Article

Bayesian Nonparametric Modeling of Categorical Data for Information Fusion and Causal Inference [†]

Sihan Xiong [1,‡], Yiwei Fu [1] and Asok Ray [1,2,*]

[1] Department of Mechanical Engineering, Pennsylvania State University, University Park, PA 16802-1412, USA; xiongsihan@gmail.com (S.X.); yxf118@psu.edu (Y.F.)
[2] Department of Mathematics, Pennsylvania State University, University Park, PA 16802-1412, USA
[*] Correspondence: axr2@psu.edu

[†] This work has been supported in part by the U.S. Air Force Office of Scientific Research (AFOSR) under Grant No. FA9550-15-1-0400 in the area of dynamic data-driven application systems (DDDAS). Any opinions, findings and recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

[‡] Current address: Siemens Corporation Technology, Princeton, NJ 08540, USA.

Received: 25 February 2018; Accepted: 17 May 2018; Published: 23 May 2018

Abstract: This paper presents a nonparametric regression model of categorical time series in the setting of conditional tensor factorization and Bayes network. The underlying algorithms are developed to provide a flexible and parsimonious representation for fusion of correlated information from heterogeneous sources, which can be used to improve the performance of prediction tasks and infer the causal relationship between key variables. The proposed method is first illustrated by numerical simulation and then validated with two real-world datasets: (1) experimental data, collected from a swirl-stabilized lean-premixed laboratory-scale combustor, for detection of thermoacoustic instabilities and (2) publicly available economics data for causal inference-making.

Keywords: Bayesian nonparametric; information fusion; causal inference; conditional tensor factorization; Bayes factor; sequential classification; thermoacoustic instability

1. Introduction

Modeling and decision-making in complex dynamical systems (e.g., distributed physical processes [1], macro-economy [2] and human brain [3]) often rely on time series collected from heterogeneous sources. Fusion of the information extracted from an ensemble of time series is a critical ingredient for better prediction and causal inference.

In many dynamical systems, the characteristic time of the physical process under consideration is small (e.g., around 2 ms in a typical combustion process) relative to the time-scale of respective decision-making (e.g., tenths of a second for active combustion control). Therefore, fast and accurate prediction of the system states and estimation of the associated parameters is essential for online monitoring and active control of the dynamical system; for example, real-time prediction of future states can significantly improve active control of thermoacoustic instabilities [4]. One way to achieve this is to make predictions based on different but correlated information sources. Although several methods have been proposed for prediction based on fusion of heterogeneous time series (e.g., [5–7]), they lack a coherent probabilistic interpretation and may not be able to accommodate more general interactions between current measurements and the measurement history. Furthermore, these methods may not be sequentially implementable and hence they may not be very useful for real-time applications.

Identification of causal relationships is essential for understanding the consequences of transitions from empirical findings to actions and thus forms a significant part of knowledge discovery. Various analytical techniques (e.g., [8–10]) have been proposed for causal inference-making; among these techniques, the concept of causality introduced by Granger [11], hereafter called Granger causality, is apparently one of the most widely used in time series analysis [12]. Granger causality does not rely on the specification of a scientific model and thus is particularly applicable to investigation of empirical cause-effect relationships. It is noted that Granger causality is especially suited for continuous-valued data based on frequentist hypothesis testing.

The goal of this paper is to develop a flexible and parsimonious model of categorical time series in a Bayesian nonparametric setting for fusion of correlated information from heterogeneous sources (e.g., sensors of possibly different modalities), which can be used for sequential classification and causal inference. From this perspective, major contributions of the paper are delineated as follows:

1. By introducing latent variables and sparsity inducing priors, a flexible and parsimonious model is developed for fusion of correlated information from heterogeneous sources (e.g., sensors of possibly different modalities), which can be used to improve the performance of sequential classification tasks.
2. By testing the dimension of latent variables in the setting of Bayes factor analysis [13], Granger causality [11] is extended to categorical time series.
3. Validation of the above concept with experimental data, generated from a swirl-stabilized lean-premixed laboratory-scale combustor [14], for real-time detection of thermoacoustic instabilities.
4. Testing of the underlying algorithm with public economics data to infer the causal relationship between two categorical time series.

The paper is organized into eight sections including the current section. Section 2 introduces the concept of Granger causality and develops the model. Section 3 discusses the algorithm for posterior computation using Gibbs sampling, and hypothesis testing using Bayes factor analysis. Section 4 presents the sequential classification algorithm with the proposed model. The underlying algorithms are tested with simulation data in Section 5 while Section 6 validates the proposed method with some experimental data, collected from a swirl-stabilized lean-premixed laboratory-scale combustor, for thermoacoustic instabilities early detection. Section 7 validates the proposed concept on publicly available economics data. Section 8 concludes the paper and provides a few recommendations for future research. The nomenclature and list of acronyms are provided at the end before the list of references.

2. Model Development

This section first introduces the concept of Granger causality and the corresponding regression model. Next, the underlying model's algebraic and statistical specifications are elaborated.

Definition 1. *(Granger Causality) Let $\{y_t\}_{t=1}^T$ and $\{\theta_t\}_{t=1}^T$ be two (statistically) stationary categorical time series. Then, the variable θ Granger-causes the variable y if the past values of θ contain statistically significant information for predictions of y besides those contained in the past values of y. Similarly, y Granger-causes θ if the past values of y contain statistically significant information for predictions of θ besides those contained in the past values of θ.*

Remark 1. *The following are four types of Granger causality relationship between θ and y:*

1. *θ Granger-causes y but not the vice versa;*
2. *y Granger-causes θ but not the vice versa;*
3. *θ and y Granger-cause each other;*
4. *θ does not Granger-cause y and vice versa.*

However, in practice, only finitely many past values of y and θ are considered. To test the null hypothesis that θ does not Granger-cause y, the following regression model is constructed:

$$p(y_t \mid y_{t-1}, \dots, y_{t-D_y}, \theta_{t-1}, \dots, \theta_{t-D_\theta}) \tag{1}$$

where in this model, predictors y_{t-1} to y_{t-D_y} represent variable y's time lags; and predictors θ_{t-1} to θ_{t-D_θ} represent variable θ's time lags. In the sequel, for simplicity of notations, predictors $\mathbf{z}_t \equiv (z_{1,t}, \dots, z_{q,t})$ are substituted for $(y_{t-1}, \dots, y_{t-D_y}, \theta_{t-1}, \dots, \theta_{t-D_\theta})$.

Remark 2. *If the explanatory power of $\theta_{t-1}, \dots, \theta_{t-D_\theta}$ to the regression is significant, then the null hypothesis (that θ does not Granger-cause y) is rejected and the alternative hypothesis (that θ Granger-causes y) is accepted. Hypothesis tests on the significance of time-lags are elaborated later in Equation (15) (see Section 3.2).*

Remark 3. *If y and θ are correlated in the sense of Granger causality, the information contained in one source can be used to predict the future values in another source. Accordingly, information fusion of different sources enables fast and accurate prediction because of Granger-causality. It is noted that if the information contained in two sources is statistically independent, then information fusion cannot enhance prediction accuracy.*

2.1. Conditional Tensor Factorization

This subsection addresses fusion of different sources of information by making use of the concept of conditional probability tensor that was first reported in [15], a formal definition of conditional probability tensor follows.

Definition 2. *(Conditional probability tensor) Let C_0 denote the number of categories of the (one-dimensional) variable, y_t, and let C_j denote the number of categories of $z_{j,t}$ for $j = 1, \dots, q$, where is the number of predictors. The quantity $p(y_t \mid \mathbf{z}_t)$ is treated as a $(q+1)$th order tensor in the $C_0 \times C_1 \cdots \times C_q$ dimensional space, hereafter called the conditional probability tensor.*

Let C_y and C_θ respectively denote the numbers of categories of the variables, y and θ. It follows from Definition 2 that $C_1 = \cdots = C_{D_y} = C_y$ and $C_{D+1} = \cdots = C_q = C_\theta$. Then, each one of these conditional probability tensors has a higher order singular value decomposition (HOSVD) of the following form [15]:

$$p(y_t \mid \mathbf{z}_t) = \sum_{s_1=1}^{k_1} \cdots \sum_{s_q=1}^{k_q} \lambda_{s_1,\dots,s_q}(y_t) \prod_{j=1}^{q} \omega_{s_j}^{(j)}(z_{j,t}) \tag{2}$$

where $1 \le k_j \le C_j$ for $j = 1, \dots, q$; and each of the parameters $\lambda_{s_1 \dots s_q}(y_t)$ and $\omega_{s_j}^{(j)}(z_{j,t})$ is non-negative while the following constraints are satisfied:

$$\sum_{y_t=1}^{C_0} \lambda_{s_1,\dots,s_q}(y_t) = 1, \quad \text{for each } (s_1, \dots, s_q) \tag{3}$$

$$\sum_{s_j=1}^{k_j} \omega_{s_j}^{(j)}(z_{j,t}) = 1, \quad \text{for each } (j, z_{j,t}) \tag{4}$$

Remark 4. *Since there exists a factorization as in Equation (2) for each one of the conditional probability tensors, the two constraints Equations (3) and (4) are not restrictive. Furthermore, it is ensured that $\sum_{y_t=1}^{C_0} p(y_t \mid \mathbf{z}_t) = 1$.*

2.2. Bayesian Nonparametric Modeling

In order to build a statistically interpretable model, two techniques can be used to convert the tensor factorization in Equation (2) to a Bayes network, i.e., (1) introduce latent allocation-class variables; (2) assign sparsity-inducing priors. To this end, T pairs of variables and their respective predictors are collected in one dataset, and it is rearranged as $\{y_t, \boldsymbol{z}_t\}_{t=1}^{T}$, where t is an index with range from 1 to T.

The conditional probability $p(y_t \mid \boldsymbol{z}_t)$, factorized as in Equation (2), is then reorganized in the following form:

$$p(y_t \mid \boldsymbol{z}_t) = \int_{x_{1,t}} \cdots \int_{x_{q,t}} p(y_t \mid \boldsymbol{x}_t) \prod_{j=1}^{q} p(x_{j,t} \mid z_{j,t}) \tag{5}$$

where $\boldsymbol{x}_t \equiv (x_{1,t}, \ldots, x_{q,t})$ denotes the latent class-allocation variables.

For index $j = 1, \ldots, q$ and index $t = 1, \ldots, T$, it then follows that

$$x_{j,t} \mid \boldsymbol{\omega}^{(j)}, z_{j,t} \sim \mathbf{Mult}(\boldsymbol{\omega}^{(j)}(z_{j,t})) \tag{6}$$

$$y_t \mid \tilde{\boldsymbol{\lambda}}, \boldsymbol{x}_t \sim \mathbf{Mult}(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{x}_t}) \tag{7}$$

where $\mathbf{Mult}(\bullet)$ is the multinomial distribution [16] and $\boldsymbol{\omega}^{(j)} \equiv \{\{\omega_s^{(j)}(c)\}_{s=1}^{k_j}\}_{c=1}^{C_j}$ is the mixture probability matrix. The cth row $\boldsymbol{\omega}^{(j)}(c) \equiv \{\omega_s^{(j)}(c)\}_{s=1}^{k_j}$ in this mixture probability matrix is a probability vector itself (i.e., it sums to 1). Moreover, $\tilde{\boldsymbol{\lambda}} \equiv \{\boldsymbol{\lambda}_{s_1,\ldots,s_q}\}_{(s_1 \ldots s_q)}$ is a conditional probability tensor where $\boldsymbol{\lambda}_{s_1,\ldots,s_q} \equiv \{\lambda_{s_1,\ldots,s_q}(c)\}_{c=1}^{C_0}$ is a probability vector for each string $(s_1, \ldots, s_q)$.

The hierarchical reformulation of HOSVD above illustrates the following features of this model in Equation (5):

- Soft clustering for each one of the predictors $\boldsymbol{z}_j \equiv \{z_{j,t}\}_{t=1}^{T}$ is implemented following Equation (6). This allows for inheritance of statistical strengths across different categories.
- The distribution of variable y_t is determined by a probability tensor $\tilde{\boldsymbol{\lambda}}$ of reduced order, following Equation (7).
- In order to capture the interactions among different predictors, class assignment variables $\boldsymbol{x}_j \equiv \{x_{j,t}\}_{t=1}^{T}$ are used. They work in an implicit and parsimonious way by allowing the latent populations with the index of $(s_1, \ldots, s_q)$ to be shared across various state combinations of predictors.

Remark 5. *Here it is very critical to distinguish these two different concepts: (1) the number of clusters $\tilde{k}_j$ generated by the latent class variables $\boldsymbol{x}_j$ and (2) the dimensions k_j of the probability vector $\boldsymbol{\omega}^{(j)}(c)$ from the mixture probability matrix. The former one represents the number of groups generated by the data, and is smaller than the latter. It should be noted that $\tilde{k}_j$ determines if the predictor $\boldsymbol{z}_j$ should be included in the model, because $p(y_t \mid \boldsymbol{z}_t)$ will not change with $z_{j,t}$ if $\boldsymbol{z}_j$ has just a single latent cluster. Thus the significance of some particular predictor could be tested using on $\tilde{k}_j$, which is later elaborated in Section 3.2.*

In many real-world applications, the tensor $\tilde{\lambda}$ often has more components than needed, since the product $\prod_{j=1}^{q} k_j$ can be large even for modest values of q and C_j. To deal with this problem, tensor $\tilde{\lambda}$ is then clustered within different combinations of $(s_1, \ldots, s_q)$ nonparametrically by imposing a Pitman-Yor process prior [17]. Then, by using the stick-breaking representation of the Pitman-Yor process [18], it follows that

$$\boldsymbol{\lambda}_l \mid \gamma \sim \mathbf{Dir}(\alpha), \quad \text{for } l = 1, \ldots, \infty \tag{8}$$

$$V_k \mid a, b \sim \mathbf{Beta}(1 - b, a + kb), \quad \text{for } k = 1, \ldots, \infty \tag{9}$$

$$\pi_l = V_l \prod_{k=1}^{l-1} (1 - V_k), \quad \text{for } l = 1, \ldots, \infty \tag{10}$$

where the bold symbols $\mathbf{Dir}(\bullet)$ and $\mathbf{Beta}(\bullet)$ represents the uniform Dirichlet distributions and Beta distributions [16] respectively, and $\boldsymbol{\lambda}_l \equiv (\lambda_l(1), \ldots, \lambda_l(C_0))$. Moreover, $0 \leq b < 1$ and $a > -b$. For each combination $(s_1, \ldots, s_q)$, it follows that

$$\phi_{s_1,\ldots,s_q} \mid \boldsymbol{\pi} \sim \mathbf{Mult}(\pi) \tag{11}$$

where $\boldsymbol{\pi} \equiv (\pi_1, \pi_2, \ldots)$. For $t = 1, \ldots, T$,

$$y_t \mid \boldsymbol{\lambda}\,, \boldsymbol{\phi}, \boldsymbol{x}_t \sim \mathbf{Mult}(\lambda_{\phi_{x_t}}) \tag{12}$$

where $\boldsymbol{\lambda} \equiv \{\lambda_l\}_{l=1}^{\infty}$ and $\boldsymbol{\phi} \equiv \{\phi_{s_1,\ldots,s_q}\}_{(s_1,\ldots,s_q)}$.

The next step assigns priors to the mixture probability matrix $\boldsymbol{\omega}^{(j)}$. Here the dimension of $\boldsymbol{\omega}^{(j)}$ grows linearly as k_j increases (unlike the tensor $\tilde{\lambda}$). Therefore, further clustering of $\boldsymbol{\omega}^{(j)}$ is not necessary. Hence, we assign independent priors to the rows of $\boldsymbol{\omega}^{(j)}$ for $j = 1, \ldots, q$ in the following way:

$$\boldsymbol{\omega}^{(j)}(c) \mid k_j, \beta_j \sim \mathbf{Dir}(\beta_j), \quad \text{for } c = 1, \ldots, C_j \tag{13}$$

Lastly, we assign priors to the dimension of the mixture probability vector k_j, i.e., for $j = 1, \ldots, q$,

$$p(k_j = k \mid \mu_j) \propto \exp(-\mu_j k), \quad \text{for } k = 1, \ldots, C_j \tag{14}$$

where $\mu_j \geq 0$ and $\boldsymbol{k} \equiv \{k_j\}_{j=1}^{q}$.

Remark 6. *As the parameter μ_j grows larger, the exponential prior in Equation (14) will assign increasing probabilities to smaller values of k_j, and it becomes a uniform prior distribution on $\{1, \ldots, C_j\}$ when μ_j is zero. Commonly, people have prior beliefs that as time lags increase, they will a have vanishing impact on the distribution of the current response variable. To impose this prior belief, we can assign larger μ_j to time lags further back in the history.*

By combining Equations (6)–(14) together, a Bayes network representation of the model is created and Figure 1 illustrates its structures.

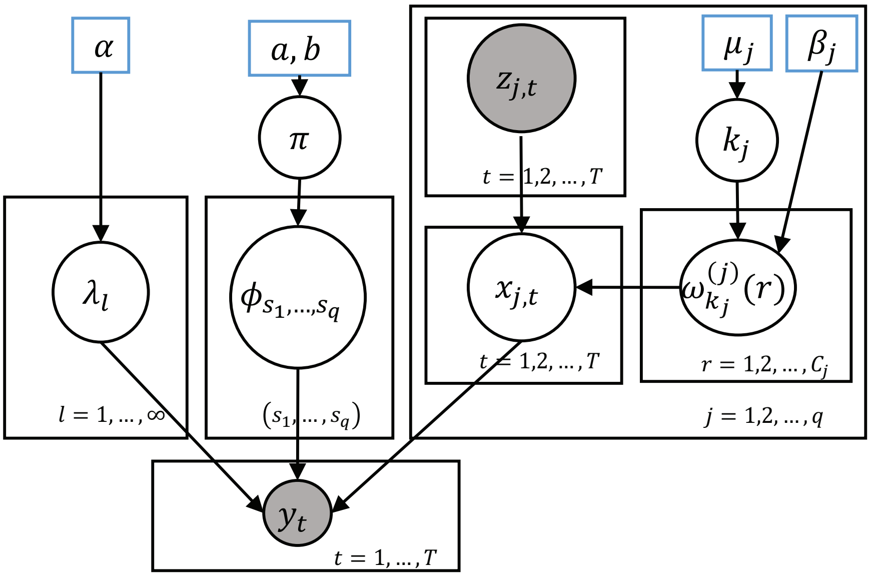

Figure 1. Bayes network representation of the model in the form of a graph. Deterministic hyperparameters are those that are enclosed by blue rectangles. Unobserved random variables are enclosed by transparent (unshaded) circles, and observed random variables are enclosed by shaded circles.

3. Estimation and Inference

This section presents the details of an algorithm for computing posteriors as well as Bayesian hypothesis testing by using Bayes factors.

3.1. Posterior Computation

Despite the fact that the posterior distribution does not have any specific analytical form, we can still perform the inference of the corresponding Bayes network by using Gibbs sampling method. Because the dimension of $\boldsymbol{\omega}^{(j)}$ may vary with k_j, constructing a stationary Markov chain by plain Gibbs sampling is difficult. To infer a model with variable dimensions, a common analytical tool—the reversible jump Monte Carlo Markov chain (MCMC) [19], which does trans-dimensional exploration in the model space—is often used.

Product partition modeling [20,21] can help alleviate difficulties occurring in trans-dimensional modeling by constructing a stationary Markov chain on the clustering space. For this proposed method, the dimension $\boldsymbol{\omega}^{(j)}$ is being integrated out for the sampling of k_j directly from $p(k_j \mid \boldsymbol{x}_j, \boldsymbol{z}_j)$, which will create a partially collapsed Gibbs sampler [22] that alternates between these two spaces: (1) the space with all the variables and (2) the space with all the variables but $\boldsymbol{\omega} = \{\boldsymbol{\omega}^{(j)}\}_{j=1}^{q}$.

To compute the posterior probabilities of the Pitman-Yor process, the infinite-dimensional tensors $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ after their Lth component are truncated, as performed in [18]. For achieving desired accuracy, an appropriate L needs to be chosen. Other than this, the posterior sampling is rather straightforward. The detailed process is presented in Algorithm 1, in which it is not explicitly mentioned that $\boldsymbol{x} \equiv \{\boldsymbol{x}_t\}_{t=1}^{T}$ and $\boldsymbol{\xi}$ collects the variables.

Algorithm 1 Gibbs sampling for the proposed method

```
Input: Datasets $\{y_t, z_t\}_{t=1}^{T}$; hyperparameters a, b, $\alpha$, $\{\mu_j\}_{j=1}^{q}$, $\{\beta_j\}_{j=1}^{q}$; number of truncating components $L$; number of all
    samples $N$; initial sample ${}_{(0)}\boldsymbol{\phi}$, ${}_{(0)}\boldsymbol{\pi}$, ${}_{(0)}\boldsymbol{\lambda}$, ${}_{(0)}\boldsymbol{\omega}$, ${}_{(0)}\boldsymbol{x}$, ${}_{(0)}\boldsymbol{k}$.
Output: All posterior samples $\{{}_{(n)}\boldsymbol{\phi},\ {}_{(n)}\boldsymbol{\pi},\ {}_{(n)}\boldsymbol{\lambda},\ {}_{(n)}\boldsymbol{\omega},\ {}_{(n)}\boldsymbol{x},\ {}_{(n)}\boldsymbol{k}\}_{n=1}^{N}$

for $n = 1$ to $N$ do
    For each one string $(s_1, \ldots, s_q)$, collect a sample $\phi_{s_1,\ldots,s_q}$ from its multinomial full conditional
        $p(\phi_{s_1,\ldots,s_q} = l \mid \boldsymbol{\xi}) \propto \pi_l \prod_{c=1}^{C_0} \{\lambda_l(c)\}^{n_{s_1,\ldots,s_q}(c)}$
    where $n_{s_1,\ldots,s_q}(c) = \sum_1^T \mathbf{1}\{x_{1,t} = s_1, \ldots, x_{q,t} = s_q, y_t = c\}$.
    For $l = 1, \ldots, L$, update $\pi_l$ by the following rules
        $V_l \mid \boldsymbol{\xi} \sim \mathbf{Beta}(1 - b + n_l, a + lb + \sum_{k>l} n_k), \quad l < L$
        $V_L = 1, \quad \pi_l = V_l \prod_{k=1}^{l-1} (1 - V_k)$
    where $n_l = \sum_{(s_1,\ldots,s_q)} \mathbf{1}\{\phi_{s_1,\ldots,s_q} = l\}$.
    For $l = 1, \ldots, L$, collect samples $\boldsymbol{\lambda}_l$ from their respective Dirichlet full conditionals
        $\boldsymbol{\lambda}_l \mid \boldsymbol{\xi} \sim \mathbf{Dir}\{\alpha + n_l(1), \ldots, \alpha + n_l(C_0)\}$
    where $n_l(c) = \sum_{(s_1,\ldots,s_q)} \mathbf{1}\{\phi_{s_1,\ldots,s_q} = l\} n_{s_1,\ldots,s_q}(c)$.
    For $j = 1, \ldots, q$, for $c = 1, \ldots, C_j$, collect samples
        $\boldsymbol{\omega}^{(j)}(c) \mid \boldsymbol{\xi} \sim \mathbf{Dir}\{\beta_j + n_{j,c}(1), \ldots, \beta_j + n_{j,c}(k_j)\}$
    where $n_{j,c}(s_j) = \sum_{t=1}^T \mathbf{1}\{x_{j,t} = s_j, z_{j,t} = c\}$.
    For $j = 1, \ldots, q$, for $t = 1, \ldots, T$, collect samples $x_{j,t}$ from their corresponding multinomial full conditionals
        $p(x_{j,t} = s \mid \boldsymbol{\xi}, x_{i,t} = s_i, i \neq j) \propto \omega_s^{(j)}(z_{j,t}) \lambda_{\phi_{s_1,\ldots,s,\ldots,s_q}}(y_t)$
    For $j = 1, \ldots, q$, collect samples $k_j$ from their respective multinomial full conditionals
        $p(k_j = k \mid \boldsymbol{\xi}) \propto \exp(-\mu_j k) \prod_{c=1}^{C_j} n_{j,c}^{-k\beta_j},\ k_j = \max_t \{x_{j,t}\}, \ldots, C_j$
    where $n_{j,c} = \sum_{t=1}^T \mathbf{1}\{z_{j,t} = c\}$.
end for
```

To successfully run Algorithm 1, certain hyperparameters need to be chosen. The aforementioned determination of μ_j and L have been carefully discussed along with their implications, so we focus on the other hyperparameters. Among those hyperparameters, a and b will determine the clustering ability of the Pitman-Yor process (which are set to be 1 and 0 in this case), rendering it a Dirichlet process; this is sufficient for applications discussed in this paper. It should be noted that α and β_j are Dirichlet Distribution's hyperparamters and serve the role of pseudo-counts. The determination of these reflects the users' prior belief. They are often manually chosen to be some small values without additional information which can justify larger values. In the following sections, they are chosen to be: $\alpha = 1$ and $\beta_j = 1/C_j$ across different applications.

3.2. Bayesian Factor and Hypothesis Testing

This subsection discusses hypothesis testing techniques on the significance of all the predictors to the regression Equation (1). It can be used to make causal inference in order to provide a better understanding of the model and to better allocate computational resources for the sequential classification task by including only the important predictors (and discard the unimportant ones). As previously noted, a particular predictor z_j is considered important if and only if the number of clusters $\tilde{k}_j$ formed by their corresponding latent class allocation variables x_j is greater than 1.

Let $\Lambda \subset \{1, \ldots, q\}$ be the set of predictors under consideration. To perform the Bayesian hypothesis testing, we only need to compute the Bayes factor [23] in favor of $H_1 : \tilde{k}_j > 1$ for some $j \in \Lambda$ against $H_0 : \tilde{k}_j = 1$ for any $j \in \Lambda$, given by

$$BF_{10} = \frac{p(H_1|\boldsymbol{y}, \boldsymbol{z})/p(H_1)}{p(H_0|\boldsymbol{y}, \boldsymbol{z})/p(H_0)} \tag{15}$$

where $\boldsymbol{y} \equiv \{y_t\}_{t=1}^{T}$, $\boldsymbol{z} \equiv \{\boldsymbol{z}_t\}_{t=1}^{T}$; and $p(H_0|\boldsymbol{y}, \boldsymbol{z})$, $p(H_1|\boldsymbol{y}, \boldsymbol{z})$ are numerically computed as the fraction of samples in which the $\tilde{k}_j$'s conform to H_0 and H_1, respectively; the prior probabilities $p(H_0)$ and $p(H_1)$ can be obtained by the following probability equation:

$$\begin{aligned} p(\tilde{k}_j = 1) &= \sum_{k=1}^{C_j} p(k_j = k) \sum_{l=1}^{k} p(x_{j,t} = l \;\forall\; t | k_j = k) \\ &= \Big(\prod_{r=1}^{C_j} \gamma_j^{(n_{j,c})}\Big)\Big(\sum_{k=1}^{C_j} \frac{p(k_j = k)k}{\prod_{k=1}^{C_j} (k\gamma_j)^{(n_{j,c})}}\Big) \end{aligned}$$

Specifically, to test whether θ Granger-causes y, it is only necessary to choose

$$\Lambda = \{D_1 + 1, \ldots, q\}.$$

4. Sequential Classification

In Section 3, a Gibbs sampling algorithm is developed to infer the posterior distribution of model parameters given the observed data. In this section, a classification algorithm for dynamical systems based on the posterior predictive distribution, which is derived by marginalizing the likelihood of unobserved data over the posterior distribution of model parameters, is proposed. This algorithm consists of two phases: (1) off-line training phase and (2) online testing phase. Suppose there are M different classes of dynamical systems that are of interest, $\mathcal{C}_i, i = 1, 2, \cdots, M$, for each of them we collect a training set ${}^{(i)}\mathcal{D}_{T_i} = \{{}^{(i)}y_t, {}^{(i)}z_t\}_{t=1}^{T_i}$. The requirement for this dataset is that the data are categorical (e.g., quantized categories from continuous data), and for each class they have an identical number of categories of predictors and variables.

During the training phase, training set ${}^{(i)}\mathcal{D}_{T_i}$ is used to compute the posterior of samples

$$\{{}^{(i)}_{(n)}\boldsymbol{\phi}, {}^{(i)}_{(n)}\boldsymbol{\lambda}, {}^{(i)}_{(n)}\boldsymbol{\omega}\}_{n=1}^{M}$$

for each one of the class $\mathcal{C}_i$, as previously described in Algorithm 1. Then, during the test phase, the test set $\mathcal{D}_T$ will be classified. Among these M classes, one will be identified as the class to which $\mathcal{D}_T$ most likely belongs. In order to do so, the following conditional probability $p(\mathcal{D}_T \mid {}^{(i)}\mathcal{D}_{T_i})$ will be computed:

$$p(\mathcal{D}_T \mid {}^{(i)}\mathcal{D}_{T_i}) = \prod_{t=1}^{T} p(y_t \mid z_t; {}^{(i)}\mathcal{D}_{T_i}) \tag{16}$$

$$p(y_t \mid z_t; {}^{(i)}\mathcal{D}_{T_i}) \approx \frac{1}{N} \sum_{n=1}^{N} \Big(\sum_{s_1=1}^{k_1} \cdots \sum_{s_q=1}^{k_q} {}^{(i)}_{(n)}\lambda_{{}^{(i)}_{(n)}\phi s_1,\ldots,s_q}(y_t) \prod_{j=1}^{q} {}^{(i)}_{(n)}\omega^{(j)}_{s_j}(z_{j,t}) \Big) \tag{17}$$

Following the above calculation of conditional probabilities $p(\mathcal{D}_T \mid {}^{(i)}\mathcal{D}_{T_i})$, the posterior probability of the test data $\mathcal{D}_T$ belonging to class $\mathcal{C}_i$ (denoted as $p(\mathcal{C}_i \mid \mathcal{D}_T)$) can be then calculated as:

$$p(\mathcal{C}_i \mid \mathcal{D}_T) = \frac{p(\mathcal{D}_T \mid {}^{(i)}\mathcal{D}_{T_i}) p(\mathcal{C}_i)}{\sum\limits_{r=1}^{M} p(\mathcal{D}_T \mid {}^{(r)}\mathcal{D}_{T_r}) p(\mathcal{C}_r)} \tag{18}$$

where $p(\mathcal{C}_i)$ is the prior probability of the class $\mathcal{C}_i$. Next, the classification result is generated by:

$$D_{\text{class}} = \arg\max_i p(\mathcal{C}_i \mid \mathcal{D}_T) \tag{19}$$

The prior probability $p(\mathcal{C}_i)$ reflects user's subjective beliefs and can also be designed to optimize some objective criterion. The reason that the detection algorithm is "sequential" is due to the fact the conditional probability $p(\mathcal{D}_T \mid {}^{(i)}\mathcal{D}_{T_i})$ is evaluated one by one as shown in Equation (16). In real-world applications, values of $p(y_t \mid z_t; {}^{(i)}\mathcal{D}_{T_i})$ in Equation (17) are often precomputed and stored for various values of (y_t, z_t), in order to achieve faster computations.

For the binary classification case, we can construct the likelihood ratio test [24] as:

$$\frac{p(\mathcal{D}_T \mid {}^{(1)}\mathcal{D}_{T_1})}{p(\mathcal{D}_T \mid {}^{(0)}\mathcal{D}_{T_0})} \underset{0}{\overset{1}{\gtrless}} \Theta \tag{20}$$

where in this equation Θ is a certain threshold. To choose the threshold Θ, one could rely on the receiver operating characteristic (ROC). ROC curves are often obtained by changing Θ in order to make a trade-off between the probability of (successful) detection $p_D = Prob(\text{decide } 1 \mid 1 \text{ is true})$ and the false alarm probability $p_F = Prob(\text{decide } 1 \mid 0 \text{ is true})$. Using those ROC curves, an optimal combination of p_D and test set data length for a given p_F can be selected, which would then determine the threshold Θ.

5. Numerical Example

This section presents a numerical example which utilizes the proposed method to infer causal relationships between two categorical time series. In this example, the data generation model is known and thus can be compared with the results from the proposed algorithm for evaluation of performance. The data generation details are given below.

In this particular numerical example, there are two binary sequences of symbols y_t and θ_t. Symbol sequences y_t are generated using a known Markov model $p(y_t \mid y_{t-1}, y_{t-3}, y_{t-4})$, where only the time-lags $y_{t-1}, y_{t-2}, y_{t-5}$ are important predictors. Symbol sequences θ_t are generated from another Markov model $p(\theta_t \mid \theta_{t-1}, \theta_{t-2}, y_{t-1}, y_{t-3})$, where $\theta_{t-1}, \theta_{t-2}$ and y_{t-1}, y_{t-3} are the key predictors. In other words, the variable y Granger-causes the variable θ but not the other way around because y only depends on its own past. Table 1 lists the transition probabilities for y_t, where it is seen that the predictors are $y_{t-1}, y_{t-3}, y_{t-4}$ only. Table 2 lists the transition probabilities for θ_t, where the predictors are $y_{t-1}, y_{t-3}, \theta_{t-1}$, and θ_{t-2} only.

Table 1. Transition Probabilities for y_t in the Numerical Example.

y_{t-1}	y_{t-3}	y_{t-4}	$p(y_t = 1)$	$p(y_t = 0)$
0	0	0	0.20	0.80
1	0	0	0.75	0.25
0	1	0	0.70	0.30
1	1	0	0.35	0.65
0	0	1	0.40	0.60
1	0	1	0.38	0.62
0	1	1	0.33	0.67
1	1	1	0.71	0.29

Table 2. Transition Probabilities for θ_t in the Numerical Example.

y_{t-1}	y_{t-3}	θ_{t-1}	θ_{t-2}	$p(\theta_t = 1)$	$p(\theta_t = 0)$
0	0	0	0	0.40	0.60
1	0	0	0	0.65	0.35
0	1	0	0	0.70	0.30
1	1	0	0	0.40	0.60
0	0	1	0	0.50	0.50
1	0	1	0	0.47	0.53
0	1	1	0	0.33	0.67
1	1	1	0	0.69	0.31
0	0	0	1	0.45	0.55
1	0	0	1	0.75	0.25
0	1	0	1	0.30	0.70
1	1	0	1	0.50	0.50
0	0	1	1	0.75	0.25
1	0	1	1	0.66	0.34
0	1	1	1	0.65	0.35
1	1	1	1	0.20	0.80

To estimate the regression model in Equation (1) with the parameter $T = 1005$, samples of $\{y_t\}_{t=1}^{1005}$ and $\{\theta_t\}_{t=1}^{1005}$ are being collected simultaneously. Based on the prior belief that y_{t-D} and θ_{t-D} are no longer important for making predictions about y_t and θ_t when D is greater than 5, predictors for both y_t and θ_t are set as follows:

$$z_t \equiv (y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}, \theta_{t-1}, \theta_{t-2}, \theta_{t-3}, \theta_{t-4}, \theta_{t-5}) \tag{21}$$

From these data sets, 1000 training samples are chosen for testing the proposed algorithm.

To calculate posteriors using Algorithm 1 for $p(y_t|z_t)$, since there is no other prior knowledge, μ_j is set to be 1 across $j = 1, \ldots, 10$. Initially, 200,000 samples are used in a burn-in period: they are fed into the algorithm and then discarded. The next 50,000 samples (after burn-in) are downsampled further by taking every 5th sample to reduce their autocorrelation. Figure 2 summarizes the results, in which Figure 2a displays the log-likelihood for 10,000 iterations of this model and Figure 2b illustrates the ability to correctly identify all the important predictors for the proposed method. For this example, the key predictors should be 1, 3 and 4, and the results from the prediction (y_{t-1}, y_{t-3} and y_{t-4}) are the same as the ground truth. Figure 2c shows the relative frequency of number of predictors that are important. Furthermore, the proposed method also creates parsimonious representations of the model as seen in Figure 2d,e. As previously discussed in Section 2.2, the tensor $\lambda_{s_1 \ldots s_q}(y_t)$ has more components than needed but it can be clustered in a nonparametric way to reduce the number of combinations. Referring to [13], Figure 2f shows the Bayes factors calculation as mentioned in Section 3.2 for all of the predictors. Bayes factor BF_{10} in Equation (15) can be regarded as the evidence against H_0. After setting a commonly-used threshold of $t = 20$, it can be concluded that those predictors with higher BF_{10} have implications of their evidences being strong. Furthermore, having $BF_{10} > 150$ indicates even stronger evidence against the hypothesis H_0 [13]. It should be noted here that when the inclusion proportions of different lags in Figure 2b are equal to 1, then their corresponding Bayes factors in Figure 2f should tend to infinity (as for predictors 1, 3 and 4 in this example).

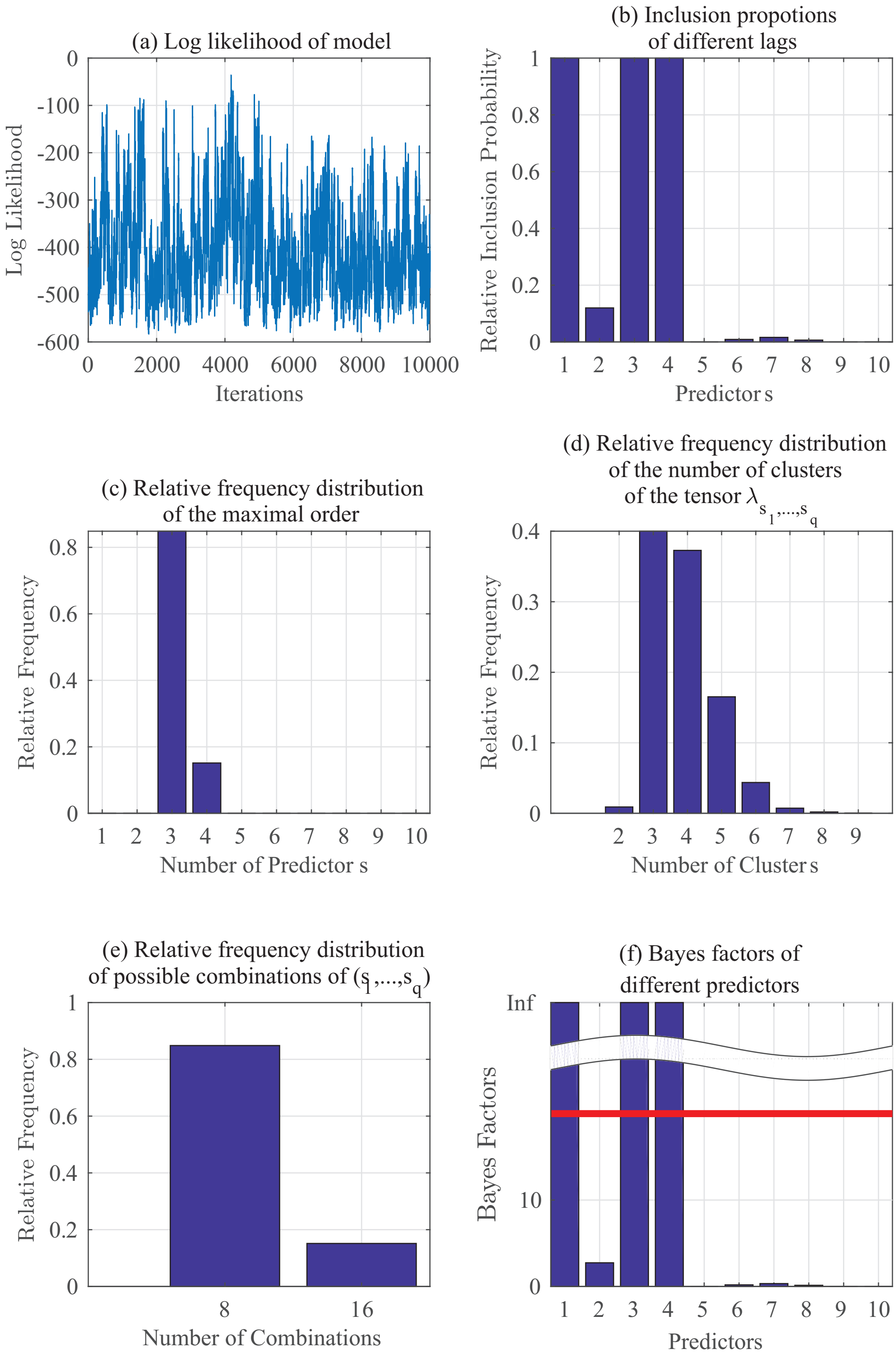

Figure 2. Gibbs sampling results: Numerical example for $p(y_t \mid y_{t-1}, \ldots, y_{t-5}, \theta_{t-1}, \ldots, \theta_{t-5})$.

Similarly, Figure 3 shows the results using the same set of data as in Figure 2 but instead of estimating $p(y_t|z_t)$, we are estimating $p(\theta_t|z_t)$ here. Figure 3a–f have the same implications as those previously stated for Figure 2a–f. It can be seen that in this case for $p(\theta_t)$, the key predictors should be 1, 3, 6 and 7, and the results confirm this in Figure 3.

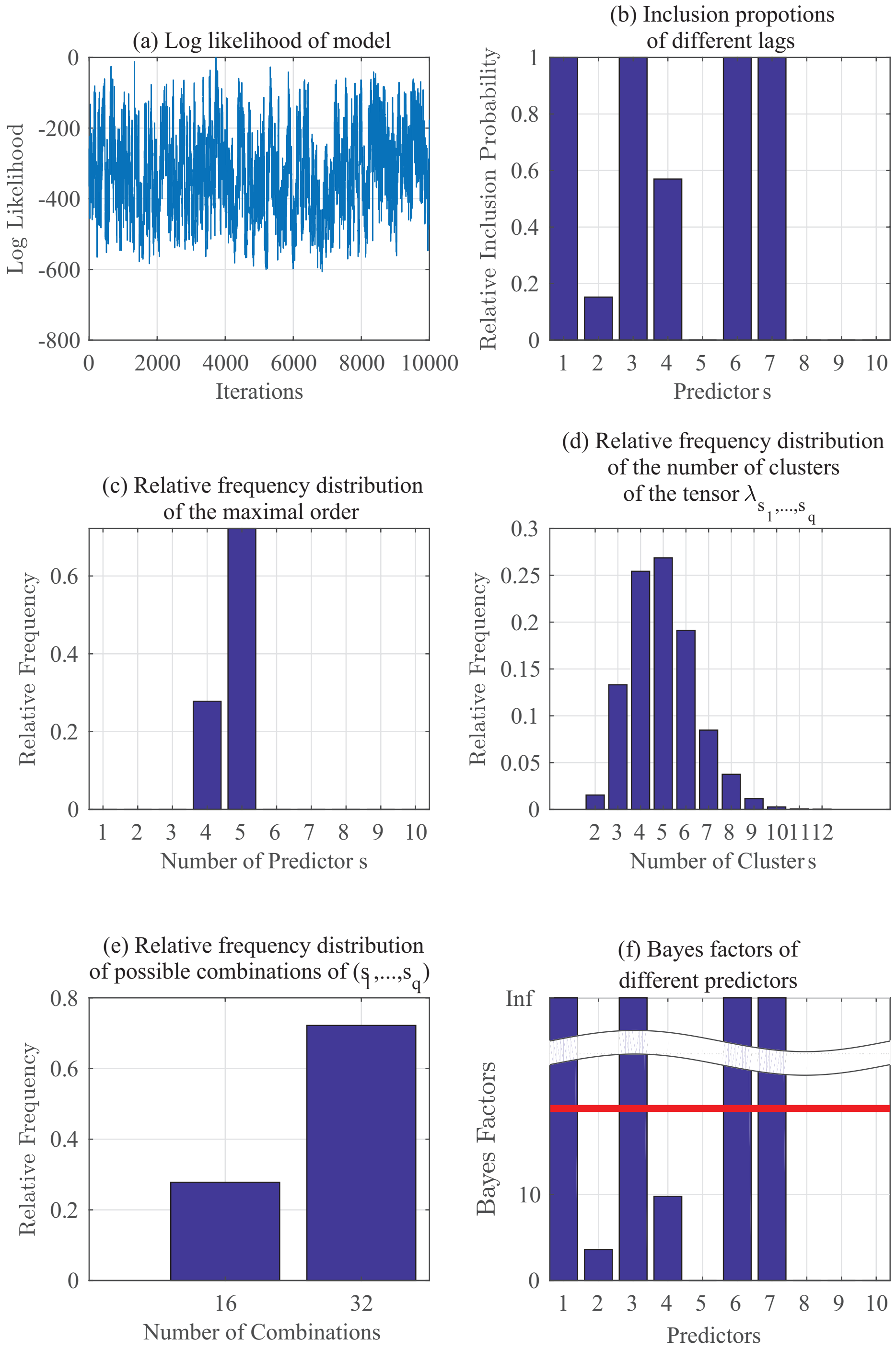

Figure 3. Gibbs sampling results: Numerical example for $p(\theta_t \mid y_{t-1}, \ldots, y_{t-5}, \theta_{t-1}, \ldots, \theta_{t-5})$.

Besides the ability to correctly identify the structure of the model, the proposed method can also perform transition probability estimation. Figure 4 illustrates two arbitrarily selected cases from Table 1 and Table 2. Setting $y_{t-1} = 0$, $y_{t-3} = 1$, and $y_{t-4} = 0$, from Table 1 we can get the transition probability of the model of $(y_t = 1)$ is 0.70. Similarly, setting $y_{t-1} = 1$, $y_{t-3} = 0$, $\theta_{t-1} = 1$, and $\theta_{t-2} = 0$, from Table 2 we can get the transition probability of $(y_t = 1)$ is 0.50. In Figure 4, the estimated transition probability using the proposed method is displayed along with their running mean as well as their 5% and 95% percentiles. From both subplots of Figure 4, it is observed that the running mean of the transition probability is actually close to the true transition probability as given in the data generation tables. Even with a limited amount of data, the proposed method can not only estimate the transition probabilities, but also give an uncertainty bound in terms of their respective quantiles.

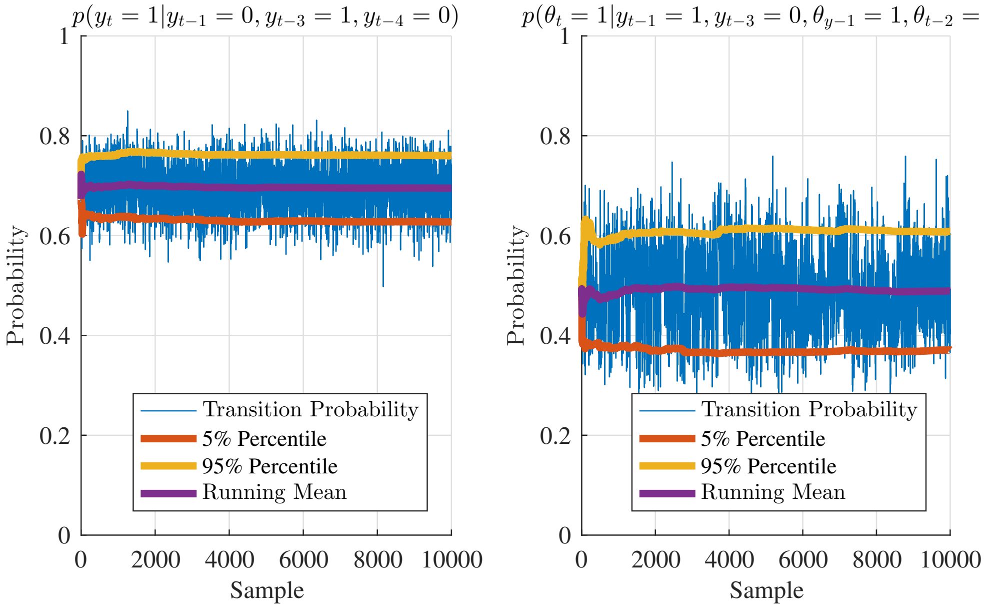

Figure 4. Transition probabilities in the numerical example.

The causal relationship between $\boldsymbol{y}$ and θ is identified by Bayes factor analysis (see Section 3.2. The results are summarized in Table 3, which show that y Granger-causes θ but not the other way, which is in line with the ground truth.

Table 3. Hypothesis Testing of Granger causality in the Numerical Example.

Null Hypothesis	**Bayes Factor BF_{10}**
θ does not Granger-cause y	0.43
y does not Granger-cause θ	Infinity

6. Validation with Experimental Data: The Combustor Apparatus

This section validates the nonparametric regression model with experimental data generated from a swirl-stabilized lean-premixed laboratory-scale combustor apparatus [14].

6.1. Background and Description of the Experimental Procedure

This subsection presents a brief background of thermoacoustic instabilities in the combustor apparatus along with the experimental details for data collection. Thermoacoustic instabilities occur from highly nonlinear coupled phenomena that evolve from mutual interactions among thermofluid dynamics, unsteady heat release, and acoustics of the combustor chamber. The resulting self-sustained high-amplitude pressure oscillations often impose severe negative impacts on the performance and operational life of gas turbine engines [25–27].

Technical literature abounds with studies on combustion instabilities and their early detection by time series analysis, especially by using Markov chains [28,29]. However, current methods are largely limited to individual investigations of pressure or chemiluminescence measurements, and have apparently not taken the machine-learning-theoretic approach to information fusion into consideration; consequently, fast detection of thermoacoustic instabilities may not be achieved to the full extent based on the individual information of different sources only. Moreover, parameter estimation is difficult in current methods, even for moderately high-order Markov chains, due to the paucity of data, let alone a more sophisticated information fusion model. As for the detection procedure, empirical thresholds are often used in existing literature, without taking advantage of methods in statistical detection theory (such as sequential testing techniques); therefore, those applications are very limited in real-time detection cases.

Figure 5 presents a schematic diagram of the combustor apparatus [14] that consists of an inlet section, an injector, a combustion chamber, and an exhaust section. The combustor chamber consists of an optically-accessible quartz section followed by a variable-length steel section.

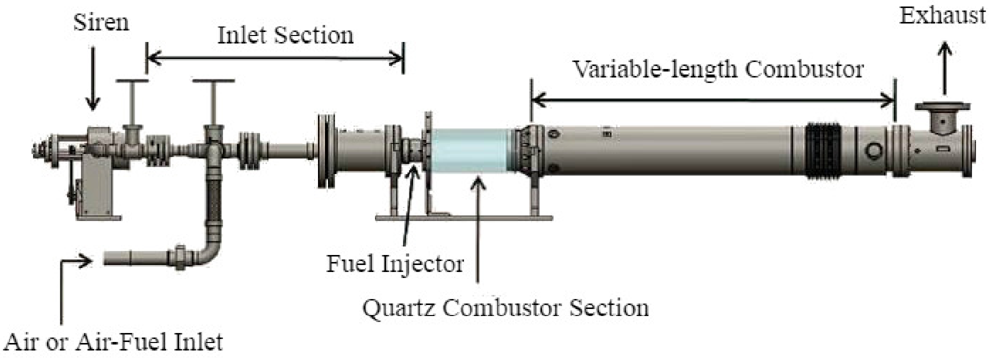

Figure 5. Schematic diagram of the combustor apparatus.

Experiments have been conducted at 62 different operating conditions by varying the equivalence ratio and percentage of pilot fuel, as listed in Table 4. Under each operating condition, 8 s of pressure and chemiluminescence measurements have been collected at the sampling rate of 8192 Hz, where stable and/or unstable modes are recorded along with each time series data. To alleviate the problem of (possible) oversampling, the pressure and chemiluminescence measurements from combustors are first downsampled, which is obtained from first minimum of the average mutual information [30]. Then, the continuously varying time series data for both stable and unstable modes are quantized using maximum entropy partitioning [31,32] with a ternary alphabet $\Sigma = \{1, 2, 3\}$. The quantized pressure measurements are denoted as y_t and the chemiluminescence measurements are denoted as θ_t at time instant t.

Table 4. Operating conditions.

	Parameters	**Values**
Variables	Equivalence Ratio Pilot Fuel (percent)	0.525, 0.538, 0.575, 0.625 0–9% (0.5% increment)
Fixed Conditions	Inlet Temperature Inlet Velocity Combustor Length	250 °C 40 m/s 0.625 m

6.2. Training Phase

This subsection describes details in the nonparametric regression model training, wherein 500 samples have been used after downsampling the quantized pressure time series data under stable and unstable conditions. The maximum memory D of each of y_t and θ_t in this dataset is observed to be generally limited to 5 for both stable and unstable cases. Hence, predictors of y_t or θ_t are set to be $z_t \equiv (y_{t-1}, y_{t-2}, \ldots, y_{t-5}, \theta_{t-1}, \theta_{t-2}, \ldots, \theta_{t-5})$ and the corresponding regression model is hereafter referred to as "full order model". Since y_t and θ_t has three categories, it follows that $C_y = C_\theta = 3$.

To compute posteriors, as in Algorithm 1, the values

$$[1,\ 1.5,\ 2.0,\ 2.5,\ 3.0,\ 1.0,\ 1.5,\ 2.0,\ 2.5,\ 3.0]$$

are assigned to μ_j for $j = 1, \dots, 10$. After discarding 200,000 data points during the burn-in period, remaining 50,000 samples are then downsampled by taking every 5th data point to reduce their autocorrelation. Gibbs sampling results of pressure data are represented as $p(y_t \mid y_{t-1}, \dots, y_{t-5}, \theta_{t-1}, \dots, \theta_{t-5})$ in Figure 6a,b for a stable mode and in Figure 6c,d for an unstable mode. Similarly, Gibbs sampling results of chemiluminescence data are represented as $p(\theta_t \mid y_{t-1}, \dots, y_{t-5}, \theta_{t-1}, \dots, \theta_{t-5})$ in Figure 7a,b for a stable mode and in Figure 7c,d for an unstable mode.

Figures 6a,c and 7a,c show the log likelihood with different iterations for pressure and chemiluminescence data under stable and unstable conditions, respectively. Similarly, Figures 6b,d and 7b,d illustrate the Bayes factors of predictors for pressure and chemiluminescence data under stable and unstable conditions, respectively. Based on the Bayes factor analysis, the important predictors for stable pressure data are identified as:

$$y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, \theta_{t-1}, \theta_{t-3} \text{ and } \theta_{t-4}$$

while those for unstable pressure data are identified as

$$y_{t-1}, y_{t-3}, y_{t-4}, y_{t-5}, \theta_{t-1}, \theta_{t-3}, \theta_{t-4} \text{ and } \theta_{t-5}$$

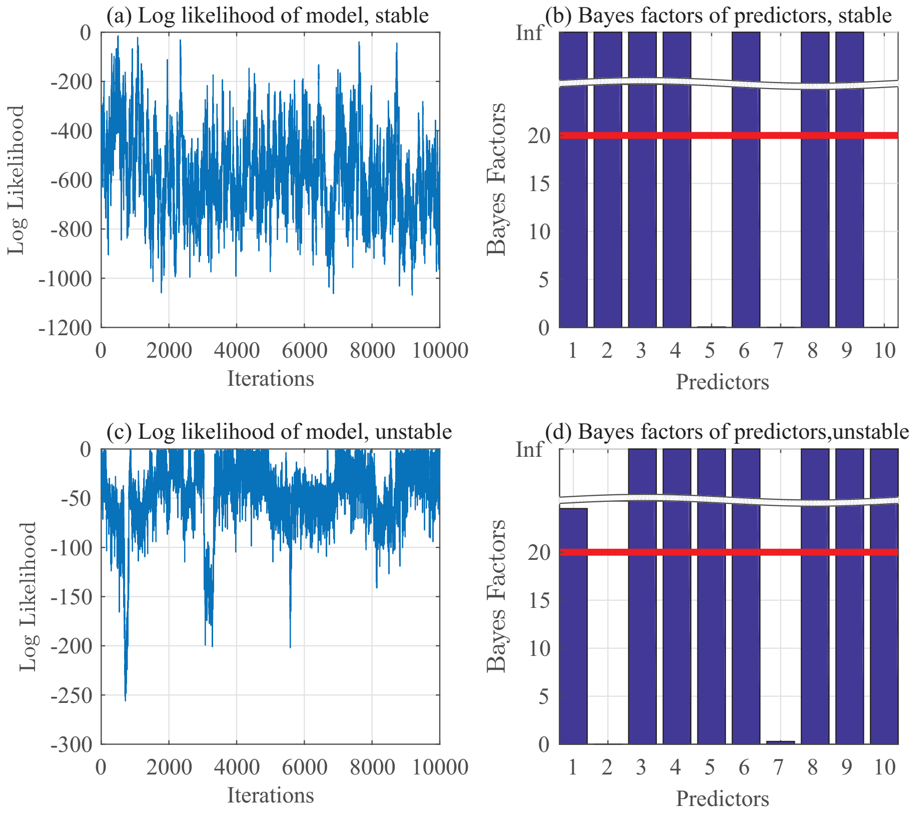

Figure 6. Gibbs sampling of pressure data.

Using the identical set of hyperparameters and number of iterations, Gibbs sampling has been performed on the same set with pressure data y_t only; this is referred to as the "reduced order model" in the following text. In this case the predictors are set as $z_t \equiv (y_{t-1}, y_{t-2}, \dots, y_{t-5})$. The stable and unstable cases are shown in Figure 8a–d respectively. The important predictors for y_t using this reduced order model are: y_{t-2}, y_{t-4}, and y_{t-5} for the stable mode, and $y_{t-1}, y_{t-2}, y_{t-4}$, and y_{t-5} for the unstable mode.

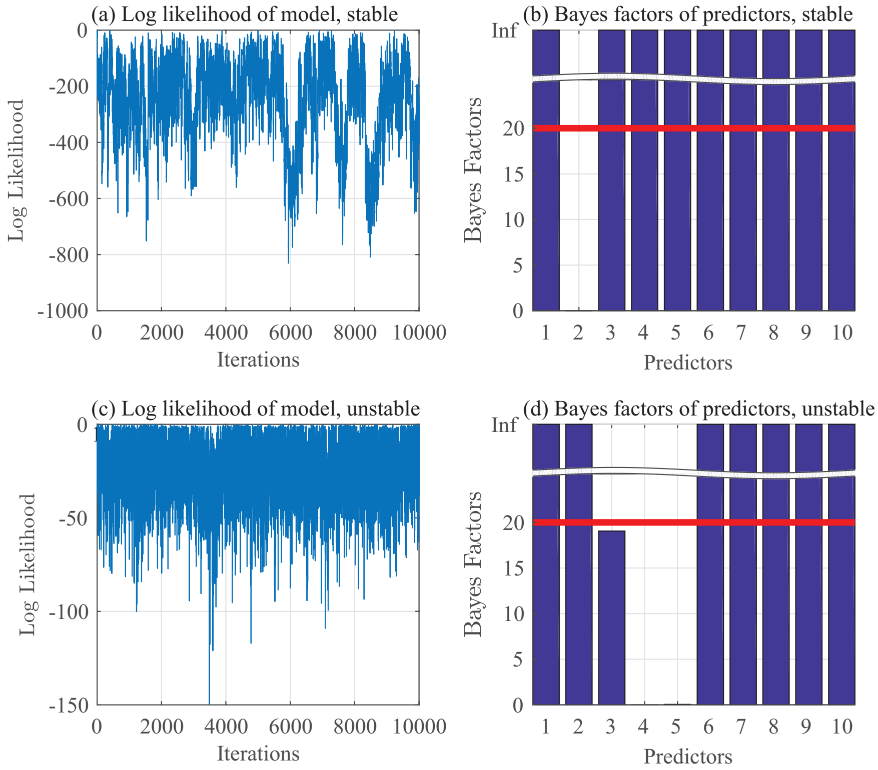

Figure 7. Gibbs sampling of chemiluminescence data.

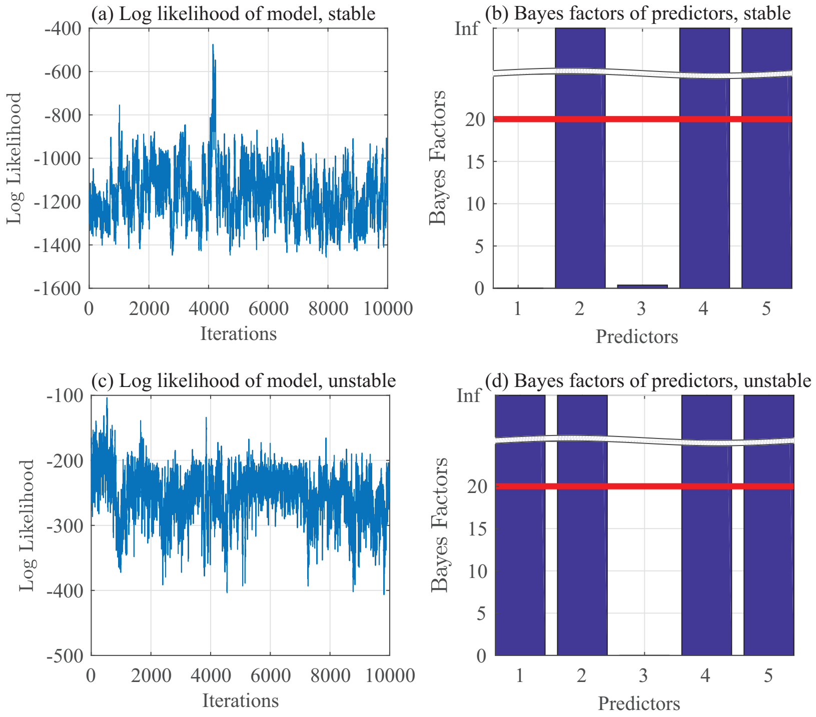

Figure 8. Gibbs sampling of the reduced-order model.

Similarly, for chemiluminescence data, the important predictors are identified as:

$$y_{t-1}, y_{t-3}, y_{t-4}, y_{t-5}, \theta_{t-1}, \theta_{t-2}, \theta_{t-3}, \theta_{t-4} \text{ and } \theta_{t-5}$$

while those for unstable chemiluminescence data are identified as:

$$y_{t-1}, y_{t-2}, \theta_{t-1}, \theta_{t-2}, \theta_{t-3}, \theta_{t-4} \text{ and } \theta_{t-5}$$

6.3. Granger Causality

To identify the Granger causal relationship between pressure and chemiluminescence data, Bayes factor analysis has been performed for both stable and unstable cases as described in Section 3.2. The results are summarized in Table 5, which show that pressure and chemiluminescence measurements Granger-cause each other under both stable and unstable conditions; this implies that fusion of these two measurements can enhance the accuracy of prediction. This kind of mutual interaction between pressure and chemiluminescence measurements could be caused by a third unknown physical quantity, the exploration of which is a topic of future research.

Table 5. Hypothesis Testing of Granger Causality.

Null Hypothesis	Operating Condition	BF_{10}
θ does not Granger-cause y	Stable	Infinity
y does not Granger-cause θ	Stable	Infinity
θ does not Granger-cause y	Unstable	Infinity
y does not Granger-cause θ	Unstable	Infinity

6.4. Sequential Classification

For evaluation of the performance of the sequential classification for thermoacoustic instability identification, 100 instances of 50-sample datasets, which are not included in the training set, have been selected (also from their downsampled quantized pressure measurements for both stable and unstable modes). Figure 9 exhibits the profiles of posterior probability of each class as a function of the length of the observed data, where the top plate (i.e., Figure 9a) uses the full order model, and the bottom plate (i.e., Figure 9b) uses the reduced-order model for the same test data sequence. While the test sequences are correctly classified by both models, the reduced-order model is slower than the full-order model that contains more information.

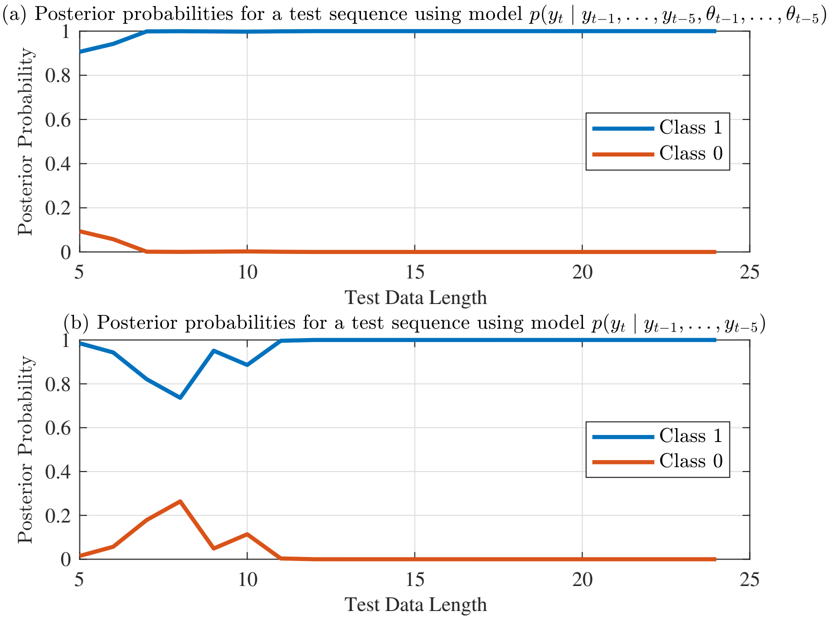

Figure 9. Posterior probabilities using different models.

Figure 10 shows the receiver operating characteristic (ROC) curves for the proposed detection algorithm with different lengths of the test data. These ROC curves are plotted for both full-order and reduced-order models to show that, when testing with the same dataset, the full order model achieves better detection performance in terms of the area under the ROC. In other words, the full-order model may achieve the same performance as the reduced order model in a shorter time, which is desirable for active control of thermoacoustic instabilities in real time. It is also observed that the ROC curves tend to improve (i.e., move toward the top left corner) considerably as the length of test data is increased from 5 to 9. This is expected because the information contents monotonically increase with the length of test data and hence better results are obtained.

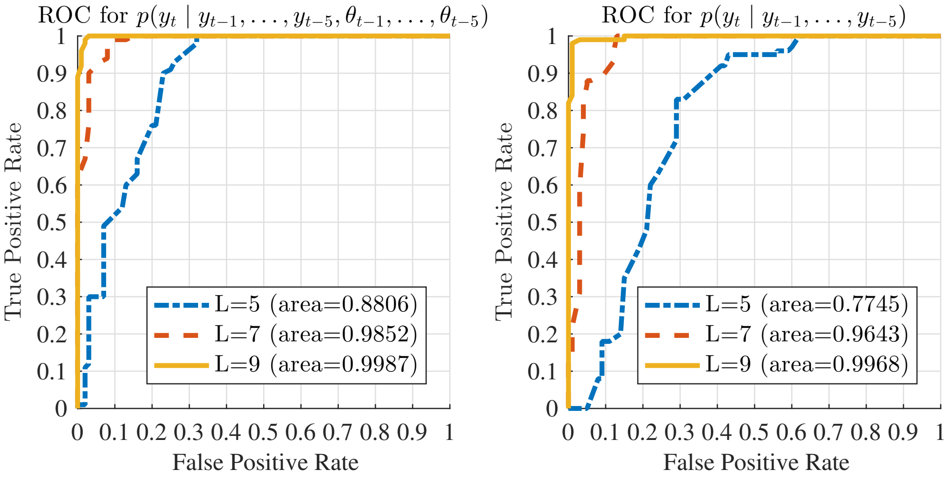

Figure 10. ROC curves with different test data length L.

7. Validation with Economics Data

This section validates the nonparametric regression model with (publicly available) real-world economics data. Specifically, monthly data of the U.S. consumer price index (CPI) and the U.S. Dollar London Interbank Offered Rate (LIBOR) interest rate index with one-month maturity from January 1986 to December 2016 are used. It is noted that: (i) U.S. CPI is a measure of the average change over time in the prices paid by urban consumers for U.S. market of consumer goods and services, and (ii) U.S. Dollar LIBOR is a benchmark for short-term interest rates around the world, which is not a monetary measure associated with any country, and which does not reflect any institutional mandate in contrast to, e.g., when the Federal Reserve sets interest rates. Economics theory [33] indicates that low interest rates can cause high inflation, and empirical research [34] has been conducted to investigate the causal relationship between inflation and nominal or real interest rates for the same country or region.

To avoid spurious regression [35], the raw data of U.S. CPI and U.S. Dollar LIBOR are preprocessed to achieve stationarity. U.S. CPI raw data are used to calculate the monthly percentage increase, and then this percentage increase is converted into a categorical variable by discretizing to quintiles (e.g., 5-quantiles in this study) that are denoted as y_t; the rationale for discretization of (noise-contaminated) continuously varying data is to improve the signal-to-noise ratio [36]. Similarly, U.S. LIBOR raw data are used to calculate the monthly difference, and then this difference is convertedin to a categorical variable by discretizing to quintiles, denoted as θ_t. The entire dataset is used for training the proposed algorithm.

To estimate the regression model in Equation (1), based on the assertion that y_{t-D_y} and θ_{t-D_θ} are not important for predicting y_t and θ_t if both D_y and D_θ are greater than 6 (i.e, six months for both CPI and LIBOR), the predictor for y_t and θ_t is set as:

$$z_t \equiv (y_{t-1}, y_{t-2}, \ldots, y_{t-6}, \theta_{t-1}, \theta_{t-2}, \ldots, \theta_{t-6}) \tag{22}$$

To compute the posterior probabilities using the proposed Algorithm 1, μ_j are assigned to be $j/2$ for $j = 1, \ldots, 6$ and $(j-6)/6$ for $j = 7, \ldots, 12$. After the initial 100,000 samples are discarded during the burn-in period, the remaining 50,000 samples are then downsampled by taking every 5th to reduce their autocorrelation. Figures 11 and 12 respectively summarize the results for y_t and θ_t. These figures have similar characteristics to their counterparts in the numerical example in Section 5. The results show that, for y_t or CPI, the important lags are $y_{t-1}, y_{t-2}, y_{t-3}$ and θ_{t-1}. Similarly, for θ_t or LIBOR, the important lags are $\theta_{t-1}, \theta_{t-1}, \theta_{t-3}$. These results show that LIBOR Granger-cause CPI, but not vice versa. This conclusion is summarized by Bayes factor analysis in Table 6.

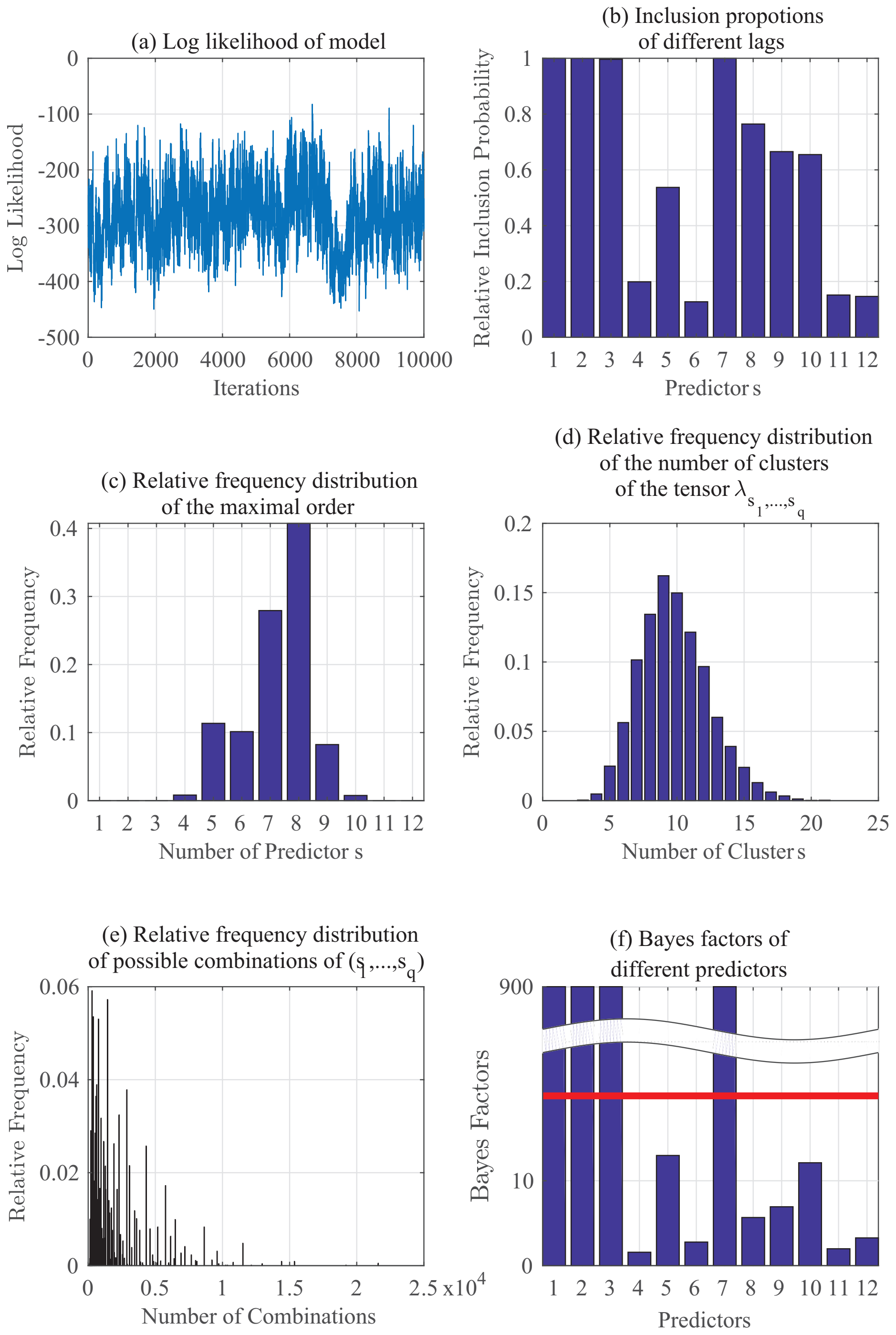

Figure 11. Gibbs sampling of economics dataset for $p(y_t \mid y_{t-1}, \ldots, y_{t-6}, \theta_{t-1}, \ldots, \theta_{t-6})$.

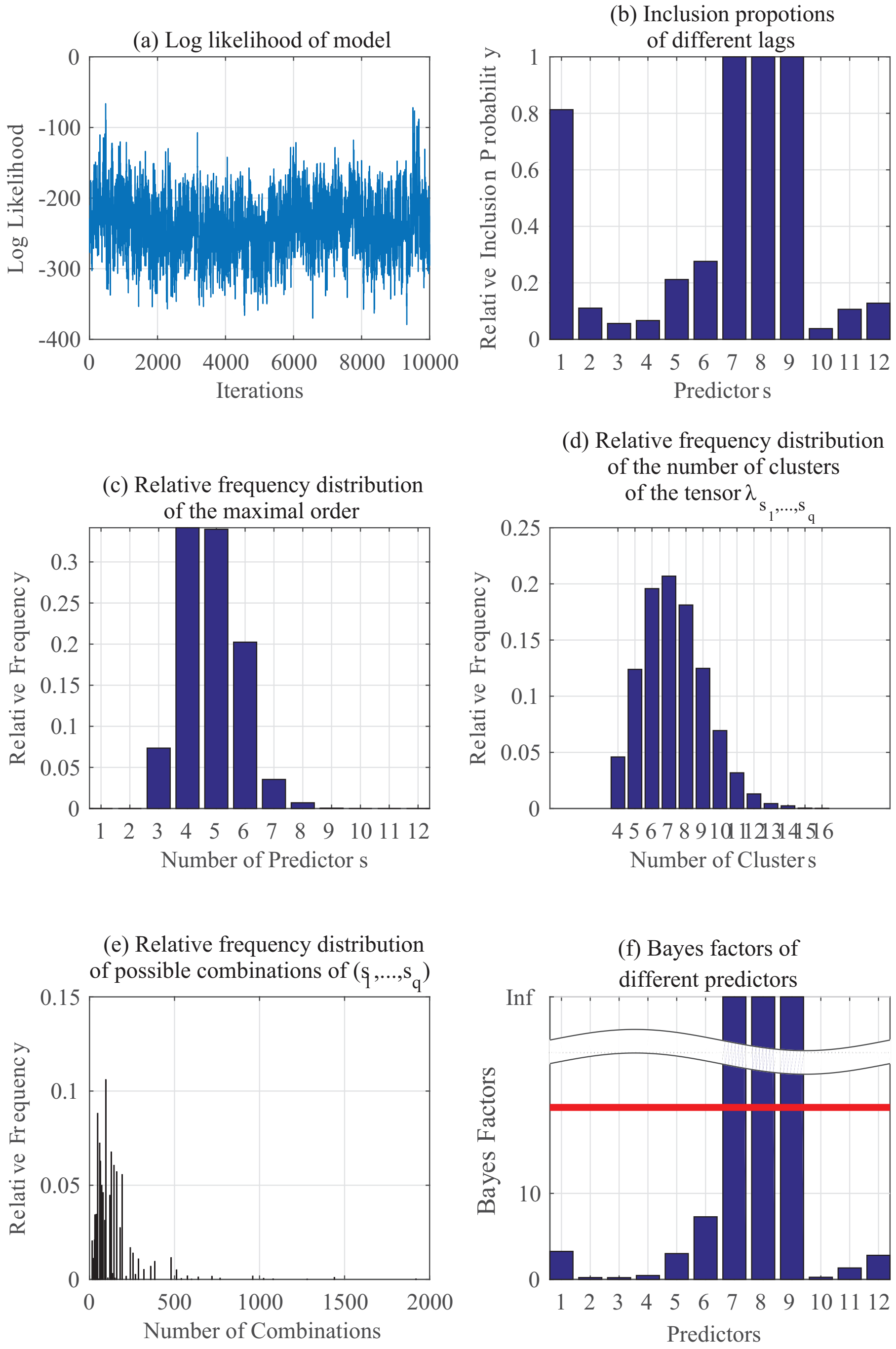

Figure 12. Gibbs sampling of economics dataset for $p(\theta_t \mid y_{t-1}, \ldots, y_{t-6}, \theta_{t-1}, \ldots, \theta_{t-6})$.

Table 6. Hypothesis test of Granger Causality for economics data.

Null Hypothesis	Bayes Factor BF_{10}
US CPI does not Granger-cause LIBOR	7.29
LIBOR does not Granger-cause US CPI	Infinity

8. Summary, Conclusions, and Future Work

The proposed Bayesian nonparametric method provides a flexible model for information fusion of heterogeneous, correlated time series data. The proposed method has been validated on a real-world application by using the experimental data collected from a laboratory-scale swirl-stabilized combustor apparatus, as well as on the publicly available economics data. It is demonstrated that the proposed method is capable of enhancing the accuracy for real-time detection of thermoacoustic instabilities and correctly identifying the Granger causal relationship between key economic variables.

There are many promising directions in which the proposed model can be further explored, such as:

1. Variational inference algorithm development for the proposed model [37].
2. Extension of the present analysis to hidden Markov models (HMM) [38] and information transfer [39].
3. Exploration of an unknown physical quantity that may cause the appearance of mutual interactions between pressure and chemiluminescence measurements.
4. Investigation of the empirical performance of the proposed approach utilizing extensive simulation studies.

Author Contributions: Conceptualization, S.X. and A.R.; Methodology, S.X. and A.R.; Software, Y.F. and S.X.; Validation, S.X., Y.F. and A.R.

Acknowledgments: The authors are grateful to Domenic Santavicca and Jihang Li from Pennsylvania State University for kindly providing the experimental data that were used to validate the theoretical results.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature of Pertinent Parameters

a	Hyperparameter of prior on probability vector $\boldsymbol{\pi}$
b	Hyperparameter of prior on probability vector $\boldsymbol{\pi}$
C_j	Number of categories of the jth predictor
$\mathcal{C}_i$	ith class of dynamical systems
D_y	Number of time-lags of variable y
D_θ	Number of time-lags of variable θ
$\tilde{k}_j$	Number of clusters formed by $\boldsymbol{x}_j$
k_j	Dimension of the jth mixture probability vector
$\boldsymbol{k}$	Vector $\{k_j\}_{j=1}^q$
L	Number of truncations in a Pitman-Yor process
N	Number of iterations in Algorithm 1
q	Number of predictors
s	Realization of a latent allocation-class variable
T	Number of pairs of variables and predictors
$x_{j,t}$	jth latent allocation-class variables at time t
$\boldsymbol{x}_j$	jth latent allocation-class variables $\{x_{j,t}\}_{t=1}^T$
$\boldsymbol{x}_t$	Latent allocation-class variables $\{x_{j,t}\}_{j=1}^q$ at time t
$\boldsymbol{x}$	Latent allocation-class variables $\{\boldsymbol{x}_t\}_{t=1}^T$
y_t	Variable y at time t
$\boldsymbol{y}$	Variables $\{y_t\}_{t=1}^T$
$z_{j,t}$	jth predictor at time t
$\boldsymbol{z}_j$	jth predictors $\{z_{j,t}\}_{t=1}^T$
$\boldsymbol{z}_t$	Predictors $\{z_{j,t}\}_{j=1}^q$ at time t
$\boldsymbol{z}$	Predictors $\{\boldsymbol{z}_t\}_{t=1}^T$
α	Hyperparameter of prior on $\boldsymbol{\lambda}$
β_j	Hyperparameter of prior on $\boldsymbol{\omega}_j$
θ_t	Variable θ at time t
$\boldsymbol{\Theta}$	Threshold
$\boldsymbol{\lambda}_{s_1,\ldots,s_q}$	Probability vector $\{\lambda_{s_1,\ldots,s_q}(c)\}_{c=1}^{C_0}$
Λ	Set of predictors
$\tilde{\boldsymbol{\lambda}}$	Conditional probability tensor $\{\boldsymbol{\lambda}_{s_1,\ldots,s_q}\}_{s_1,\ldots,s_q}$

$\boldsymbol{\lambda}_l$	Probability vector $\{\lambda_l(c)\}_{c=1}^{C_0}$
$\boldsymbol{\lambda}$	Sequence $\{\boldsymbol{\lambda}_l\}_{l=1}^{\infty}$
μ_j	Hyperparameter of prior on k_j
$\boldsymbol{\pi}$	Probability vector $\{\pi_l\}_{l=1}^{\infty}$
$\boldsymbol{\phi}$	Collection $\{\phi_{s_1,\ldots,s_q}\}_{s_1,\ldots,s_q}$
$\boldsymbol{\psi}^{(k)}$	Time-invariant spatial variables for kth experiment
$\boldsymbol{\omega}^{(j)}(c)$	Mixture probability vector $\{\omega_s^{(j)}(c)\}_{s=1}^{k_j}$
$\boldsymbol{\omega}^{(j)}$	Mixture probability matrix $\{\boldsymbol{\omega}_s^{(j)}(c)\}_{c=1}^{C_j}$
$\boldsymbol{\omega}$	Mixture probability tensor $\{\boldsymbol{\omega}^{(j)}\}_{j=1}^{q}$

Pertinent Acronyms

BF	Bayes Factor
Beta	Beta Distribution
Dir	Uniform Dirichlet Distribution
HOSVD	Higher order singular value decomposition
Mult	Multinomial Distribution
ROC	Receiver operating characteristic

References

1. Sarkar, S.; Virani, N.; Ray, A.; Yasar, M. Sensor fusion for fault detection and classification in distributed physical processes. *Phys. C Supercond.* **2014**, *1*, 369–373. [CrossRef]
2. Kónya, L. Exports and growth: Granger causality analysis on oecd countries with a panel data approach. *Econ. Model.* **2006**, *23*, 978–992. [CrossRef]
3. Seth, A.K.; Barrett, A.B.; Barnett, L. Granger causality analysis in neuroscience and neuroimaging. *J. Neurosci.* **2015**, *35*, 3293–3297. [CrossRef] [PubMed]
4. Annaswamy, A.M.; Ghoniem, A.F. Active control of combustion instability: Theory and practice. *IEEE Control Syst.* **2002**, *22*, 37–54. [CrossRef]
5. Fujimaki, R.; Nakata, T.; Tsukahara, H.; Sato, A.; Yamanishi, K. Mining abnormal patterns from heterogeneous time-series with irrelevant features for fault event detection. *Stat. Anal. Data Min.* **2009**, *2*, 1–17. [CrossRef]
6. Virani, N.; Marcks, S.; Sarkar, S.; Mukherjee, K.; Ray, A.; Phoha, S. Dynamic data driven sensor array fusion for target detection and classification. *Proc. Comput. Sci.* **2013**, *18*, 2046–2055. [CrossRef]
7. Iyengar, S.; Varshney, P.; Damarla, T. A parametric copula-based framework for hypothesis testing using heterogeneous data. *IEEE Trans. Signal Process.* **2011**, *59*, 2308–2319. [CrossRef]
8. Spirtes, P. Introduction to causal inference. *J. Mach. Learn. Res.* **2010**, *11*, 1643–1662.
9. Eichler, M. Causal inference in time series analysis. *Causal. Stat. Perspect. Appl.* **2012**, 327–354. [CrossRef]
10. Athey, S. Machine learning and causal inference for policy evaluation. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 5–6.
11. Granger, C.W. Causality, cointegration, and control. *J. Econ. Dyn. Control* **1988**, *12*, 551–559. [CrossRef]
12. Tank, A.; Fox, E.; Shojaie, A. Granger causality networks for categorical time series. *arXiv* **2016**, arXiv:1706.0278. [CrossRef]
13. Kass, R.E.; Raftery, A.E. Bayes factors. *J. Am. Stat. Assoc.* **1995**, *90*, 773–795. [CrossRef]
14. Kim, K.; Lee, J.; Quay, B.; Santavicca, D. Response of partially premixed flames to acoustic velocity and equivalence ratio perturbations. *Combust. Flame* **2010**, *157*, 1731–1744. [CrossRef]
15. Yang, Y.; Dunson, D.B. Bayesian conditional tensor factorizations for high-dimensional classification. *J. Am. Stat. Assoc.* **2016**, *111*, 656–669. [CrossRef]
16. Wilks, S. *Mathematical Statistics*; John Wiley: New York, NY, USA, 1963.
17. Ferguson, T.S. A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1973**, *1*, 209–230. [CrossRef]
18. Ishwaran, H.; James, L.F. Gibbs sampling methods for stick-breaking priors. *J. Am Stat. Assoc.* **2001**, *96*, 161–173. [CrossRef]
19. Green, P.J. Reversible jump markov chain monte carlo computation and Bayesian model determination. *Biometrika* **1995**, *82*, 711–732. [CrossRef]

20. Pitman, J. Exchangeable and partially exchangeable random partitions. *Probab. Theory Relat. Fields* **1995**, *102*, 145–158. [CrossRef]
21. Miller, J.W.; Harrison, M.T. Mixture models with a prior on the number of components. *arXiv* **2015**, arXiv:1502.06241. [CrossRef]
22. Van Dyk, D.A.; Park, T. Partially collapsed gibbs samplers: Theory and methods. *J. Am. Stat. Assoc.* **2008**, *103*, 790–796. [CrossRef]
23. Akaike, H. Factor analysis and aic. *Psychometrika* **1987**, *52*, 317–332. [CrossRef]
24. Poor, H.V. *An Introduction to Signal Detection and Estimation*; Springer Science & Business Media: Heidelberg/Berlin, Germany, 2013.
25. Lieuwen, T.; Torres, H.; Johnson, C.; Zinn, B.T. A mechanism of combustion instability in lean premixed gas turbine combustors. In *ASME 1999 International Gas Turbine and Aeroengine Congress and Exhibition*; American Society of Mechanical Engineers: New York, NY, USA, 1999.
26. Dowling, A.; Hubbard, S. Instability in lean premixed combustors. *Proc. Inst. Mech. Eng. Part A J. Power Energy* **2000**, *214*, 317–332. [CrossRef]
27. Huang, Y.; Yang, V. Dynamics and stability of lean-premixed swirl-stabilized combustion. *Prog. Energy Combust. Sci.* **2009**, *35*, 293–364. [CrossRef]
28. Jha, D.; Virani, N.; Reimann, J.; Srivastav, A.; Ray, A. Symbolic analysis-based reduced order Markov modeling of time series data. *Signal Process.* **2018**, *149*, 68–81. [CrossRef]
29. Sarkar, S.; Ray, A.; Mukhopadhyay, A. Sen, S. Dynamic data-driven prediction of lean blowout in a swirl-stabilized combustor. *Int. J. Spray Combust. Dyn.* **2015**, *7*, 209–241. [CrossRef]
30. Abarbanel, H.D.; Brown, R.; Sidorowich, J.J.; Tsimring, L.S. The analysis of observed chaotic data in physical systems. *Rev. Mod. Phys.* **1993**, *65*, 1331. [CrossRef]
31. Rajagopalan, V.; Ray, A. Symbolic time series analysis via wavelet-based partitioning. *Signal Process.* **2006**, *86*, 3309–3320. [CrossRef]
32. Mukherjee, K.; Ray, A. State splitting and merging in probabilistic finite state automata for signal representation and analysis. *Signal Process.* **2014**, *104*, 105–119. [CrossRef]
33. Blanchard, O.J.; Fischer, S. *Lectures on Macroeconomics*; MIT Press: Cambridge, UK, 1989.
34. Eichler, M. Granger causality and path diagrams for multivariate time series. *J. Econ.* **2007**, *137*, 334–353. [CrossRef]
35. Österholm, P. The Taylor rule: A spurious regression? *Bull. Econ. Res.* **2005**, *57*, 217–247. [CrossRef]
36. Beim Graben, P. Estimating and improving the signal-to-noise ratio of time series by symbolic dynamics. *Phys. Rev. E* **2001**, *64*, 51104. [CrossRef] [PubMed]
37. Hoffman, M.D.; Blei, D.M.; Wang, C.; Paisley, J. Stochastic variational inference. *J. Mach. Learn. Res.* **2013**, *14*, 1303–1347.
38. Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 267–286. [CrossRef]
39. Schreiber, T. Measuring Information Transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464. [CrossRef] [PubMed]