$, if element q belongs to the set $\tilde{q}$, else it is 0. The derivation of the Markov model $\Pr(\widetilde{Q}_{k+1} \mid \widetilde{Q}_k)$ using $\Pr(Q_{k+1} \mid Q_k)$, stationary probability vector $\boldsymbol{p}$, and assignment map f is shown below.

$$\begin{aligned}
\Pr(\widetilde{Q}_{k+1} \mid \widetilde{Q}_k) &= \sum_{q \in \mathcal{Q}} \Pr(\widetilde{Q}_{k+1}, Q_{k+1} = q \mid \widetilde{Q}_k) && \text{(Marginalization)} \\
&= \sum_{q \in \mathcal{Q}} \Pr(Q_{k+1} = q \mid \widetilde{Q}_k) \Pr(\widetilde{Q}_{k+1} \mid \widetilde{Q}_k, Q_{k+1} = q) && \text{(Chain rule of probability)} \\
&= \sum_{q \in \mathcal{Q}} \Pr(Q_{k+1} = q \mid \widetilde{Q}_k) \Pr(\widetilde{Q}_{k+1} \mid Q_{k+1} = q) && \text{(Factorization using Figure 3)} \\
&= \sum_{q \in \mathcal{Q}} \Pr(Q_{k+1} = q \mid \widetilde{Q}_k)\, \mathrm{I}_{\widetilde{Q}_{k+1}}(q) && \text{(using Eq. (9))} \\
&= \sum_{q \in \widetilde{Q}_{k+1}} \Pr(Q_{k+1} = q \mid \widetilde{Q}_k).
\end{aligned} \tag{10}$$

Since the assignment of state of a full-order model to a cluster is a deterministic many-to-one mapping, the states $\widetilde{Q}_{k+1}$ and $\widetilde{Q}_k$ are conditionally independent given Q_{k+1}.

The density $\Pr(Q_{k+1} \mid \widetilde{Q}_k)$ is obtained by Bayes' rule as

$$\Pr(Q_{k+1} \mid \widetilde{Q}_k) = \frac{\Pr(\widetilde{Q}_k \mid Q_{k+1}) \Pr(Q_{k+1})}{\sum_{q \in \mathcal{Q}} \Pr(\widetilde{Q}_k \mid Q_{k+1} = q) \Pr(Q_{k+1} = q)}. \tag{11}$$

Similar to the steps to obtain (10), the following expression is derived as

$$\Pr(\widetilde{Q}_k \mid Q_{k+1}) = \sum_{q \in \widetilde{Q}_k} \Pr(Q_k = q \mid Q_{k+1}). \tag{12}$$

Then $\Pr(Q_k \mid Q_{k+1})$ is obtained from Bayes' rule using transition matrix and stationary probability vector as follows:

$$\Pr(Q_k \mid Q_{k+1}) = \frac{\Pr(Q_{k+1} \mid Q_k) \Pr(Q_k)}{\sum_{q \in \mathcal{Q}} \Pr(Q_{k+1} \mid Q_k = q) \Pr(Q_k = q)}. \tag{13}$$

Now the desired state transition matrix $\widetilde{\boldsymbol{\Pi}}$ of the reduced order model is obtained by combining Eqs. (10), (11), (12), and (13) together. Once the state cluster set $\widetilde{\mathcal{Q}}$ and state transition matrix $\widetilde{\boldsymbol{\Pi}}$ are available, the reduced order model is completely defined.

3.3. Model selection using information-theoretic criteria

This subsection describes the model selection process during the underlying state merging process for model inference. After computation of "penalized" likelihood estimates for different models, the model with the lowest score is selected as the optimal model. This is explained next.

The (unpenalized) log-likelihood of a symbol sequence $\vec{s}$ given a Markov model $\mathcal{M}$ is computed as follows:

$$\mathcal{L}(\vec{s} | \mathcal{M}) = \log \Pr(\vec{s} | G) \tag{14}$$

This expression is simplified by using the Markov property that the symbol emission probabilities are independent given the current state of the model $\mathcal{M}$.

$$\begin{aligned}
\mathcal{L}(\vec{s} | \mathcal{M}) &\cong \log \prod_{k=1}^{N} \Pr(s_k | q_k) \\
&\cong \sum_{k=1}^{N} \log \Pr(s_k | q_k),
\end{aligned} \tag{15}$$

where the effects of the initial state are ignored because they become insignificant for long statistically stationary symbol sequences. It is noted that, with a finite symbol sequence, the log-likelihood is always finite. Furthermore, with the Markov models considered in this paper, this sum could be further simplified using the fact that the states are completely observable and are defined by a finite collection of past symbols, i.e., $q_k = s_{k-1}, \ldots, s_{k-D}$. Under this assumption for the current class of Markov models, Eq. (15) is further simplified as follows:

$$\mathcal{L}(\vec{s}|\mathcal{M}) \cong \sum_{k=D+1}^{N} \log \Pr(s_k|s_{k-1}, \ldots, s_{k-D}) \tag{16}$$

As discussed earlier in Section 3.1, the states are merged using hierarchical clustering and thus, for every desired number of final states, the map $f_{N_{\max}}$ determines how the original states are partitioned using the hierarchical clustering. This map is known for every terminal number of states and thus, the current state in the reduced model can be estimated by using the map $f_{N_{\max}}$. Once the current state is determined, the log-likelihood is estimated by using Eqs. (15) and (16).

$$\mathcal{L}(\vec{s}|\widetilde{\mathcal{M}}) \cong \sum_{k=D+1}^{N} \log \Pr(s_k|\tilde{q}_k = f_{N_{\max}}(q_k)) \tag{17}$$

where $\tilde{q}_k$ is a state of the reduced-order model and q_k is a state of the original full-order model. Alternatively, the likelihood can also be computed with the state transition matrix of the reduced-order model (see Section 3.2) as follows:

$$\mathcal{L}(\vec{s}|\widetilde{\mathcal{M}}) \cong \sum_{k=D+1}^{N} \log \Pr(\tilde{q}_{k+1} = f_{N_{\max}}(q_{k+1})|\tilde{q}_k = f_{N_{\max}}(q_k)). \tag{18}$$

Once the log-likelihood for all possible reduced-order models are estimated, a penalty on the complexity of the models is introduced based on their size as explained below.

In the next step of the model selection process, a "complexity penalty" is added to the log-likelihood estimates, thereby balancing goodness of fit against the complexity of the model to prevent over-fitting. This paper has adopted two widely-used model selection functions, namely the Akaike information criterion (AIC) [36] and the Bayesian information criterion (BIC) [37]:

1. $\mathcal{M}_{\text{BIC}} = -2\mathcal{L}(\vec{s}|\widetilde{\mathcal{M}}) + K\log(N)$, where K is the number of free parameters and N is the number of observations.
2. $\mathcal{M}_{\text{AIC}} = -2\mathcal{L}(\vec{s}|\widetilde{\mathcal{M}}) + 2K$, where K is the number of free parameters.

The free parameters to be estimated from the data are those of the symbol emission parameters, i.e., the number of such parameters is $K = \mid \mathcal{A} \mid\mid \tilde{\mathcal{Q}} \mid$. It is noted that this procedure facilitates model selection for individual symbol sequences. The criterion here allows a terminal condition for state merging; however, different symbol sequences may have different models. The model with the minimum score is selected as the best model. Based on the results presented in next sections, it will be shown that the temporal and predictive capabilities are preserved for the reduced-order models with a very small number of states as compared to the original model.

It is noted that the calculation of log-likelihood for the reduced-order models is independent of the choice of penalty imposed on the model complexity. In this paper has used the AIC and BIC to arrive at a final model size. This is based on the results from literature that criterion result in consistent model selection results. However, since a theoretical analysis of the models estimated by AIC and BIC is not within the scope of this paper, it is considered as a topic of future research.

Remark 3.1. The final Markov model is a finite-depth approximation of the original time-series data. However, compared to the PFSA-based D-Markov machines [2,3], the current aggregated model has a non-deterministic algebraic structure, i.e., the same symbol emissions from a state can lead to different states. While this leads to some loss in predictive capability as compared to the models in [2,3], the size of the model is reduced as per the requirement at hand. This allows faster convergence rates for the symbol emission probabilities as fewer parameters are required to estimate from data, which would lead to faster decisions during testing.

The rest of this section presents a Hamming distance-based bound for distortion in the predictive capabilities of reduced models and demonstrate the utility of these models in practical problems of fault/anomaly detection from time-series data.

3.4. Analysis of the proposed algorithm

This subsection derives an upper bound on the distortion of the model due to the reduction of state-space of the Markov model using Hamming distance between two symbol sequences. Pinsker's inequality [32] is first presented, which relates the information divergence to the variational distance between probability measures defined on arbitrary spaces. This is followed by another theorem which is used to derive Hamming distance bounds using the informational divergence.

Theorem 3.1 (Pinsker's inequality)**.** *[32] Let P and Q be two probability distributions on a measurable space $(\mathbb{X}, \Sigma)$. Then, the following is true*

$$d_{TV}(P, Q) \leq \sqrt{\frac{1}{2} D_{\text{KL}}(P\|Q)} \tag{19}$$

where $d_{TV}(P, Q) = \sup_{A\in\Sigma}\{|P(A) - Q(A)|\}$ is the total variation distance.

Theorem 3.2 [38]**.** *Let $\mathbb{X}$ be a countable set and let us denote by x^n the sequence $(x_1, x_2, \ldots, x_n) \in \mathbb{X}^n$. Let q^n be a Markov measure on $\mathbb{X}^n$, that is, $q(x^n) = q(x_1)\prod_{i=2}^{n} q_i(x_i|x_{i-1})$. Then for any probability measure p^n on $\mathbb{X}^n$, the following is true*

$$\bar{d}(p^n, q^n) \leq \left[\frac{1}{2n} D_{\text{KL}}(p^n\|q^n)\right]^{1/2} \tag{20}$$

where, $\bar{d}$ denotes the normed Hamming distance on $\mathbb{X}^n \times \mathbb{X}^n$:

$$\bar{d}(x^n, y^n) = n^{-1}\sum_{n=1}^{n} d(x_i, y_i), \tag{21}$$

where $d(x_i, y_i) = 1$ if $x_i \neq y_i$ and 0 otherwise. The $\bar{d}$-distance between p^n and q^n is

$$\bar{d}(p^n, q^n) = \min E\bar{d}(\hat{X}^n, X^n), \tag{22}$$

where min *is taken over all joint distributions with marginals $p^n = \text{dist}\ \hat{X}^n$ and $q^n = \text{dist}\ X^n$ and E denotes the expectation operator.*

Theorems 3.1 and 3.2 provide a means to bound Hamming distance between sequences generated by two different distributions. This leads to a bound on the Hamming distance between the symbol sequences generated by the reduced-order Markov model and the original model, which is obtained by estimating the K–L distance between the measure on symbol sequences induced by these models. The following paragraph explains how an approximate estimate of the K–L distance between the original and a reduced model is constructed.

Let the original D-Markov model be denoted by $\mathcal{M}$ and the reduced-order model by $\widetilde{\mathcal{M}}$. The Markov measure on the probability space $(\mathbb{S}^n, \mathcal{E}, P)$, where the set $\mathbb{S}^n$ consists of sequences

of length n from an alphabet $\mathcal{A}$ could be estimated using the symbol emission probabilities. More explicitly, the Markov measure of a sequence S_n on $\mathbb{S}^n$ induced by $\mathcal{M}$ is given by $P_{\mathcal{M}}(S_n) = \Pr(q_1)\prod_{i=D+1}^{n}\Pr(s_i \mid q_i)$, where D is the depth of the model. Then, the K–L divergence between $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ is given by the following expression:

$$D_{\mathrm{KL}}(P^n_{\mathcal{M}} \| P^n_{\widetilde{\mathcal{M}}}) = \sum_{S_n \in \mathbb{S}^n} P_{\mathcal{M}}(S_n) \log\left(\frac{P_{\mathcal{M}}(S_n)}{P_{\widetilde{\mathcal{M}}}(S_n)}\right). \tag{23}$$

The above expression is simplified as follows:

$$\log\left(\frac{P_{\mathcal{M}}(S_n)}{P_{\widetilde{\mathcal{M}}}(S_n)}\right) = \log(\Pr(q_1)) - \log(\Pr(\tilde{q}_1)) + \sum_{i=D+1}^{n} \log(\Pr(s_i \mid q_i)) - \log(\Pr(s_i \mid \tilde{q}_i)),$$

where $\tilde{q}_i$ is the merged state of the reduced-order model and q_i is the original state of the full-order model. The merged state may consist of multiple original states and then $\Pr(\tilde{q}_1) = \sum_{q \in \tilde{q}_1} \Pr(q) \geq \Pr(q_1)$ as $q_1 \in \tilde{q}_1$. Using this inequality and monotonicity of the logarithm function, the first term $\log(\Pr(q_1)) - \log(\Pr(\tilde{q}_1))$ in the above equation is non-positive, and the relation (24) is obtained. The expression on the right in (24) could be further bounded by using the Lipschitz constant for the logarithm function and under the assumption that $\log(\Pr(s_j|q_i)) \neq 0 \ \forall q_i \in \mathcal{Q}$ and all $s_j \in \mathcal{A}$.

$$\log\left(\frac{P_{\mathcal{M}}(S_n)}{P_{\widetilde{\mathcal{M}}}(S_n)}\right) \leq \sum_{i=D+1}^{n} \log(\Pr(s_i \mid q_i)) - \log(\Pr(s_i \mid \tilde{q}_i)) \tag{24}$$

$$\leq \sum_{i=D+1}^{n} \left(\frac{\Pr(s_i \mid q_i) - \Pr(s_i \mid \tilde{q}_i)}{\Pr(s_i \mid q_i)}\right) \tag{25}$$

$$\leq (n - D - 1)\kappa, \tag{26}$$

where, $\kappa = \max_{q \in Q, s \in \mathcal{A}} \frac{\Pr(s|q) - \Pr(s|\tilde{q})}{\Pr(s|q)}$. In the above inequalities, Eq. (25) is obtained from equation (24) by using the observation that $\Pr(s_i \mid \tilde{q}_i) = \Pr(s_i \mid q_i) + \eta$, where η is the perturbation in the symbol emission probability from q_i when it is clustered into a new state $\tilde{q}_i$. Hence, the K-L distance in Eq. (23) could be bounded by the following term.

$$\begin{aligned} D_{\mathrm{KL}}(P^n_{\mathcal{M}} \| P^n_{\widetilde{\mathcal{M}}}) &\leq \sum_{S_n \in S^n} P_{\mathcal{M}}(S_n)(n - D - 1)\kappa \\ &= (n - D - 1)\kappa \sum_{S_n \in S^n} P_{\mathcal{M}}(S_n) \\ &= (n - D - 1)\kappa. \end{aligned} \tag{27}$$

Thus, a uniform bound on the Hamming distance between the original and the final model could then be obtained as follows.

$$\bar{d}(P_{\mathcal{M}(S_n)}, P_{\widetilde{\mathcal{M}}}(S_n)) \leq \sqrt{\frac{(n - D - 1)\kappa}{2n}} \tag{28}$$

The above inequality thus, allows comparison of models with different state-space based on the predictive accuracy of a reduced model when compared to the original model. As compared to the earlier information theoretic criteria, which were based on the efficiency of data compression by different models, the inequality in (28) allows to compare them based on their symbol emission statistics and thus, is computationally efficient. It is possible to find a rather tighter bound in an expected sense by using the stationary distribution of the two Markov chains to find an expected bound on Hamming distance. However, finding the same is left as an exercise for future work. Using the above bound for selection of models could be more efficient than the information theoretic metrics (as it can estimated by using the symbol emission probabilities instead of the penalized likelihoods); however, finding a penalized version of the bound for model selection is also left as a future exercise. In the following sections, we present modeling and analysis of two different data sets using the proposed algorithms and present relevant inference.

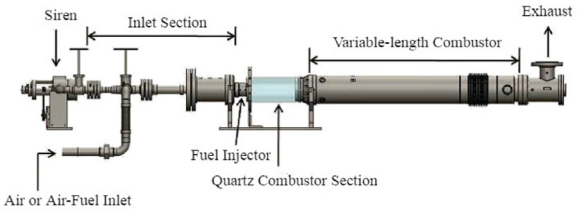


Fig. 4. Schematic diagram of the combustion apparatus.

Table 1
Operating conditions.

Parameters	Value
Equivalence ratio	0.525, 0.55, 0.60, 0.65
Inlet velocity	25–50 m/s in 5 m/s increments
Combustor length	25–59 inch in 1 inch increments

4. Concept validation on time series of pressure oscillations in a combustor

This section validates the proposed concepts on experimental data from a laboratory-scale combustion apparatus [15], where the objective is to investigate instabilities in lean-premixed combustion. A reduced order data-driven model has been used for correct labeling of transition of the combustion process from the stable to unstable phase, because a sufficiently accurate physics-based model of the combustion process is not available. The following subsections demonstrate that the proposed algorithm is capable of model order reduction to achieve trade-offs between predictive accuracy and model complexity. Results on classification and anomaly detection show that the proposed method of model learning yields good performance in terms of the machine learning objectives of class separability and anomaly detection.

4.1. Experimental data on combustion

A swirl-stabilized, lean-premixed, laboratory-scale combustor has been used for collection of experimental data for validation of the proposed algorithm. Fig. 4 depicts the schematic diagram of the variable-length combustor apparatus, consisting of an inlet section, an injector, a combustion chamber, and an exhaust section. There is an optically-accessible quartz section followed by a variable-length steel section. Further details of the combustion apparatus are available in [15].

The laboratory-scale combustor, shown in Fig. 4, was used to generate the experimental data. Tests were conducted at a nominal combustor pressure of 1 atm over a range of operating conditions, as listed in Table 1.

In each test, pressure dynamics in the combustion chamber and the global OH and CH chemiluminescence intensity were measured to study the mechanisms of combustion instability. The measurements were made simultaneously at a sampling rate of 8192 Hz (per channel), and the data were collected for 8 seconds, for a total of 65,536 measurements (per channel). A total of 780 sets of data were collected from all the tests; in every test, the combustion process was driven from stable to unstable by changing the equivalence ratio ϕ and the combustor length. It is noted that each data set includes a variety of the process behavior over a

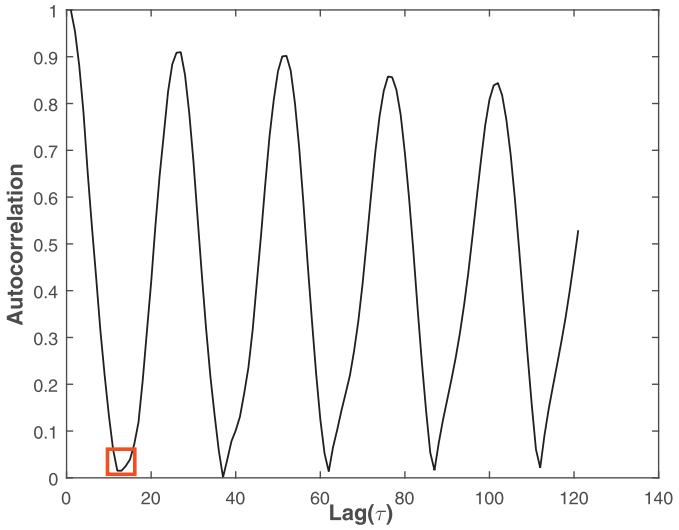


Fig. 5. Autocorrelation function of time-series data in the unstable phase of combustion. The time-series data is down-sampled by the lag marked in the red square. It is noted that the individual time-series have their own down-sampling lags. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

large number of operating conditions and thus provides rich information contents to test the efficacy of the algorithm in detecting classes, irrespective of the underlying operating conditions.

4.2. Markov modeling of combustion data

This subsection presents the results for modeling and analysis of the combustion data generated from the experimental apparatus in Fig. 4. The time-series data are first normalized by subtracting the mean and dividing by the standard deviation of its elements; this step corresponds to bias removal and variance normalization. Data from engineering systems are typically oversampled to ensure that the underlying dynamics can be captured, which is 8192 Hz in the current experiments. Due to coarse-graining from the symbolization process, an over-sampled time-series may mask the true nature of the system dynamics in the symbolic domain (e.g., occurrence of self loops and irrelevant spurious transitions in the Markov chain). The procedure is outlined below.

A time-series is first down-sampled to find the next crucial observation. The first minimum of auto-correlation function generated from the observed time-series is obtained to find the uncorrelated samples in time. The data sets are then down-sampled by this lag. Fig. 5 shows the autocorrelation function for a typical time-series in the region of unstable combustion, where the data are downsampled by the lag marked in red rectangles. To avoid the risk of discarding significant amount of data due to downsampling, the down-sampled data using different initial conditions have been concatenated. Further details of this preprocessing are reported in [11].

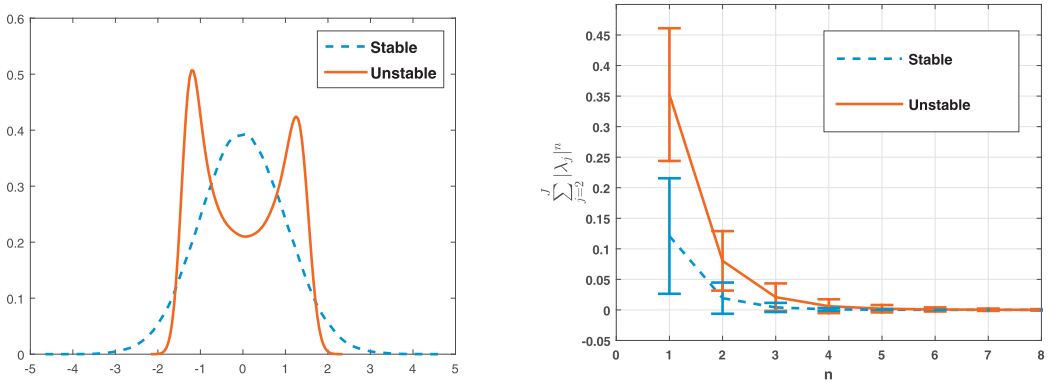
The continuous time-series data set is then partitioned using maximum entropy partitioning (MEP) [34], where the information rich regions of the data set are partitioned finer and those with sparse information are partitioned coarser. In essence, each cell in the partitioned data set contains (approximately) equal number of data points under MEP. A ternary alphabet with $\mathcal{A} = \{0, 1, 2\}$ has been used to symbolize the continuous combustion instability data. Data sets from different phases are analyzed as the process evolves from a stable region through the transient phase to the unstable region; here, the ground truth is decided using the RMS-values of pressure.

Fig. 6a shows the observed changes in the behavior of the data as the combustion operating condition changes from stable to unstable; a change in the empirical distribution of data from unimodal to bi-modal is observed as the system makes this transition. In these experiments, 150 samples of pressure data were selected for each transition from the stable to the unstable phase for analysis and comparison. First, the expected sizes of temporal memory are compared during these two phases of operation. There are changes in the eigenvalue decomposition rate for the 1-step stochastic matrix calculated from the data during the stable and unstable phase, irrespective of the combustor length and inlet velocity. During a stable operation, some of the eigenvalues very quickly go to zero as compared to the unstable operating condition, as seen in Fig. 6b, which suggests that the size of temporal memory of the discretized data increases as the system moves to the unstable operating condition. Therefore, under the stable condition, the discretized data appear to behave as symbolic noise as the predictive power of Markov models remain unaffected even if the order of the Markov model is increased. In contrast, the predictive power of the Markov model can be enhanced by increasing the order of the Markov model during an unstable operating condition, indicating a more deterministic behavior.

A threshold $\epsilon = 0.05$ is chosen to estimate the depth of the Markov models for both stable and unstable phases. The value of ϵ is estimated by the behavior of the decay of eigenvalues with n (see Eq. (7) in Section 3.1) which is shown in Fig. 6b. Correspondingly, the depth was calculated as 2 and 3 for the stable and unstable conditions, respectively (see Fig. 6). The corresponding $D(\epsilon)$ is used to construct the Markov models next. First, a PFSA (whose states are words of length $D(\epsilon)$ over $\mathcal{A}$) is created and the corresponding parameters, $\boldsymbol{M}$ and $\boldsymbol{\Pi}$, are estimated. Then, the hierarchical clustering algorithm using K-L distance is used to cluster and aggregate the states. It is noted that individual models are constructed for every sample (i.e., every sample is partitioned individually) so that the symbols have different meaning (e.g., they represent different regions in the measurement space of the signals) for every sample. Consequently, each sample may have a different state-space structure when viewed in the continuous domain. Thus, the mean behavior of the samples is not shown during any operating regime, because the state-space could be inconsistent even though the cardinality may remain the same.

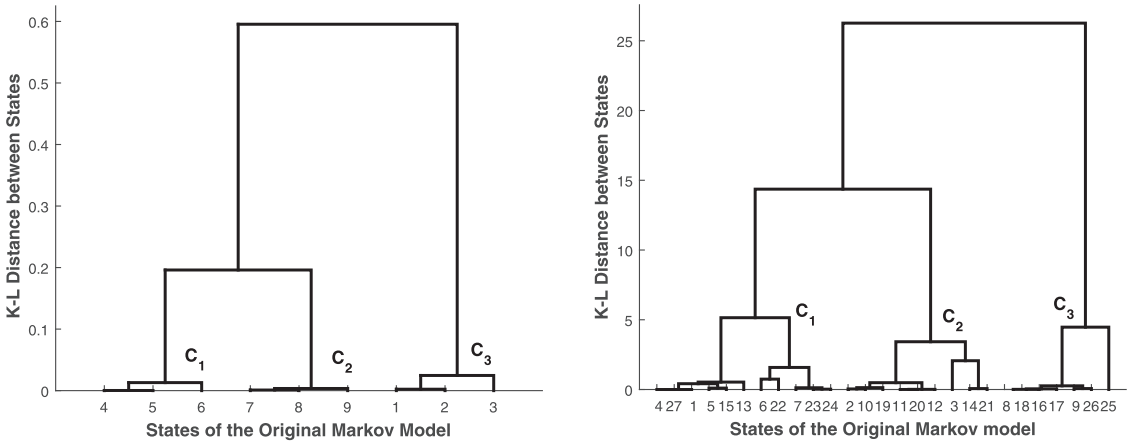
Fig. 7 shows the hierarchical cluster tree of the PFSA with depth $D(\epsilon)$ for a typical sample during stable and unstable conditions. The tree structure displays the order of splits in the states of the full-order Markov model using the metric defined in Equation (8) in Section 3.1. The figure also displays the states which are clustered together if the reduced model has 3 states (denoted as C_1, C_2, C_3 in the figure). The cluster tree also suggests the symbolic noise behavior of the data during the stable regime, where the states are very close to each other based on the K-L distance. However, clearly a coarse clustering of states in the model during the unstable behavior would lead to significant information loss, because the states are statistically different. However, to compare the two Markov models, the state cardinalities of the respective final models are kept the same. For example, the algorithm (see Appendix) is terminated with 3 states in the final Markov model during the stable as well as the unstable regime, and the final aggregated states are the three clusters depicted in Fig. 7. Once the final aggregated states are obtained, the model parameters are estimated using the Bayesian inference as discussed in Section 3.2.

Next, results for model selection are presented using the information-theoretic criteria discussed earlier in Section 3.3, where AIC and BIC are used to select the model that achieves the minimum score. These scores are estimated by first calculating the log-likelihood of all the possible reduced models using Eq. (17) in Section 3.3 and then using a penalty governed by AIC and BIC cri-



(a) Probability density function for the pressure time series data (b) Spectral decomposition of the stochastic matrix for 1-step Markov model

Fig. 6. Profiles of probability density function. The left-hand plate shows the change in the empirical density calculated for the pressure time-series data as the process deviates from a stable operating condition to an unstable operating condition. The right-hand plate shows the spectral decomposition of the 1-step stochastic matrix for the data under stable and unstable operating conditions.



(a) Hierarchical cluster tree of stable states (b) Hierarchical cluster tree of unstable states

Fig. 7. State clustering under stable and unstable conditions.

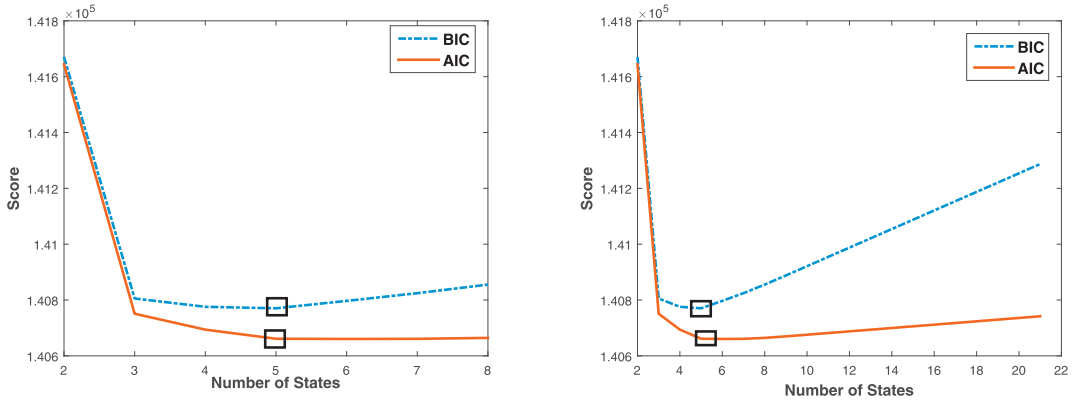
teria. As seen in Figs. 8a and b, a model structure with 5 states is selected for both stable and unstable conditions; it is noted that the original model for the stable condition had 9 states with a depth of 2 and the unstable model had 27 states with a depth of 3. In contrast to cross-validation, these criteria provide an unsupervised way for model selection, which implies that a much smaller state-space is able to preserve the temporal statistics of the data. It is interesting to note that while in general, the AIC and BIC criteria might lead to different final models, we get the same final model using both of them.

Fig. 9 shows the Hamming distance between the sequences generated by the original model and the reduced models for a typical sample, one each for stable and unstable combustion. The box-plots are generated by simulating the original model and the reduced-order model to generate symbol sequences of length 1000 from 100 different initial states (i.e., a total of 100 strings are generated) and the Hamming distance between the respective pairs is calculated. A bound on the Hamming distance between the sequences generated by the original model and final model is also calculated by using the inequality in Eq. (28), and the results are shown in Fig. 9. Although the proposed Hamming distance metric can be used to select a final model, it only measures the distance between the distributions induced by the Markov models. It is noted that the bounds on Hamming distance can provide a computationally convenient way to select model scores based on the symbol emission probabilities of the model, instead of relying on the likelihood of the symbol sequences. This issue has not been addressed here and is suggested as a topic of future research in Section 6.

4.3. Classification and anomaly detection results on combustion data

This subsection presents the results for anomaly detection and classification using the pressure time-series data to infer the underlying reduced-order Markov model. As discussed earlier in Section 4.1, since the exact transition point of the system from stable to unstable is unknown, the results are presented for anomaly detection and segregation of the data into different clusters, which can be then be associated with stable and unstable classes. Two different metrics are presented for anomaly detection for comparison of models having different state-spaces and algebraic structures. It is noted that the word metric is used here in a loose sense; it is meant to be a divergence that could be used to compare two different Markov models.

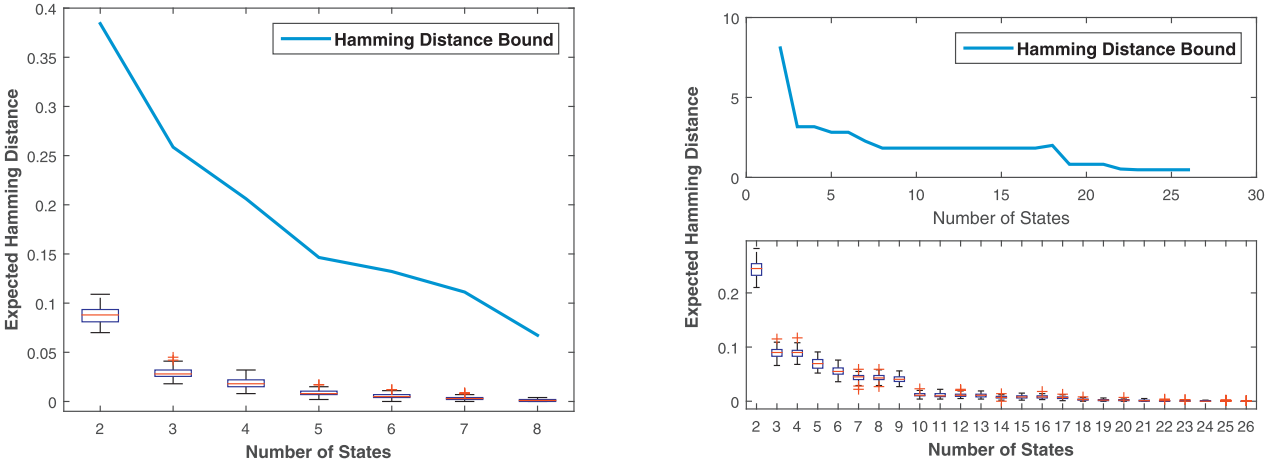
As individual time-series have different state spaces, appropriate metrics are introduced to compare them. These metrics reflect



(a) Model scores using the BIC and AIC criterion during a typical stable condition

(b) Model scores using the BIC and AIC criterion during a typical unstable condition

Fig. 8. Unsupervised model selection under stable and unstable conditions.



(a) Hamming distance between the original and final models for a typical stable combustion sample

(b) Hamming distance between the original and final models for a typical unstable combustion sample

Fig. 9. Box plot for Hamming distance between the original and reduced-order models obtained after merging.

changes in the information complexity of Markov models and reveal different behavior of the combustion process based on the changes in the inferred data model. In particular, the following two measures are introduced.

1. *Cluster divergence*: This divergence is defined for individual Markov models based on the cluster structure of the state-space model. It represents the maximum statistical difference between the states of the Markov model $\mathcal{M}$ using the K-L divergence as follows:

$$\Delta_{\mathcal{M}} = \max_{q_i, q_j \in \mathcal{Q}} d(q_i, q_j) \tag{29}$$

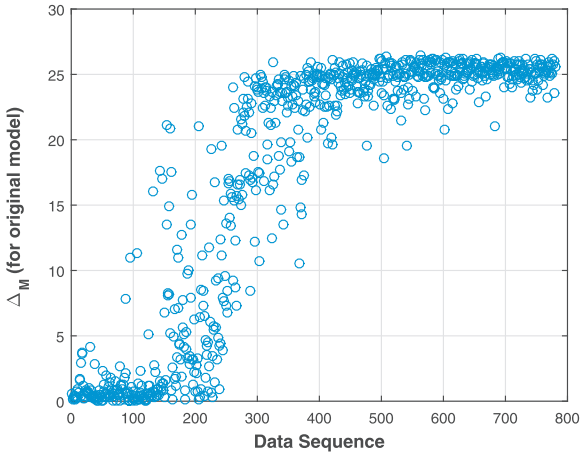
where $d(\cdot, \cdot)$ is defined in Eq. (8).

2. *Discrepancy statistics*: The discrepancy between the i.i.d. statistics and the Markov statistics is measured for the discretized data, which could also be interpreted as the information gain for Markov models; this measure also represents information complexity of data. If the i.i.d. statistics and the Markov statistics are very close, then the data has no significant temporal statistics; however, an increase in this measure would indicate the information gain by creating a temporal Markov model of the data, which is given by the following equation:

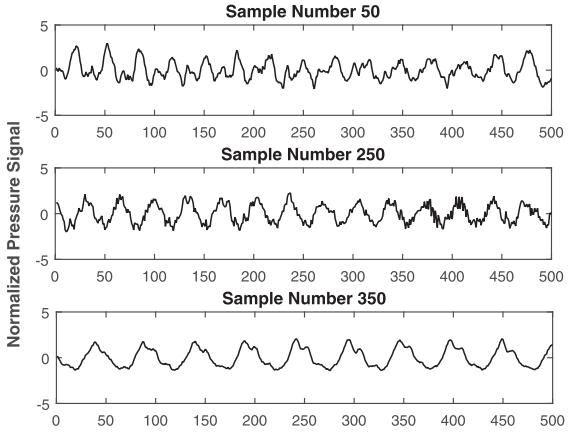
$$H_{\mathcal{M}} = \sum_{q \in \mathcal{Q}} \Pr(q) D_{KL}(\Pr(\mathcal{A} \mid q) \| \Pr(\mathcal{A})) \tag{30}$$

where $\Pr(\mathcal{A} \mid q)$ represents the symbol emission probability conditioned on a state q of the Markov model and $\Pr(\mathcal{A})$, and the term D_{KL} represents the symmetric K–L distance between two distributions.

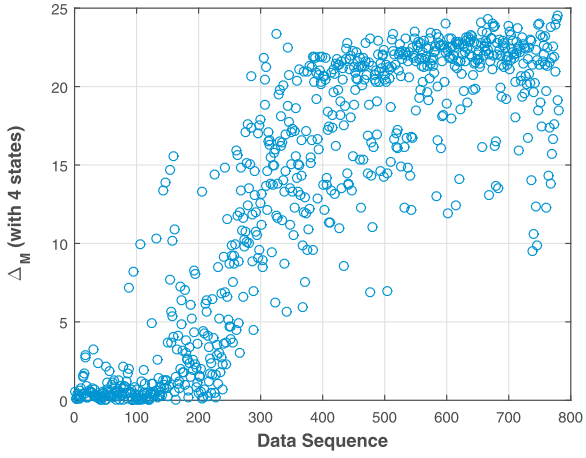
Fig. 10 presents results to show the behavior of $\Delta_{\mathcal{M}}$ with increasing pressure fluctuations. It is noted that each model has been created in an unsupervised fashion by first discretizing and then estimating the memory of the discrete data sequence. As seen in Fig. 10a, there are three distinct behaviors that can be associated with $\Delta_{\mathcal{M}}$. With low pressure fluctuations, the metric is very close to 0, indicating that the states of the model are statistically very similar. This is seen until the data number 200 with corresponding $P_{rms} \sim 0.065$ psig, which leads to a gradual change to a point where $\Delta_{\mathcal{M}}$ saturates with $P_{rms} \sim 0.12$ psig, when the process becomes unstable. Thus, different behaviors of the process can be associated with the gradual trend of increasing pressure fluctuations. However, as is seen in Fig. 10a, the transition from stable to unstable



(a) $\Delta_{\mathcal{M}}$ for the full state model for the time-series data with increasing pressure root mean square



(b) Typical pressure signals from the three clusters seen in Figure 10a



(c) $\Delta_{\mathcal{M}}$ for models with 4 states for the time-series data with increasing pressure RMS

Fig. 10. Anomalous behavior of data in the combustion process.

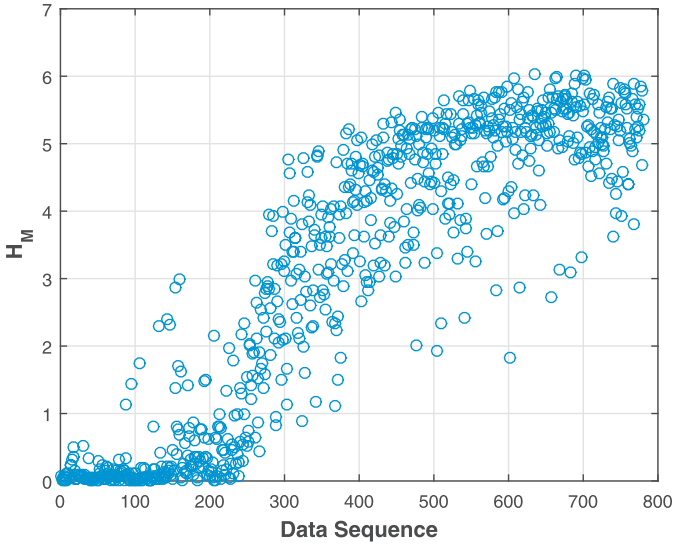


Fig. 11. Variation of discrepancy statistics $H_{\mathcal{M}}$ with increasing pressure fluctuations in the combustion process. This also shows an anomaly around the point 200 and qualitatively agrees to the behavior of $\Delta_{\mathcal{M}}$.

behavior is not clearly defined and is very difficult to label during the experiments because the process is very fast. Fig. 10b shows the pressure signals from the three different clusters, where the sample number 250 approaches an approximate limit-cycle behavior. An important point to note is that this phenomenon is independent of the operating conditions and only depends on stability (or instability) of the process; the associated metric may thus be used for anomaly detection. Fig. 10c shows the statistics of $\Delta_{\mathcal{M}}$ with four states, which is subjected to (possible) loss of information due to state merging in the unstable class, while the stable cluster remains unchanged implying that the states are statistically similar and that the model distortion due to state merging is insignificant. Thus, $\Delta_{\mathcal{M}}$ can be reliably used to detect departure from a stable behavior.

Variations of discrepancy statistics for full state models are shown in Fig. 11 that qualitatively agrees with the earlier results on $\Delta_{\mathcal{M}}$. From these plots, it is inferred that the Markov statistics for the stable cluster are very similar to the i.i.d. statistics and thus the data are very much independently distributed. Conditioning on the inferred states of the Markov models does not improve predictability (or information complexity) of the temporal model. Thus, these two measures help infer the changes in the behavior of the data during the combustion process and are useful for anomaly detection.

The underlying changes in the models are now visualized in the information space during stable and unstable phases by reducing the state space of the models to just 2 states and estimating the corresponding emission parameters. As the models have three symbols, the emission matrix has 2 rows and each row corresponds to the symbol emission probabilities conditioned on the two states. Each of these rows is plotted on a single simplex plane for 100 stable cases and 100 unstable cases. This is displayed in Fig. 12 that shows the clusters of stable and unstable cases in the information space. The figure also shows that the model with even 2 states are clustered separately and that that there is a structured change in the temporal dynamics of data at the two phases, where the inferred Markov models are able to capture this change. Furthermore, the distinctive features of the models are sufficiently retained even after significant reduction in the state-space of the models.

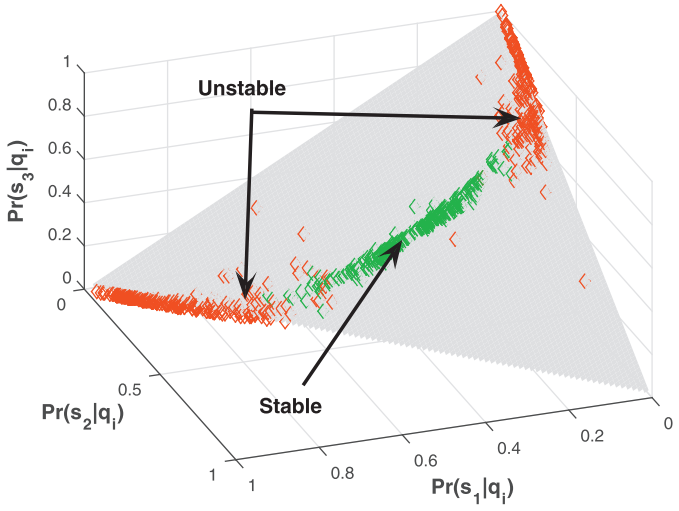


Fig. 12. Clusters of stable and unstable combustion in information space. Each point is a row of the emission matrix for the reduced Markov model with 2 states. The plot shows the change in the Markov model as the process moves from stable and unstable. Red diamonds represent the unstable phase while green diamonds represent the stable phase. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Performance of classifiers with different number of states. Mean Error= Lower is better.

Number of States	Classifier	Classification Error (%)
9	SVM	3.48 ± 0.74
	DT	9.83 ± 3.24
8	SVM	3.62 ± 0.71
	DT	9.38 ± 3.11
7	SVM	2.87 ± 0.68
	DT	7.70 ± 2.61
6	SVM	2.48 ± 0.61
	DT	7.00 ± 2.55
5	SVM	2.05 ± 0.54
	DT	6.10 ± 2.17
4	SVM	1.86 ± 0.43
	DT	4.72 ± 2.29
3	SVM	1.69 ± 0.45
	DT	5.56 ± 1.90
2	SVM	1.67 ± 0.43
	DT	4.83 ± 1.80

4.4. Classification

These models are then used to train classifiers using support vector machines (SVM) and decision trees (DT) [1]. The rationale behind using multiple classifiers is to show that the performance of the Markov models is independent of the classification technique (i.e., it works equally well with maximum-margin classifiers or decision-tree classifiers). The SVM classifier is trained using a radial basis function kernel while the decision tree is trained using the standard Euclidean distance. The classifiers are trained with 100 data points from each class and are tested on the remaining data (around 80 and 380 for stable and unstable cases, respectively). The tests are repeated for 100 different training and testing data sets from the total data. The results of classification accuracy are listed in Table 2. The SVM classifier is able to achieve around 1.67% error using models with 2 states while the decision-tree classifier is able to achieve around 4.70% error using models with 4 states. This provides another way of selecting the final model for state merging in a setting of supervised learning. It is noted that the original models contain 9 states for the stable and 27 states for the unstable class.

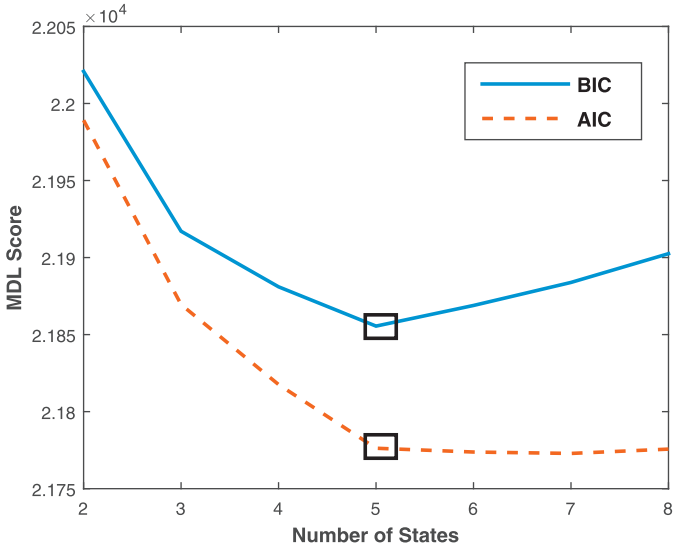


Fig. 13. Model scores using the BIC and AIC criteria for prognosis of rolling-element bearings; selected models are depicted by black rectangles.

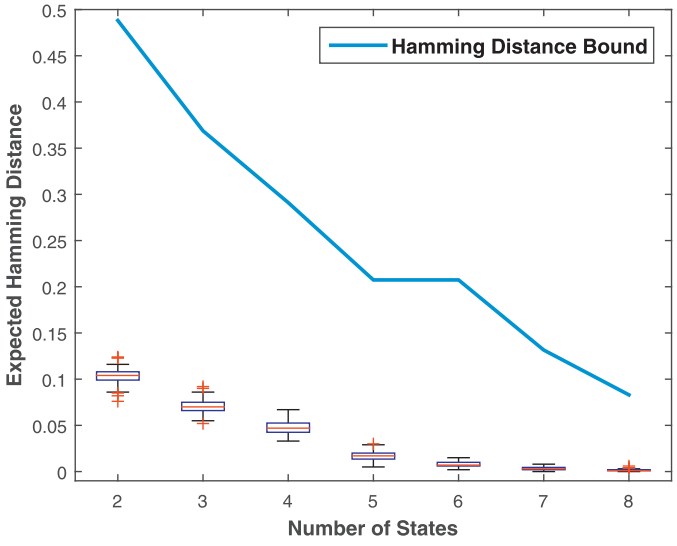


Fig. 14. Box plot of the Hamming distance between the original and reduced-order models along with the analytical bound for prognosis of rolling-element bearings.

5. Validation on a benchmark public data set

This section validates the proposed concepts on a benchmark public data set, where the objective is to predict service life of rolling element bearings in rotating machinery. It is demonstrated that the proposed algorithm is capable of model order reduction to achieve trade-offs between predictive accuracy and model complexity. Results on classification and anomaly detection demonstrate that the proposed method of model learning yields good performance in terms of the machine learning objectives of class separability and anomaly detection.

In this paper, NASA's prognostics data repository [16,39] is the source of the data set on rolling element bearings; a detailed description of the experiments is given in [28]. The bearing test rig hosts four test bearings on one shaft which is driven by an AC motor at a constant speed. A constant force is applied on each of the bearings, and the accelerometer data were collected at every bearing at a sampling rate of 20kHz for about 1 s. The tests were carried for 35 days until a significant amount of debris was found in

the magnetic plug of the test bearing. A defect in at least one of the bearings was found at the end of every test. This paper uses the data from Bearing 3, which show anomalous behavior in later parts of the test.

5.1. Test results of rolling element bearing data

This subsection presents the results on modeling of the bearing data using the same procedure as explained earlier for combustion data. The same procedure of downsampling and depth estimation is followed for analysis of bearing data as was described in Section 4.2 for combustion. A ternary alphabet is also chosen here to discretize the continuous data after downsampling and the maximum entropy partitioning is used to find the partitions. Using the spectral method, a depth of 2 (i.e., a total of $3^2 = 9$ states) is estimated for an $\epsilon = 0.02$; however, the plot of spectral decomposition plot is not icluded here for brevity.

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) scores for the different models are shown in Fig. 13 and the model with five states is selected using the obtained scores (marked in black rectangle). In Fig. 14, we show the Hamming distance between the sequences generated by the original model (with 9 states) and the reduced models and the corresponding bounds obtained by inequality (28).

6. Summary, conclusions and future work

In recent times the idea of representation learning has become very popular in the machine learning literature as it allows decoupling of data for model learning from the end-objectives like classification or clustering. This paper has presented a technique for Markov modeling of time-series data using concepts of symbolic dynamics, which allows inference of model structure as well as parameters for compact data representation. The proposed technique first estimates the memory size of the discretized time-series data. Then, the size of memory is estimated using spectral decomposition properties of the one-step Markov model created from the symbol sequence. A second pass of data is made to infer the model with the right memory and the corresponding symbol emission matrix is estimated. Then, the equivalence class of states based on K–L distance between the states are estimated using hierarchical clustering of the corresponding states of the Markov model.

The proposed concept has been validated using two different datasets– combustion instability and service life of rolling element bearing. Since modeling of instability phenomena still remains a puzzle in the combustion community, the Markov modeling technique in the symbolic domain has been used to analyze the problem of combustion instability in this paper. The proposed concepts have been tested on experimental data from a swirl-stabilized combustor [15] that was constructed in Penn State laboratories to investigate unstable thermo-acoustic phenomena in combustion processes. The proposed approach is capable of quantifying the complexity of time-series data based on the inferred Markov model. Two different metrics have been proposed for anomaly detection and classification of the stable and unstable classes. While the results presented in this paper are encouraging as the inferred models are able to identify the stable and unstable phases independent of any other operating condition, further theoretical research and experimental validation are deemed necessary before the proposed methodology can be used to develop codes for use in real-life applications. A few topics of future research are suggested below.

1. Simultaneous optimization of discretization and memory estimation for model inference.
2. Comparison of the proposed models with hidden Markov models (HMM) of similar state-space size.
3. Theoretical analysis of the models estimated by Akaike information criterion (AIC) [36] and the Bayesian information criterion (BIC) [37].
4. Analysis of transient data for prognosis and control of combustion instabilities.

Acknowledgments

The authors would like to thank Professor Domenic Santavicca and Mr. Jihang Li of Center for Propulsion, Penn State for kindly providing the experimental data for combustion used in this work.

Appendix : Pseudocode of the main algorithm

This appendix presents Algorithm 1 as a the pseudo-code, which has been used to find the model parameters in the training phase. The parameters in the testing phase are estimated using the clustering map $f_{N_{\max}}$. Algorithm 1 has been executed on two different data-sets to find the model parameters in the training phase. The parameters in the testing phase are estimated using the clustering map $f_{N_{\max}}$.

Algorithm 1: Reduced order Markov modeling.

Input: The observed symbol sequence $\vec{s} = \{\ldots s_1 s_2 s_3 \ldots | s_i \in \mathcal{A}\}$

Output: The final Markov model, $\mathcal{M} = (\tilde{\mathcal{Q}}, \tilde{\boldsymbol{M}}, \tilde{\boldsymbol{\Pi}})$

Estimate the $\boldsymbol{\Pi}$ matrix for 1-step Markov model using frequency counting with an uniform prior;
Estimate the size of temporal memory, $D(\epsilon)$ for $\vec{s}$ using equation (7);
Estimate $\boldsymbol{M}$ and $\boldsymbol{\Pi}$ for the $D(\epsilon)$-Markov model using frequency counting with an uniform prior;
$\mathcal{C}_{|\mathcal{Q}|} = \{q_i \mid q_i \in \mathcal{Q}\}$;
for $i = | \mathcal{Q} | -1, \ldots, 1$ **do**
    find distinct clusters $A, B \in \mathcal{C}_{i+1}$ minimizing $d(A \cup B)$;
    $\mathcal{C}_i := (\mathcal{C}_{i+1} \setminus \{A, B\}) \cup \{A \cup B\}$
return $\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{Q}|}$ *and* $f_i : \mathcal{Q} \to \mathcal{C}_i \; \forall i \in \{1, \ldots, | \mathcal{Q} |\}$
Calculate the parameters of reduced model using $\tilde{\mathcal{Q}} = \mathcal{C}_{N_{\max}}$, $f_{N_{\max}}$ and equations (10) through (13);
Calculate the Log-likelihood for models with Equation (17);
The final model is selected using the Akaike information criterion (AIC) or Bayesian information criterion (BIC) criteria explained in Section 3-C;

References

[1] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[2] A. Ray, Symbolic dynamic analysis of complex systems for anomaly detection, Signal Process. 84 (7) (2004) 1115–1130.
[3] K. Mukherjee, A. Ray, State splitting and merging in probabilistic finite state automata for signal representation and analysis, Signal Process. 104 (2014) 105–119.
[4] C.R. Shalizi, K.L. Shalizi, Blind construction of optimal nonlinear recursive predictors for discrete sequences, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, UAI '04, 2004, pp. 504–511.
[5] I. Chattopadhyay, H. Lipson, Abductive learning of quantized stochastic processes with probabilistic finite automata, Philosop. Trans. R. Soc. Lon. A 371 (1984) (2013) 20110543.
[6] E. Vidal, F. Thollard, C. De La Higuera, F. Casacuberta, R.C. Carrasco, Probabilistic finite-state machines-part i, IEEE Trans. Pattern Anal. Mach. Intell. 27 (7) (2005) 1013–1025.
[7] B.C. Geiger, T. Petrov, G. Kubin, H. Koeppl, Optimal kullback–leibler aggregation via information bottleneck, IEEE Trans. Automat. Contr. 60 (4) (2015) 1010–1022.
[8] M. Vidyasagar, A metric between probability distributions on finite sets of different cardinalities and applications to order reduction, IEEE Trans. Automat. Contr. 57 (10) (2012) 2464–2477.

[9] Y. Xu, S.M. Salapaka, C.L. Beck, Aggregation of graph models and markov chains by deterministic annealing, Autom. Contr. IEEE Trans. 59 (10) (2014) 2807–2812.
[10] F. Darema, Dynamic data driven applications systems: New capabilities for application simulations and measurements, in: 5th International Conference on Computational Science - ICCS 2005, 2005. Atlanta, GA; United States.
[11] A. Srivastav, Estimating the size of temporal memory for symbolic analysis of time-series data, in: American Control Conference, Portland, OR, USA, 2014, pp. 1126–1131.
[12] D.K. Jha, A. Srivastav, K. Mukherjee, A. Ray, Depth estimation in Markov models of time-series data via spectral analysis, in: American Control Conference (ACC), 2015, IEEE, 2015, pp. 5812–5817.
[13] N. Virani, D.K. Jha, A. Ray, Sequential hypothesis tests using Markov models of time series data, Workshop on Machine Learning for Prognostics and Health Management at 2016 KDD, San Francisco, CA, 2016.
[14] N. Virani, Learning Data-Driven Models for Decision-Making in Intelligent Physical Systems, The Pennsylvania State University, 2017.
[15] K. Kim, J. Lee, B. Quay, D. Santavicca, Response of partially premixed flames to acoustic velocity and equivalence ratio perturbations, Combust. Flame 157 (9) (2010) 1731–1744.
[16] NASA, Prognostic data repository: Bearing data set nsf i/ucrc center for intelligent maintenance systems, 2010.
[17] J. O'Connor, V. Acharya, T. Lieuwen, Transverse combustion instabilities: acoustic, fluid mechanic, and flame processes, Prog. Energy Combust. Sci. 49 (2015) 1–39.
[18] Sé, b. Ducruix, T. Schuller, D. Durox, Sé, b. Candel, Combustion dynamics and instabilities: elementary coupling and driving mechanisms, J. Propul. Power 19 (5) (2003) 722–734.
[19] S. Candel, D. Durox, T. Schuller, J.-F. Bourgouin, J.P. Moeck, Dynamics of swirling flames, Annu. Rev. Fluid Mech. 46 (2014) 147–173.
[20] Y. Huang, V. Yang, Dynamics and stability of lean-premixed swirl-stabilized combustion, Prog. Energy Combust. Sci. 35 (4) (2009) 293–364.
[21] J.P. Moeck, J.-F. Bourgouin, D. Durox, T. Schuller, S. Candel, Nonlinear interaction between a precessing vortex core and acoustic oscillations in a turbulent swirling flame, Combust. Flame 159 (8) (2012) 2650–2668.
[22] A. Banaszuk, P.G. Mehta, C.A. Jacobson, A.I. Khibnik, Limits of achievable performance of controlled combustion processes, IEEE Trans. Control Syst. Technol. 14 (5) (2006) 881–895.
[23] A. Banaszuk, P.G. Mehta, G. Hagen, The role of control in design: from fixing problems to the design of dynamics, Control Eng. Pract. 15 (10) (2007) 1292–1305.
[24] D.K. Jha, A. Srivastav, A. Ray, Temporal learning in video data using deep learning and Gaussian processes, Workshop on Machine Learning for Prognostics and Health Managament at 2016 KDD, San Francisco, CA, 2016.
[25] S. Sarkar, S.R. Chakravarthy, V. Ramanan, A. Ray, Dynamic data-driven prediction of instability in a swirl-stabilized combustor, Int. J. Spray Combust. Dyn. (2016). 1756827716642091.
[26] V. Nair, G. Thampi, R. Sujith, Intermittency route to thermoacoustic instability in turbulent combustors, J. Fluid Mech. 756 (2014) 470–487.
[27] M. Murugesan, R. Sujith, Combustion noise is scale-free: transition from scale-free to order at the onset of thermoacoustic instability, J. Fluid Mech. 772 (2015) 225–245.
[28] H. Qiu, J. Lee, J. Lin, G. Yu, Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics, J. Sound Vib. 289 (4) (2006) 1066–1090.
[29] J. Lin, E. Keogh, L. Wei, S. Lonardi, Experiencing SAX: a novel symbolic representation of time series, Data Min. Knowl. Discov. 15 (2) (2007) 107–144.
[30] A. Berman, P.R. J, Nonnegative Matrices in the Mathematical Sciences, John Wiley & Sons, 2012.
[31] D. Lind, B. Marcus, An introduction to symbolic dynamics and coding, Cambridge University Press, 1995.
[32] R.M. Gray, Entropy and information, Springer, 1990.
[33] T.M. Cover, J.A. Thomas, Elements of information theory, John Wiley & Sons, 2012.
[34] V. Rajagopalan, A. Ray, Symbolic time series analysis via wavelet-based partitioning, Signal Process. 86 (11) (2006) 3309–3320.
[35] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (3) (2005) 645–678.
[36] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Contr. 19 (6) (1974) 716–723.
[37] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (2) (1978) 461–464.
[38] K. Marton, Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration, Ann. Probab. 24 (2) (1996) 857–866.
[39] D.A. Tobon-Mejia, K. Medjaher, N. Zerhouni, G. Tripot, A data-driven failure prognostics method based on mixture of gaussians hidden markov models, IEEE Trans. Reliab. 61 (2) (2012) 491–503.