Bayesian Nonparametric Regression Modeling of Panel Data for Sequential Classification

Sihan Xiong, Yiwei Fu, and Asok Ray, *Fellow, IEEE*

***Abstract*— This paper proposes a Bayesian nonparametric regression model of panel data for sequential pattern classification. The proposed method provides a flexible and parsimonious model that allows both time-independent spatial variables and time-dependent exogenous variables to be predictors. Not only this method improves the accuracy of parameter estimation for limited data, but also it facilitates model interpretation by identifying statistically significant predictors with hypothesis testing. Moreover, as the data length approaches infinity, posterior consistency of the model is guaranteed for general data-generating processes under regular conditions. The resulting model of panel data can also be used for sequential classification. The proposed method has been tested by numerical simulation, then validated on an econometric public data set, and subsequently validated for detection of combustion instabilities with experimental data that have been generated in a laboratory environment.**

***Index Terms*— Bayes factor, Bayesian nonparametric, conditional tensor factorization, panel data, posterior consistency, regression model, thermoacoustic instability.**

I. Introduction

STATISTICAL analysis of panel data has evolved as a subdiscipline in pattern analysis with applications to both scientific and social disciplines, such as ecology, meteorology, and health economics. Panel data may contain observations of multiple phenomena over multiple time periods for the same set of units (e.g., experimental conditions), naturally arising from complex systems, where the key variables may interact with each other and evolve with time. Several methods have been proposed to model panel data in econometric literature. Cross-lagged structural equation model [1] facilitates investigation of causal relationships between two variables through regression on the lagged score of both variables, including fixed-effects model [2] and random-effects model [3]. These developments are aimed for continuously varying data from the perspectives of frequentist estimation methods (e.g., maximum likelihood estimation and generalized methods of moments). Often these methods have relatively poor performance for small-size data and hence restrict the underlying models to be of low order.

Manuscript received January 9, 2017; revised June 5, 2017, August 3, 2017, and September 1, 2017; accepted September 8, 2017. Date of publication October 12, 2017; date of current version August 20, 2018. This work was supported by the U.S. Air Force Office of Scientific Research under Grant FA9550-15-1-0400. *(Corresponding author: Asok Ray.)*

The authors are with the Department of Mechanical Engineering, Pennsylvania State University, University Park, PA 16802-1412 USA (e-mail: sux101@psu.edu; yxf118@psu.edu; axr2@psu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2017.2752005

In the machine learning discipline, there exists a plethora of methods for pattern classification of high-dimension variables, such as support vector machines (SVMs) [4], deep neural networks [5], and extreme learning machines [6]. Even though these algorithms may achieve good accuracy in prediction tasks, it is often difficult for users to interpret the results. Recently, Bayesian nonparametric methods have received much attention from the research community. For example, Dirichlet process priors are assigned on latent variables for classification of ordinal or categorical variables [7], [8]. However, these methods are only suitable for independent identically distributed (IID) data and lack systematic hypothesis testing methods.

Sarkar and Dunson [9] have proposed a Bayesian nonparametric modeling of high-order ergodic Markov chain, which facilitate calculation of Bayes factors [10] for a variety of hypotheses of interest. Since large-sample properties (e.g., consistency [11]) of this method are not guaranteed in general for panel data, they are not suitable for modeling of data generated from systems under different experimental conditions. To alleviate the above limitations, this paper proposes a Bayesian nonparametric regression model of categorical panel data for applications to pattern classification in dynamic data-driven application systems [12]. The merits of the proposed method compared with [9] are delineated in the following.

1) Development of a flexible model containing multiple time-independent spatial variables and time-dependent exogenous variables, which is capable of capturing spatio-temporal characteristics for sequential classification (e.g., detection of combustion instabilities [13]).
2) Theoretically rigorous guarantee of posterior consistency for general (e.g., nonstationary) data-generating processes under regular conditions.
3) Numerical and experimental validation of the proposed method in various applications, which shows superior performance compared with [9] for a real-time application.

This paper is organized in seven sections including this section. Section II develops the model and establishes the results of posterior consistency. Section III presents the Gibbs sampling algorithm for posterior computation and Bayes factor analysis for hypothesis testing. Section IV describes the sequential classification method based on the proposed model. The algorithms of the proposed method are tested by numerical simulation in Section V and with a public data set in Section VI, while Section VII validates the proposed

method with experimental data, generated on a swirl-stabilized lean-premixed laboratory-scale combustor, for early detection of combustion instabilities. Finally, Section VIII concludes this paper along with a few recommendations for the future research.

II. Model Development From Panel Data

This section first describes the procedure of data collection and the development of a regression model for panel data. Then, algebraic and statistical specifications of the proposed model are elaborated. Finally, the posterior consistency of the proposed model is presented.

A. Construction of a Regression Model

Let $\boldsymbol{\psi} \equiv (\psi_1, \ldots, \psi_N)$ represent time-independent spatial variables having K possible combinations, each of which denotes an experimental condition. For the kth experiment $\boldsymbol{\psi}^{(k)} \equiv (\psi_1^{(k)}, \ldots, \psi_N^{(k)})$, the collected categorical time-series data are represented as: $\{y_t^{(k)}, \boldsymbol{\theta}_t^{(k)}\}_{t=1}^{T_k}$, where $y_t^{(k)}$ is the response variable at time instants t ranging from 1 to T_k, and $\boldsymbol{\theta}_t^{(k)} = (\theta_{1,t}^{(k)}, \ldots, \theta_{M,t}^{(k)})$ represents the time-dependent exogenous variables for the kth experiment. Let $\mathcal{F}_{t-1}^{(k)} = \sigma(\{y_\tau^{(k)}\}_{\tau=1}^{t-1}, \{\boldsymbol{\theta}_\tau^{(k)}\}_{\tau=1}^{t}, \boldsymbol{\psi}^{(k)})$ and $\mathcal{F}_{t-1} = \sigma(\{\mathcal{F}_{t-1}^{(k)}\}_{k=1}^{K})$ be the filtration of interest.

The steps for constructing a regression model of categorical panel data are as follows. Given the external variables and the most recently generated D response variables, the distribution of the current response variable is independent of all other variables, that is

$$p\big(y_t^{(k)}|\mathcal{F}_{t-1}\big) = p\big(y_t^{(k)}|y_{t-1}^{(k)}, \ldots, y_{t-D}^{(k)}, \boldsymbol{\theta}_t^{(k)}, \boldsymbol{\psi}^{(k)}\big). \quad (1)$$

Essentially, given the spatial variables, $\{y_t^{(k)}\}_{t=1}^{T_k}$ is modeled as a Markov chain of order D, whose transition probability may be time-varying and is determined by the exogenous variables $\{\boldsymbol{\theta}_t^{(k)}\}_{t=1}^{T_k}$. If there is no spatial variable, the proposed model reduces to a time-series model that can be used to represent the nonstationary Markov chain. Similarly, when there is no time-dependent exogenous variable, the proposed model becomes the most basic spatio-temporal model, for which time series $\{y_t^{(k)}\}_{t=1}^{T_k}$ is a time-homogeneous Markov chain under each spatial condition $\boldsymbol{\psi}^{(k)}$.

Remark 1: Given $\boldsymbol{\theta}_t^{(k)}, \boldsymbol{\psi}^{(k)}$, the important time lags in determining the distribution of $y_t^{(k)}$ could be an arbitrary subset of $(y_{t-1}^{(k)}, \ldots, y_{t-D}^{(k)})$ and the maximal order (always less or equal than D) is the minimum order beyond which the lags are not important.

B. Conditional Tensor Factorization

For simplicity of notations, predictors $\boldsymbol{z}_t \equiv (z_{1,t}, \ldots, z_{q,t})$ are substituted for $(y_{t-1}, \ldots, y_{t-D}, \boldsymbol{\theta}_t, \boldsymbol{\psi})$, where the first D predictors represent the time lags of response variable y and the rest stand for external variables. Let y_t have C_0 categories and $z_{j,t}$ have C_j categories for $j = 1, \ldots, q$. Because the response variables y and its time lags have the same number of categories, it follows that $C_0 = C_1 = \cdots = C_D$. The quantity $p(y_t|\boldsymbol{z}_t)$ is treated as a $(q+1)$th order tensor in the $C_0 \times C_1 \ldots \times C_q$ dimensional space, called the conditional probability tensor. It was first reported in [8] that every conditional probability tensor has the following higher order singular value decomposition (HOSVD) as:

$$p(y_t|\boldsymbol{z}_t) = \sum_{s_1=1}^{k_1} \cdots \sum_{s_q=1}^{k_q} \lambda_{s_1,\ldots,s_q}(y_t) \prod_{j=1}^{q} \omega_{s_j}^{(j)}(z_{j,t}) \quad (2)$$

where $1 \le k_j \le C_j$ for $j = 1, \ldots, q$. Moreover, the parameters $\lambda_{s_1 \ldots s_q}(y_t)$ and $\omega_{s_j}^{(j)}(z_{j,t})$ are all nonnegative and satisfy the following constraints:

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

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

Because the factorization in (2) exists for every conditional probability tensor, the above-mentioned constraints are not restrictive but it is ensured that $\sum_{y_t=1}^{C_0} p(y_t|\boldsymbol{z}_t) = 1$.

C. Bayesian Nonparametric Modeling

To construct a statistically interpretable and parsimonious (i.e., low dimensional) model, the tensor factorization in (2) is converted to a Bayes network by introducing latent allocation class variables and assigning sparsity-inducing priors. More formally, T pairs of response variable and predictors are collected and the data set is arranged as $\{y_t, \boldsymbol{z}_t\}_{t=1}^{T}$, where t does not represent time but an index ranging from 1 to T. For later use, let us denote $\boldsymbol{y} \equiv \{y_t\}_{t=1}^{T}$ and $\boldsymbol{z} \equiv \{\boldsymbol{z}_t\}_{t=1}^{T}$. The conditional probability $p(y_t|\boldsymbol{z}_t)$, factorized as in (2), is rewritten in the following form:

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

where $\boldsymbol{x}_t \equiv (x_{1,t}, \ldots, x_{q,t})$ denotes the latent class allocation variables, and $\boldsymbol{x} \equiv \{\boldsymbol{x}_t\}_{t=1}^{T}$. For $j = 1, \ldots, q$ and $t = 1, \ldots, T$, it follows that:

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

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

where **Mult** denotes multinomial distribution and $\boldsymbol{\omega}^{(j)} \equiv \{\{\omega_s^{(j)}(c)\}_{s=1}^{k_j}\}_{c=1}^{C_j}$ is the mixture probability matrix such that the cth row $\boldsymbol{\omega}^{(j)}(c) \equiv \{\omega_s^{(j)}(c)\}_{s=1}^{k_j}$ is a probability vector. Moreover, $\tilde{\boldsymbol{\lambda}} \equiv \{\boldsymbol{\lambda}_{s_1,\ldots,s_q}\}_{(s_1 \ldots s_q)}$ is a conditional probability tensor such that $\boldsymbol{\lambda}_{s_1,\ldots,s_q} \equiv \{\lambda_{s_1,\ldots,s_q}(c)\}_{c=1}^{C_0}$ is a probability vector for each combination $(s_1, \ldots, s_q)$.

The above hierarchical reformulation of HOSVD illustrates how the proposed method enables the model structure to converge toward a low dimension.

1) It follows from (6) that soft clustering is implemented for each predictor $\boldsymbol{z}_j \equiv \{z_{j,t}\}_{t=1}^{T}$ to inherit statistical strength across different categories.

2) It follows from (7) that the distribution of y_t is determined by a reduced-order conditional probability tensor $\tilde{\boldsymbol{\lambda}}$.
3) The clustering assignments $\boldsymbol{x}_j \equiv \{x_{j,t}\}_{t=1}^{T}$ are used to capture the interactions among the predictors in an implicit and parsimonious manner by allowing the latent populations indexed by $(s_1, \ldots, s_q)$ to be shared among the various state combinations of the predictors.

Remark 2: It is important to distinguish between the number of clusters $\tilde{k}_j$ formed by the latent allocation variables $\boldsymbol{x}_j$ and the dimension k_j of the mixing probability vector $\boldsymbol{\omega}^{(j)}(c)$. The former refers to the number of groups formed by the data, and is always smaller than the latter. It is noted that $\tilde{k}_j$ determines the inclusion of the predictor $\boldsymbol{z}_j$ in the model, because $p(y_t|\boldsymbol{z}_t)$ does not vary with $z_{j,t}$ if $\boldsymbol{z}_j$ has only one latent cluster. Therefore, the significance of a particular predictor can be tested based on $\tilde{k}_j$, which is elaborated later in Section III-B.

In real-life applications, the tensor $\tilde{\boldsymbol{\lambda}}$ may have more components than required, because $\prod_{j=1}^{q} k_j$ could be very large even for moderate values of q and C_j. To alleviate this difficulty, $\tilde{\boldsymbol{\lambda}}$ is clustered among different combinations of $(s_1, \ldots, s_q)$ in a nonparametric way by imposing Pitman–Yor process prior [14] on it. Thus, by employing the stick-breaking representation of Pitman–Yor process [15], it follows that:

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

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

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

where **Dir** and **Beta** denote uniform Dirichlet and beta distributions, respectively, and $\boldsymbol{\lambda}_l \equiv (\lambda_l(1), \ldots, \lambda_l(C_0))$. Moreover, $0 \le b < 1$ and $a > -b$. For each combination $(s_1, \ldots, s_q)$

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

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

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

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

Next, priors are assigned on the mixture probability matrix $\boldsymbol{\omega}^{(j)}$. Unlike the tensor $\tilde{\boldsymbol{\lambda}}$, the dimension of $\boldsymbol{\omega}^{(j)}$ grows linearly as k_j increases. Thus, it is not necessary to further cluster $\boldsymbol{\omega}^{(j)}$, and hence independent priors are assigned on the row of $\boldsymbol{\omega}^{(j)}$ for $j = 1, \ldots, q$ as follows:

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

Finally, priors are assigned on the dimension of the mixture probability vector k_j, for $j = 1, \ldots, q$

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

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

Remark 3: The exponential prior in (14) assigns increasing probabilities to smaller values of k_j as the parameter μ_j becomes larger, and it is a uniform prior on $\{1, \ldots, C_j\}$ when μ_j equals to zero. To reflect the prior belief that increasing time lags will have vanishing impact on the distribution of the current response variable, one can assign larger μ_j to more distant time lags. If one has no prior information of a particular predictor (e.g., an external predictor), then setting the corresponding $\mu_j = 0$ is appropriate.

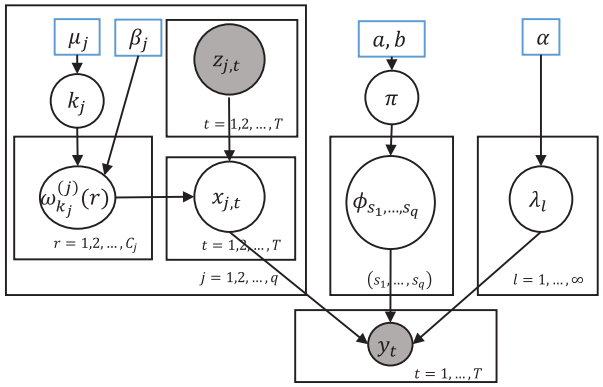


Fig. 1. Graphical representation of the Bayes network. The quantity enclosed by a rectangle represents a deterministic hyperparameter. The quantities enclosed by transparent circles denote unobserved random variables, while those enclosed by shaded circles denote observed random variables.

Combining equations from (6) to (14) yields a Bayes network representation of the model. Fig. 1 shows its dependence structure.

D. Posterior Consistency

In Bayesian paradigm, the posterior distribution is obtained by updating the prior distribution with observed data and it contains all the information for statistical inference. In this context, posterior consistency implies that if the data are indeed generated from a fixed true model, then the posterior should concentrate around it as the data length approaches infinity. Although it is an asymptotic property, posterior consistency plays a central role in Bayesian analysis for the following two reasons [11].

1) The information contents in priors are eventually dominated by those in the data.
2) It is ensured that two Bayesians starting with different priors will ultimately have very close predictive distribution for given sufficiently long data.

Denoting $\mathcal{P}$ as the space of all conditional probability tensors in the form $P(y_t|y_{t-1}, \ldots, y_{t-D}, \boldsymbol{\theta}_t, \boldsymbol{\psi})$, a metric is defined on the space $\mathcal{P}$ as the following norm:

$$||P - P_0|| = \sum_{s_0=1}^{C_0} \cdots \sum_{s_q=1}^{C_q} |P(s_0|s_1, \ldots, s_q) - P_0(s_0|s_1, \ldots, s_q)|. \tag{15}$$

Let $P_0 \in \mathcal{P}$ be the true conditional probability tensor and let p_0 be the corresponding probability law of the true data-generating process. Let Π be the prior on $\mathcal{P}$ induced by the proposed model, and let $\Pi(\bullet|\mathcal{D}_{\boldsymbol{T}})$ denote the corresponding posterior distribution given the observed data set $\mathcal{D}_{\boldsymbol{T}} = \{\{y_t^{(k)}, y_{t-1}^{(k)}, \ldots, y_{t-D}^{(k)}, \boldsymbol{\theta}_t^{(k)}, \boldsymbol{\psi}^{(k)}\}_{t=1}^{T_k}\}_{k=1}^{K}$. Having $\boldsymbol{T} \equiv (T_1, \ldots, T_K)$, we say $\boldsymbol{T} \to \infty$ if $T_k \to \infty$ for any $k \in \{1, \ldots, K\}$.

Before establishing the posterior consistency, it is noted that the proposed model is different from those in usual

cases studied in the literature [16] from the following two perspectives.

1) The collected data are not IID.
2) Since the distribution of external variables may not be known, the analysis is based on only partial information of the data-generating mechanism.

The statistical inference in this setting has been developed from the frequentist perspective to study the consistency of maximum partial likelihood estimator [17]. In this context, Theorem 4 [18] is relevant.

Theorem 4: Suppose $\mathcal{F}_1, \mathcal{F}_2, \ldots$ is a sequence of increasing σ-fields such that $R_t = \sum_{\tau=1}^{t} r_\tau$ is measurable with respect to $\mathcal{F}_t$ for every t. Let $i_t = \mathrm{E}(r_t|\mathcal{F}_{t-1})$, $I_t = \sum_{\tau=1}^{t} i_\tau$ and $j_t = \mathrm{Var}(r_t|\mathcal{F}_{t-1})$, $J_t = \sum_{\tau=1}^{t} j_\tau$. If there exist constants $\delta > 0$, $\gamma_t \to \infty$ such that

$$p(I_t/\gamma_t > \delta) \to 1 \tag{16}$$

$$J_t/\gamma_t^2 \xrightarrow{p} 0 \tag{17}$$

then $|R_t - I_t| \xrightarrow{p} 0$.

To apply Theorem 4 in the present setting, the following notations are introduced for each k:

$$r_t^{(k)}(P) = \log\frac{p_0\left(y_t^{(k)}|\mathcal{F}_{t-1}\right)}{p\left(y_t^{(k)}|\mathcal{F}_{t-1}\right)}, \quad R_t^{(k)}(P) = \sum_{\tau=1}^{t} r_\tau^{(k)}(P)$$

$$i_t^{(k)}(P) = \mathrm{E}\left(r_t^{(k)}(P)|\mathcal{F}_{t-1}\right), \quad I_t^{(k)}(P) = \sum_{\tau=1}^{t} i_\tau^{(k)}(P)$$

$$j_t^{(k)}(P) = \mathrm{Var}\left(r_t^{(k)}(P)|\mathcal{F}_{t-1}\right), \quad J_t^{(k)}(P) = \sum_{\tau=1}^{t} j_\tau^{(k)}(P).$$

It is noted that $R_t^{(k)}(P)$ is the logarithm of the partial likelihood ratio. Conditioned on $\mathcal{F}_{t-1}$, the discriminatory information between P_0 and P contained in $y_t^{(k)}$ is obtained as $i_t^{(k)}(P)$; therefore, the sum $I_t^{(k)}(P)$ is the accumulated Kullback–Leibler information [19]. Based on these concepts, Assumption 5 is made.

Assumption 5: Three regular conditions of the assumption are stated in the following.

1) $p(\delta_1^{(k)}(P) < I_{T_k}^{(k)}(P)/\gamma_{T_k}^{(k)} < \delta_2^{(k)}(P)) \to 1$ uniformly on any compact subset of $\mathcal{P}\backslash P_0$ for any k, where $\delta_1^{(k)}(P)$ and $\delta_2^{(k)}(P)$ are positive continuous functions of P, and $\gamma_{T_k}^{(k)} \to \infty$.
2) $J_{T_k}^{(k)}(P)/\gamma_{T_k}^{(k)} \xrightarrow{p} 0$ uniformly on any compact subset of $\mathcal{P} \setminus P_0$ for any k.
3) Let $B_\epsilon = \{P \in \mathcal{P} : ||P - P_0|| < \epsilon\}$, $U_\epsilon = \{P \in \mathcal{P} : \delta_2^{(k)}(P) < \epsilon \text{ for any } k\}$, and $L_\epsilon = \{P \in \mathcal{P} : \delta_1^{(k)}(P) < \epsilon \text{ for any } k\}$. Then for any $\epsilon > 0$, there exits $0 < \eta_2 < \eta_1$ such that $U_{\eta_2} \subset L_{\eta_1} \subset B_\epsilon$.

Remark 6: Condition 1 in Assumption 5 requires that the accumulated Kullback–Leibler information tends to infinity when $P \neq P_0$, and the divergence rate can be estimated. Condition 2 in Assumption 5 implies that the conditional variance does not grow too fast; and Condition 3 determines a relationship between the information concepts and the topology of $\mathcal{P}$. All three conditions in Assumption 5 are satisfied if $\{y_t^{(k)}\}_{t=1}^{T_k}$ are generated by an ergodic Markov chain for each k. Further details on verification of these conditions in various generalized linear models are reported in [18].

Theorem 7 establishes the posterior consistency of the proposed model under Assumption 5 by showing that $\Pi(\cdot|\mathcal{D}_{\boldsymbol{T}})$ concentrates in arbitrarily small neighborhoods of P_0 as $\boldsymbol{T} \to \infty$.

Theorem 7: If the true data-generating process p_0 satisfies (1) and all three conditions in Assumption 5, then

$$\forall \epsilon > 0, \quad \Pi(B_\epsilon^c|\mathcal{D}_{\boldsymbol{T}}) \xrightarrow{p} 0 \text{ with respect to } p_0 \text{ as } \boldsymbol{T} \to \infty.$$

Proof: Based on the first two conditions in Assumption 5 and Theorem 4, it follows that:

$$\left|\sum_{k=1}^{K} R_{T_k}^{(k)}(P) - \sum_{k=1}^{K} I_{T_k}^{(k)}(P)\right| \xrightarrow{p} 0$$

uniformly on any compact subset of $\mathcal{P} \setminus P_0$. Because of the third condition in Assumption 5, one may choose $0 < \eta_2 < \eta_1$ such that $U_{\eta_2} \subset L_{\eta_1} \subset B_\epsilon$ for any $\epsilon > 0$. Then

$$\begin{aligned}
\Pi(B_\epsilon^c|\mathcal{D}_T) &= \frac{\int_{B_\epsilon^c} \exp\left(-\sum_{k=1}^{K} R_{T_k}^{(k)}(P)\right) d\Pi(P)}{\int_{\mathcal{P}} \exp\left(-\sum_{k=1}^{K} R_{T_k}^{(k)}(P)\right) d\Pi(P)} \\
&= \frac{\int_{B_\epsilon^c} \exp\left(\eta_3\gamma_T - \sum_{k=1}^{K} R_{T_k}^{(k)}(P)\right) d\Pi(P)}{\int_{\mathcal{P}} \exp\left(\eta_3\gamma_T - \sum_{k=1}^{K} R_{T_k}^{(k)}(P)\right) d\Pi(P)} \\
&\leq \frac{\int_{L_{\eta_1}^c} \exp\left(\eta_3\gamma_T - \sum_{k=1}^{K} R_{T_k}^{(k)}(P)\right) d\Pi(P)}{\int_{U_{\eta_2}} \exp\left(\eta_3\gamma_T - \sum_{k=1}^{K} R_{T_k}^{(k)}(P)\right) d\Pi(P)} \\
&\equiv \frac{N(\mathcal{D}_T)}{D(\mathcal{D}_T)}
\end{aligned}$$

where $\eta_2 < \eta_3 < \eta_1$ and $\gamma_T \equiv \sum_{k=1}^{K} \gamma_{T_k}$.

For $N(\mathcal{D}_T)$, since $\sum_{k=1}^{K} R_{T_k}^{(k)}(P) \xrightarrow{p} \sum_{k=1}^{K} I_{T_k}^{(k)}(P)$ uniformly on $L_{\eta_1}^c$, and $\sum_{k=1}^{K} I_{T_k}^{(k)}(P) > \eta_1\gamma_T$ for sufficiently large $\boldsymbol{T}$, it follows that $\sum_{k=1}^{K} R_{T_k}^{(k)}(P) > \eta_1\gamma_T$ as $\boldsymbol{T} \to \infty$. Thus, the integrand of $N(\mathcal{D}_T)$ is less than $\exp(\eta_3\gamma_T - \eta_1\gamma_T) \to 0$ as $\boldsymbol{T} \to \infty$. Hence, $\limsup_{\boldsymbol{T}\to\infty} N(\mathcal{D}_T) = 0$.

Similarly for $D(\mathcal{D}_T)$, because $\sum_{k=1}^{K} R_{T_k}^{(k)}(P) \xrightarrow{p} \sum_{k=1}^{K} I_{T_k}^{(k)}(P)$ pointwise on U_{η_2}, the limit inferior of the integrand of $D(\mathcal{D}_T)$ is infinity. Because the prior Π has full support on $\mathcal{P}$, $\Pi(U_{\eta_2}) > 0$. By Fatou's lemma, we obtain that $\liminf_{\boldsymbol{T}\to\infty} D(\mathcal{D}_T) = \infty$. Now, it is concluded that $\forall \epsilon > 0$ $\Pi(B_\epsilon^c|\mathcal{D}_T) \xrightarrow{p} 0$ as $\boldsymbol{T} \to \infty$. □

III. Estimation and Inference

This section develops an algorithm for posterior computation and Bayes factor analysis for hypothesis testing.

A. Posterior Computation

Although the posterior distribution has no analytic form, the inference of the associated Bayes network can be accomplished by Gibbs sampling, i.e., sampling each variable from its full conditional in turn. Since the dimension of $\boldsymbol{\omega}^{(j)}$ varies

with k_j, it is hard to construct a stationary Markov chain using the plain Gibbs sampling. A common analytical tool to infer a variable-dimension model is reversible jump Monte Carlo Markov chain [20], which performs the transdimensional exploration in the model space. The difficulties of transdimensional modeling can be circumvented by product partition model [21], [22] that allows the construction of a stationary Markov chain on the clustering configuration space. In this paper, the varying dimension $\boldsymbol{\omega}^{(j)}$ is integrated out to sample k_j directly from $p(k_j|\boldsymbol{x}_j, \boldsymbol{z}_j)$, which forms a partially collapsed Gibbs sampler [23] that alternates between the space with all variables and the space with all variables but $\boldsymbol{\omega} = \{\boldsymbol{\omega}^{(j)}\}_{j=1}^{q}$.

To compute the posteriors of the Pitman–Yor process, the infinite dimensional $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ at Lth component are truncated, as presented in [15] to achieve the desired accuracy by choosing an appropriate L, which is chosen in this paper to be 100 to satisfy the accuracy requirements. Posterior sampling of other variables is straightforward. The details are summarized in Algorithm 1, where $\boldsymbol{\xi}$ collects the variables that are not explicitly mentioned.

To execute Algorithm 1, several hyperparameters need to be chosen. The implication and determination of μ_j and L have been addressed earlier and those of other hyperparameters are discussed here. The hyperparameters a and b determine the clustering ability of the Pitman–Yor process. A grid search for $b = 0, 0.25, 0.5, 0.75$ and $a = 1, 5, 10, 100$ has been made. It turns out that $a = 1$ and $b = 0$ suffice for all applications in this paper, which renders the Pitman–Yor process be a Dirichlet process. However, for other applications where power law is important, the Pitman–Yor process may provide more flexibility. It is noted that α and β_j are hyperparameters of Dirichlet distribution and serve as pseudocounts. Their determination is dependent on the users' prior belief and often they are chosen to be small values when no additional information is available. In this paper, they are chosen to be: $\alpha = 1$ and $\beta_j = 1/C_j$ for all applications.

B. Bayesian Hypothesis Testing

This section performs hypothesis testing on the importance of a particular predictor for interpreting the underlying model in many applications, as shown in Section VII. It is also a better utilization of computational resources for sequential classification by excluding the unimportant predictors. As mentioned earlier, a particular predictor $\boldsymbol{z}_j$ is important if and only if the number of clusters $\tilde{k}_j$ formed by the corresponding latent class allocation variables $\boldsymbol{x}_j$ is greater than 1. Therefore, to perform Bayesian tests for the hypothesis described above, one only needs to compute the Bayes factor [10] in favor of $H_1 : \tilde{k}_j > 1$ against $H_0 : \tilde{k}_j = 1$ given by

$$BF_{10} = \frac{p(H_1|\boldsymbol{y}, \boldsymbol{z})/p(H_1)}{p(H_0|\boldsymbol{y}, \boldsymbol{z})/p(H_0)} \tag{18}$$

where $p(H_0|\boldsymbol{y}, \boldsymbol{z})$ and $p(H_1|\boldsymbol{y}, \boldsymbol{z})$ are equal to the proportions of samples in which the $\tilde{k}_j$ values conform to H_0 and H_1, respectively; and the prior probabilities $p(H_0)$ and $p(H_1)$ are obtained based on the following probability:

$$p(\tilde{k}_j = 1) = \sum_{k=1}^{C_j} p(k_j = k) \sum_{l=1}^{k} p(x_{j,t} = l \quad \forall\, t | k_j = k)$$
$$= \left(\prod_{r=1}^{C_j} \gamma_j^{(n_{j,c})} \right) \left(\sum_{k=1}^{C_j} \frac{p(k_j = k)k}{\prod_{k=1}^{C_j} (k\gamma_j)^{(n_{j,c})}} \right)$$

where the pertinent parameters are defined in Algorithm 1.

Algorithm 1 Gibbs Sampling for Proposed Model

Input: Dataset $\{y_t, \boldsymbol{z}_t\}_{t=1}^{T}$, hyperparameters a, b, α, $\{\mu_j\}_{j=1}^{q}$, $\{\beta_j\}_{j=1}^{q}$, truncating components L, number of samples N, and the initial sample ${}_{(0)}\boldsymbol{\phi}$, ${}_{(0)}\boldsymbol{\pi}$, ${}_{(0)}\boldsymbol{\lambda}$, ${}_{(0)}\boldsymbol{\omega}$, ${}_{(0)}\boldsymbol{x}$, ${}_{(0)}\boldsymbol{k}$

Output: Posterior samples $\{{}_{(n)}\boldsymbol{\phi}, {}_{(n)}\boldsymbol{\pi}, {}_{(n)}\boldsymbol{\lambda}, {}_{(n)}\boldsymbol{\omega}, {}_{(n)}\boldsymbol{x}, {}_{(n)}\boldsymbol{k}\}_{n=1}^{N}$

1: **for** $n = 1$ to N **do**

2: For each $(s_1, \ldots, s_q)$, sample $\phi_{s_1,\ldots,s_q}$ from its multinomial full conditionals

$$p(\phi_{s_1,\ldots,s_q} = l|\boldsymbol{\xi}) \propto \pi_l \prod_{c=1}^{C_0} \{\lambda_l(c)\}^{n_{s_1,\ldots,s_q}(c)}$$

where $n_{s_1,\ldots,s_q}(c) = \sum_1^T \mathbf{1}\{x_{1,t}=s_1,\ldots,x_{q,t}=s_q,\ y_t=c\}$.

3: For $l = 1, \ldots, L$, update π_l as follows:

$$V_l|\boldsymbol{\xi} \sim \mathbf{Beta}(1 - b + n_l, a + lb + \sum_{k>l} n_k), \quad l < L$$

$$V_L = 1, \quad \pi_l = V_l \prod_{k=1}^{l-1}(1 - V_k)$$

where $n_l = \sum_{(s_1,\ldots,s_q)} \mathbf{1}\{\phi_{s_1,\ldots,s_q} = l\}$.

4: For $l = 1, \ldots, L$, sample $\boldsymbol{\lambda}_l$ from their Dirichlet full conditionals

$$\boldsymbol{\lambda}_l|\boldsymbol{\xi} \sim \mathbf{Dir}\{\alpha + n_l(1), \ldots, \alpha + n_l(C_0)\}$$

where $n_l(c) = \sum_{(s_1,\ldots,s_q)} \mathbf{1}\{\phi_{s_1,\ldots,s_q} = l\} n_{s_1,\ldots,s_q}(c)$.

5: For $j = 1, \ldots, q$ and $c = 1, \ldots, C_j$, sample

$$\boldsymbol{\omega}^{(j)}(c)|\boldsymbol{\xi} \sim \mathbf{Dir}\{\beta_j + n_{j,c}(1), \ldots, \beta_j + n_{j,c}(k_j)\}$$

where $n_{j,c}(s_j) = \sum_{t=1}^{T} \mathbf{1}\{x_{j,t} = s_j, z_{j,t} = c\}$.

6: For $j = 1, \ldots, q$ and for $t = 1, \ldots, T$, sample the $x_{j,t}$ from their multinomial full conditionals

$$p(x_{j,t}=s|\boldsymbol{\xi}, x_{i,t}=s_i, i \neq j) \propto \omega_s^{(j)}(z_{j,t}) \lambda_{\phi_{s_1,\ldots,s,\ldots,s_q}}(y_t)$$

7: For $j = 1, \ldots, q$, sample k_j using their multinomial full conditionals

$$p(k_j = k|\boldsymbol{\xi}) \propto \exp(-\mu_j k) \prod_{c=1}^{C_j} n_{j,c}^{-k\beta_j}, \quad k_j = \max_t\{x_{j,t}\}, \ldots, C_j$$

where $n_{j,c} = \sum_{t=1}^{T} \mathbf{1}\{z_{j,t} = c\}$.

8: **end for**

IV. Sequential Classification

This section proposes a classification algorithm for complex dynamical systems comprised of an off-line training phase and an online testing phase. Let there be C classes of dynamical systems of interest, from each of which a training data set ${}^{(i)}\mathcal{D}_{T_i} = \{{}^{(i)}y_t, {}^{(i)}\boldsymbol{z}_t\}_{t=1}^{T_i}$ is collected. It is required that all the data are categorical (e.g., quantized from continuous data), and the number of categories of predictors and response variables are identical for each class.

In the training phase, the data set ${}^{(i)}\mathcal{D}_{T_i}$ is used to compute posterior samples $\{{}^{(i)}_{(n)}\boldsymbol{\phi}, {}^{(i)}_{(n)}\boldsymbol{\lambda}, {}^{(i)}_{(n)}\boldsymbol{\omega}\}_{n=1}^{N}$ for each class i, as described in Algorithm 1. In the testing phase, a test data set $\mathcal{D}_T$ needs to be classified to belong to one of the C classes. For this purpose, the conditional probability $p(\mathcal{D}_T|{}^{(i)}\mathcal{D}_{T_i})$ is computed by the following equations:

$$p(\mathcal{D}_T|{}^{(i)}\mathcal{D}_{T_i}) = \prod_{t=1}^{T} p(y_t|\boldsymbol{z}_t; {}^{(i)}\mathcal{D}_{T_i}) \tag{19}$$

$$p(y_t|\boldsymbol{z}_t; {}^{(i)}\mathcal{D}_{T_i}) \approx \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{s_1=1}^{k_1}\cdots\sum_{s_q=1}^{k_q} {}^{(i)}_{(n)}\lambda_{{}^{(i)}_{(n)}\phi s_1,\ldots,s_q}(y_t)\prod_{j=1}^{q} {}^{(i)}_{(n)}\omega_{s_j}^{(j)}(z_{j,t})\right). \tag{20}$$

Based on the conditional probability $p(\mathcal{D}_T|{}^{(i)}\mathcal{D}_{T_i})$, the posterior probability of the observed data $\mathcal{D}_T$ belonging to the class i is denoted as $p(C_i|\mathcal{D}_T)$ and is given as

$$p(C_i|\mathcal{D}_T) = \frac{p(\mathcal{D}_T|{}^{(i)}\mathcal{D}_{T_i})p(C_i)}{\sum_{r=1}^{C} p(\mathcal{D}_T|{}^{(r)}\mathcal{D}_{T_r})p(C_r)} \tag{21}$$

where $p(C_i)$ is the prior probability of the class i. Then, the classification decision is made by the rule

$$D_{\text{class}} = \arg\max_i p(C_i|\mathcal{D}_T). \tag{22}$$

The prior probability $p(C_i)$ can be chosen to reflect user's subjective beliefs or designed to optimize certain objective criterion. This detection algorithm is sequential, because the conditional probability $p(\mathcal{D}_T|{}^{(i)}\mathcal{D}_{T_i})$ could be evaluated sequentially, as shown in (19). For fast implementation, the values of $p(y_t|\boldsymbol{z}_t; {}^{(i)}\mathcal{D}_{T_i})$ in (20) can be precomputed and stored for different values of $(y_t, \boldsymbol{z}_t)$.

For binary classification, the likelihood ratio test [24] is constructed as follows:

$$\frac{p(\mathcal{D}_T|{}^{(1)}\mathcal{D}_{T_1})}{p(\mathcal{D}_T|{}^{(0)}\mathcal{D}_{T_0})} \underset{0}{\overset{1}{\gtrless}} \Theta \tag{23}$$

where Θ is the threshold. One criterion to choose the threshold Θ is the receiver operating characteristic (ROC). The ROC curve is obtained by varying Θ, and provides a tradeoff between the probability of detection $p_D = p(\text{decide } 1|1 \text{ is true})$ and the probability of false alarms $p_F = p(\text{decide } 1|0 \text{ is true})$. Based on the ROC curves, it is possible to select an optimal combination of p_D and test data length for a given p_F, which would lead to a choice of the threshold Θ.

TABLE I
Transition Probabilities for y_t

y_{t-1}	y_{t-2}	y_{t-5}	θ_t	$p(y_t=1)$	$p(y_t=0)$
0	0	0	0	0.5	0.5
0	0	1	0	0.5	0.5
0	1	0	0	0.5	0.5
0	1	1	0	0.5	0.5
1	0	0	0	0.5	0.5
1	0	1	0	0.5	0.5
1	1	0	0	0.5	0.5
1	1	1	0	0.5	0.5
0	0	0	1	0.4	0.6
0	0	1	1	0.3	0.7
0	1	0	1	0.9	0.1
0	1	1	1	0.25	0.75
1	0	0	1	0.15	0.85
1	0	1	1	0.2	0.8
1	1	0	1	0.3	0.7
1	1	1	1	0.15	0.85

V. Numerical Example

This section evaluates finite-sample performance of the proposed method with simulated data generated from a nonstationary Markov model, which is a special case of panel data. The underlying Markov model for data generation is known and is used as the ground truth for performance evaluation. The data generation process is described next.

In this numerical experiment, sequences of binary symbols y_t are generated from a nonstationary Markov model $p(y_t|\mathcal{F}_{t-1}) = p(y_t|y_{t-1}, y_{t-2}, y_{t-5}, \theta_t)$, where only the time lags $y_{t-1}, y_{t-2}, y_{t-5}$ and the exogenous variable θ_t are important predictors. The exogenous variable θ_t itself is binary and is generated by a first-order Markov chain with the following transition probability matrix:

$$\begin{bmatrix} 0.75 & 0.25 \\ 0.08 & 0.92 \end{bmatrix}.$$

The transition probabilities for y_t are listed in Table I. In this table, whenever $\theta_t = 0$, the probability of $y_t = 0$ is 0.5, which corresponds to a white noise.

To estimate $p(y_t|\mathcal{F}_{t-1})$, 506 samples of $\{y_t\}_{t=1}^{506}$ and 500 samples of $\{\theta_t\}_{t=7}^{506}$ are collected, respectively. Based on the assertion that y_{t-D} is not important for predicting y_t if D is greater than 6, the predictors are set for y_t as $\boldsymbol{z}_t \equiv (y_{t-1}, y_{t-2}, \ldots, y_{t-6}, \theta_t)$. To compute posteriors using Algorithm 1, $j/2$ is assigned to μ_j for $j = 1, \ldots, 6$, and 0 is assigned to μ_7.

Upon performing 70 000 iterations of Gibbs sampling (with initial 20 000 samples discarded as a burn-in period), the remaining 50 000 after-burn-in samples are downsampled by taking every fifth sample to reduce the autocorrelation. Fig. 2 summarizes the Gibbs sampling results based on the downsampled after-burn-in samples. Fig. 2(a) exhibits the log likelihood of the model for 10 000 iterations. Fig. 2(b) shows the ability of the proposed method to identify the important predictors. In this case, the important predictors are $1, 2, 5, 7$, and the resulting prediction coincides with the ground truth $y_{t-1}, y_{t-2}, y_{t-5}$ and θ_t. Fig. 2(c) shows relative frequency of how many predictors are important. The proposed method also leads to parsimonious representations,

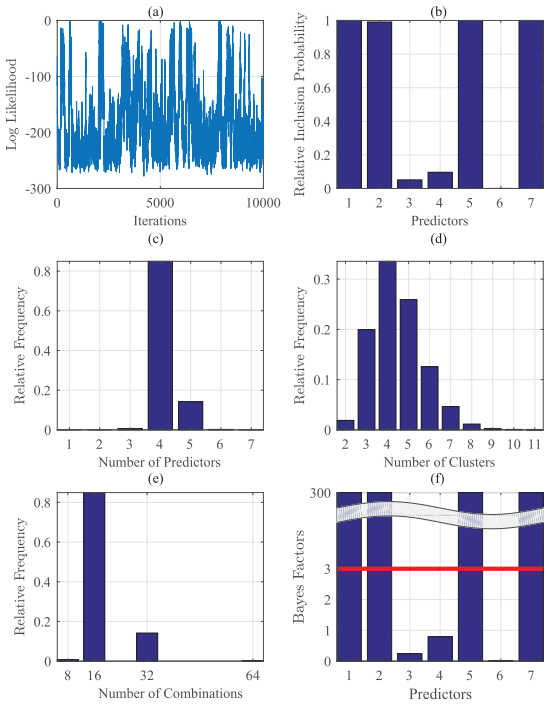


Fig. 2. Gibbs sampling results for the numerical example. The horizontal (solid red line) indicates the threshold = 3. (a) Log likelihood of model. (b) Inclusion proportions of different predictors. (c) Relative frequency distribution of the number of important predictors. (d) Relative frequency distribution of the number of clusters of the tensor $\tilde{\lambda}$. (e) Relative frequency distribution of possible combinations of $(s_1, \ldots, s_q)$. (f) Bayes factors of different predictors.

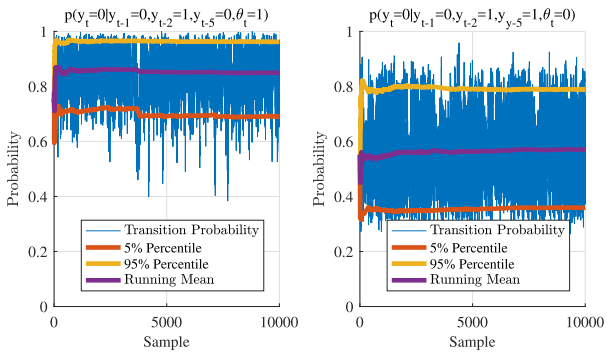


Fig. 3. Transition probabilities for the numerical example.

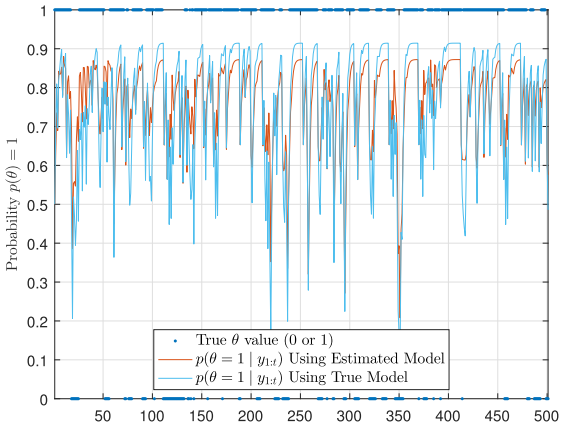


Fig. 4. $p(\theta_t|y_{1:t})$ using estimated and true models.

as shown in Fig. 2(d) and (e). As discussed in Section II-C, the tensor $\lambda_{s_1 \ldots s_q}(y_t)$ has much more components than required but they can be clustered nonparametrically to significantly reduce the needed number of components. Fig. 2(f) shows the Bayes factors as computed in Section III-B for different predictors. Bayes factor BF_{10} in (18) can be interpreted as the evidence against H_0, and a threshold of $t = 3$ indicates that this evidence is positive [25], which is used in the sequel. Furthermore, $BF_{10} > 150$ indicates very strong evidence against H_0 [25]. If the inclusion proportions in Fig. 2(b) equal to 1, then the corresponding Bayes factors in Fig. 2(f) tend to infinity, as it is the case for the predictors $1, 5, 7$. For the predictor 2, the inclusion proportion is slightly smaller than 1, thus the corresponding Bayes factor does not go to infinity although it is large (>300).

In addition to correctly identify the model structure, the proposed method estimates the transition probabilities. Fig. 3 shows two cases from Table I. When $y_{t-1} = 0$, $y_{t-2} = 1$, $y_{t-5} = 0$, and $\theta_t = 1$, it follows from Table I that the true transition probability of $y_t = 0$ is 0.9. Similarly, when $y_{t-1} = 0$, $y_{t-2} = 1$, $y_{t-5} = 1$, and $\theta_t = 0$, it also follows from Table I that the true transition probability of $y_t = 0$ is 0.5. Fig. 3 exhibits the profiles of corresponding transition probabilities, where the estimated transition probability per sample is obtained with the running mean as well as the 95% and 5% percentiles. It is seen in Fig. 3 that the running mean is close to the true transition probability, and the proposed method performs better than maximum likelihood estimation under the finite-sample settings. For example, the maximum likelihood estimation of transition probability $p(y_t = 0|y_{t-1} = 0, y_{t-2} = 1, y_{t-5} = 1, \theta_t = 0)$ is 0, as the state $y_{t-1} = 0$, $y_{t-2} = 1$, $y_{t-5} = 1, \theta_t = 0$ has not even been visited even once in the training data set. On the contrary, with limited data, the proposed method not only estimates the transition probability but also yields the uncertainty quantification in terms of the quantiles.

Robustness of the proposed method is demonstrated by comparing the estimated model with the true underlying model in a hidden state estimation task. In this scenario, a sequence of $\{y_t\}$ is generated from the true model and the task is to estimate the unobserved exogenous variable θ_t. This is accomplished by a recursive Bayes filter (RBF) to obtain $p(\theta_t|y_{1:t})$ for the estimated and true models, where the prediction and update equations of the RBF are presented as

$$p(\theta_{t+1}|y_{1:t}) = \sum_{\theta_t} p(\theta_{t+1}|\theta_t)\, p(\theta_t|y_{1:t})$$

$$p(\theta_{t+1}|y_{1:t+1}) = \frac{p(y_{t+1}|y_{1:6}, \theta_{t+1})\, p(\theta_{t+1}|y_{1:t})}{\sum_{\theta_{t+1}} p(y_{t+1}|y_{1:6}, \theta_{t+1})\, p(\theta_{t+1}|y_{1:t})}.$$

The single plate in Fig. 4 compares the profiles of probabilities $p(\theta_t = 1|y_{1:t})$ of the estimated model and the true model, respectively, for 500 samples. It is observed that the estimated model behaves very similar to the true model after the initial transients are over. To classify θ_t, let us consider a decision rule such that θ_t is classified to be 1 if the $p(\theta_t = 1|y_{1:t}) > t$ for a given threshold t, where an optimal

TABLE II
RESULTS FOR RBF

	Optimal Threshold	Error Rate
Estimated Model	0.5791	23.95%
True Model	0.5320	22.75%

threshold can be identified based on the misclassification rate. The optimal threshold and minimum error rate for both the estimated model and true model are listed in Table II. When using the true model, the error rate is 22.75%, while the error rate is 23.95% when the estimated model is used. Hence, the proposed method performs almost as good as the true model.

VI. VALIDATION ON A PUBLIC DATA SET

This section presents test results of the nonparametric regression model on (publicly available) German Health Care Usage Data from NYU Panel Data Sets.[1] This is a data set of 7293 individuals for a varying number of observations for each individual over a period of one to seven years. The raw data have been preprocessed before their usage for testing the proposed nonparametric model. The testing procedure is briefly described in the following.

The categorical response variable y_t is assigned to be the *NEWHSAT* variable in the file, which represents the health satisfaction value with coding error corrected. It ranges from 0 to 10, thus being a categorical variable with 11 possible values. A total of ten predictors $\boldsymbol{z} \equiv (z_1, \ldots, z_{10})$ are defined as: $z_1 \equiv y_{t-1}$, which is the first-order time lag of the responsible variable y; z_2 is the *FEMALE* variable in the original data set, with 0 representing male and 1 representing female; z_3 is the *MARRIED* variable in the original data set, with 1 representing married and 0 otherwise; z_4 is the discretized *AGE* variable, with $z_4 = i$ if $i \leq (AGE - 25)/5 < i + 1$ and it ranges from 0 to 7; z_5 is the discretized *HANDPER* variable which represents the degree of handicap, with $z_5 = i$ if $i - 1 < HANDPER/10 \leq i$ and it ranges from 0 to 10; z_6 corresponds to the *HHNINC* variable in the original data set, which is the household nominal monthly net income in German marks divided by 10 000, and z_6 is round to $HHNINC \times 10$ with a maximum value of 5; z_7 is a categorical variable representing the degree in the original data set, with $z_7 = 5$ if $UNIV = 1$, $z_7 = 4$ if $ABITUR = 1$, $z_7 = 3$ if $FACHHS = 1$, $z_7 = 2$ if $REALS = 1$, $z_7 = 1$ if $HAUPTS = 1$, and $z_7 = 0$ otherwise; z_8 is a categorical variable representing the employment status in the original data set, with $z_8 = 4$ if $BEAMT = 1$, $z_8 = 3$ if $SELF = 1$, $z_8 = 2$ if $WHITEC = 1$, $z_8 = 1$ if $BLUEC = 1$, and $z_8 = 0$ otherwise (unemployed); z_9 corresponds to the sum of the doctor visits and hospital visits in the original data set, with $z_9 = i$ if $i - 1 < (DOCVIS + HOSPVIS)/10 \leq i$; and z_{10} corresponds to the insurance status in the original data set, with $z_{10} = 2$ if $PUBLIC = 1$, $z_{10} = 1$ if $ADDON = 1$, and $z_{10} = 0$ otherwise (uninsured).

After data preprocessing, 2000 samples are randomly chosen from the data set with 1500 of them as training set and the remaining 500 as the test set. This procedure is repeated five times and the average test results are recorded. To compute posteriors using Algorithm 1, the hyperparameters are set as: $\mu_j = 1$ for $j = 1, \ldots, 10$.

[1] http://people.stern.nyu.edu/wgreene/Econometrics/PanelDataSets.htm

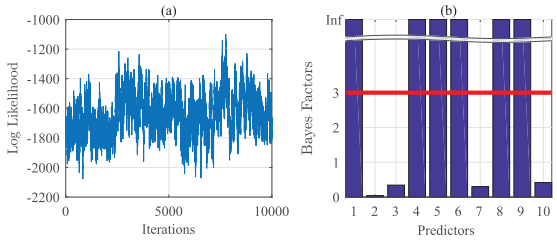


Fig. 5. Gibbs sampling results for German Health Care data. (a) Log likelihood of model. (b) Bayes factors of predictors.

TABLE III
RESULTS FOR DIFFERENT MACHINE LEARNING METHODS

	Accuracy	R^2 score
Nonparametric model	28.20%	0.3774
Logistic Regression	25.00%	0.1287
KNN ($K = 5$)	22.60%	0.2461
SVM	25.00%	0.3479
Decision Tree	24.80%	0.3382
Random Forest	25.80%	0.3481
AdaBoost	26.00%	0.3554
MLPNN	25.40%	0.3256

Gibbs sampling has been performed for 1 50 000 iterations, where the initial 1 00 000 data points are discarded as burn-in period and the after burn-in samples are downsampled by taking every fifth data point to reduce the autocorrelation. Fig. 5 shows the results of Gibbs sampling for this data set. Specifically, important predictors are: z_1 which represents y_{t-1}; z_4 which represents age; z_5 which represents degree of handicap; z_6 which represents household income; z_8 which represents employment status; and z_9 which represents the sum of hospital visits and doctor visits.

Table III presents the classification results of the test set using the proposed nonparametric model versus various standard machine learning methods [26]. Two different kinds of metrics are used to evaluate the results using the predictors as features, i.e., accuracy and R^2 score. Since the response variable is categorical ranging from 0 to 10, it can be evaluated either as a classification problem with accuracy being a metric, or as a regression problem with R^2 score being a metric. In the proposed method, the maximum *a posteriori* probability estimate is used as the predicted outcome when calculating the accuracy while the posterior mean is used when calculating R^2 score. For the K-nearest-neighbor (KNN) algorithm implementation, a value of $K = 5$ is used. For the SVM algorithm, a grid search of parameters is performed and the best results are recorded as a table using these parameters: kernel of radial basis function with parameters $\gamma = 0.001$ and $C = 10$. For random forest and AdaBoost algorithms, ten decision trees are used. For multilayer perceptron neural network (MLPNN), two hidden layers each with five neurons are used.

Upon completion of these comprehensive tests, it is concluded that the proposed nonparametric model outperforms the existing regression methods in terms of both metrics: accuracy and R^2 score. Moreover, the proposed method provides a more explainable model than most black box machine learning algorithms by explicit identification of significant predictors.

VII. Experimental Validation on a Combustor

This section presents the results of experimental validation of the nonparametric regression model on a swirl-stabilized lean-premixed laboratory-scale combustor apparatus that is described in detail in a recent publication [13]. As for the procedure for detection of combustion instability, current methods often use an empirical threshold instead of taking advantage of statistical detection theory (e.g., sequential testing). This paper accommodates the usage of both temporal and spatial variables in the proposed model and makes sequential classification of combustion instabilities, as described in Section IV.

The quantized pressure measurement is chosen as a response variable, denoted as y_t at time instant t. The continuous time series data of pressure oscillations are quantized via maximum entropy partitioning [27] with a ternary alphabet $\Sigma = \{1, 2, 3\}$ for both stable and unstable pressure oscillations.

A. Training Phase

This section describes training of the nonparametric regression model, where 50 samples are used from the downsampled ensemble of quantized pressure time series under each of the 62 operating conditions, of which 34 cases are stable and the remaining 28 cases are unstable. Based on the experimental observations, it is hypothesized that the relevant memory of y_t is limited to D, i.e., y_t is not dependent on y_{t-D} and older data. In the experimental data, it is observed that the memory D is limited to 8 for both stable and unstable time-series data. Hence, the predictors are set for y_t as $z_t \equiv (y_{t-1}, y_{t-2}, \ldots, y_{t-8}, \psi_1, \psi_2)$, where ψ_1 and ψ_2, respectively, denote the categories of fuel equivalence ratio and the percentage of pilot fuel. The different time lags of pressure measurements represent the temporal part of combustion process, while the external variables (e.g., ψ_1 and ψ_2) model the spatial part, which form a spatiotemporal model of combustion data. Since y_t has three categories, it follows that $C_0 = C_1 = \cdots = C_8 = 3$. Similarly, ψ_1 can take four different values and ψ_2 can take 19 values, which implies $C_9 = 4$ and $C_{10} = 19$, respectively. To compute posteriors, we assign $j/2$ to μ_j for $j = 1, \ldots, 8$ and 0 to μ_j for $j = 9, 10$.

Gibbs sampling has been performed for 50 000 iterations, where the initial 30 000 data points are discarded as burn-in period and the after burn-in samples are downsampled by taking every fifth data point to reduce the autocorrelation. Figs. 6 and 7 summarize the Gibbs sampling results for stable and unstable cases, respectively. Figs. 6(a) and 7(a) show the log likelihood of the model with different iterations. Figs. 6(b) and 7(b) show the proposed method's ability to identify the important predictors of the regression model. Figs. 6(c) and 7(c) show relative frequency of how many predictors are important. The proposed method also leads to parsimonious representations as shown in

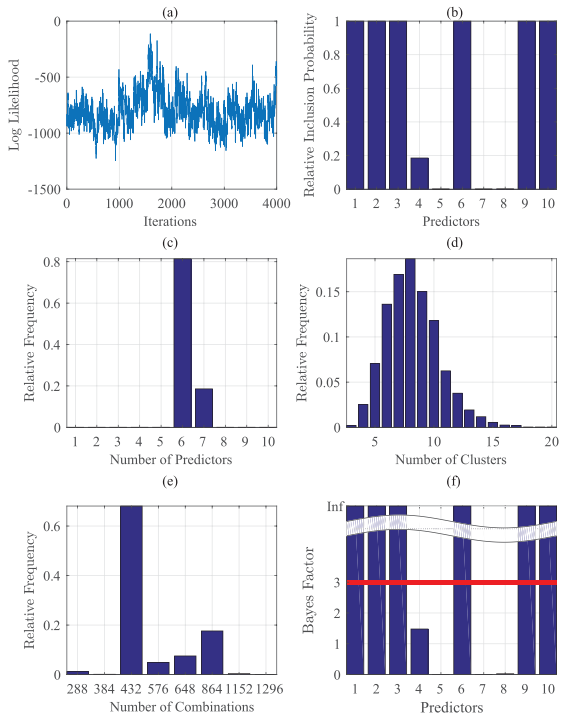


Fig. 6. Gibbs sampling results for stable pressure data. (a) Log likelihood of model. (b) Inclusion proportions of different predictors. (c) Relative frequency distribution of the number of important predictors. (d) Relative frequency distribution of the number of clusters of the tensor $\tilde{\lambda}$. (e) Relative frequency distribution of possible combinations of $(s_1, \ldots, s_q)$. (f) Bayes factors of different predictors.

Figs. 6(d) and (e) and 7(d) and (e). By comparing Figs. 6(b) and 7(b), we find that the maximal order for unstable case is 8 and is 6 for the stable case, which indicates a more deterministic behavior of unstable pressure time series and is line with previous work [13]. Another interesting observation is that the predictor 9 (equivalence ratio) is only important under the stable operating condition but not under the unstable operating condition, while predictor 10 (pilot percentage) is important in both cases. Figs. 6(f) and 7(f) show the Bayes factors of the different predictors of the stable and unstable case. The analysis of these figures is analogous to those in Section V.

B. Sequential Classification

To evaluate the performance of sequential classification for combustion instability, two instances of 50 samples (different from training samples) are selected from downsampled quantized pressure measurements under each operating condition (124 test cases in total for 62 operating conditions). Fig. 8 shows the posterior probability of each class as a function of the length of the observed data. It is seen in Fig. 8(a) that the observed sequence is correctly classified as stable because the posterior probability of the class 0 approaches one, while class 1 approaches zero very fast. Similarly in Fig. 8(b), the observed sequence is correctly classified as unstable. Fig. 9 shows a family of ROC curves for the proposed detection algorithm with varying lengths of test data. It is observed that the ROC curve improves (i.e., moves toward the top-left corner) considerably as the test data length is increased

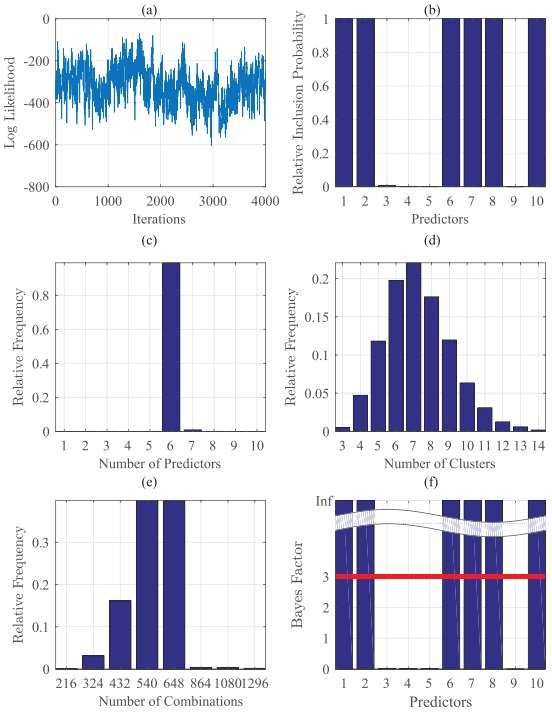


Fig. 7. Gibbs sampling results for unstable pressure data. (a) Log likelihood of model. (b) Inclusion proportions of different predictors. (c) Relative frequency distribution of the number of important predictors. (d) Relative frequency distribution of the number of clusters of the tensor $\tilde{\lambda}$. (e) Relative frequency distribution of possible combinations of $(s_1, \ldots, s_q)$. (f) Bayes factors of different predictors.

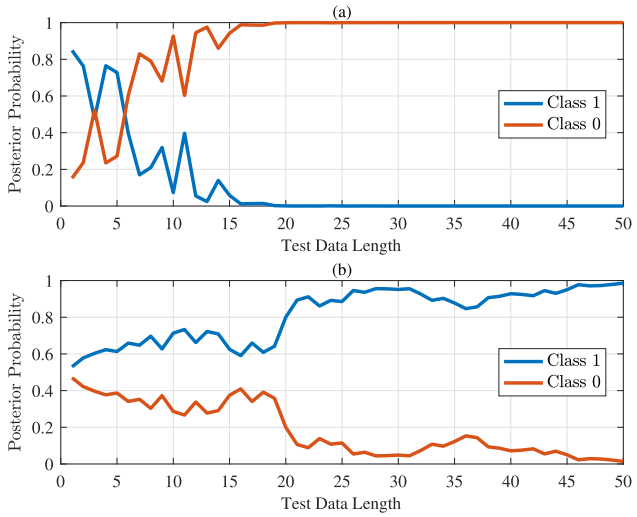


Fig. 8. Posterior probabilities for stable and unstable cases. (Class 0 represents stability and Class 1 represents instability.) (a) Posterior probabilities for a test sequence from stable conditions. (b) Posterior probabilities for a test sequence from unstable conditions.

from 10 to 50. This is reasonable since the longer the test data length, the more information is there and thus better results are expected.

The results from the proposed nonparametric Bayes method are compared with those reported in the open literature, specifically, the symbolic dynamics-based adaptive pattern classification [28]–[30] (which is called an adaptive method in this paper) and with a Markov chain method [9]. In the adaptive method, each row of the transition probability matrix

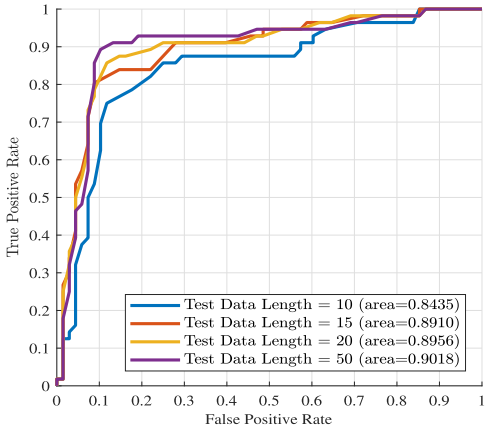


Fig. 9. ROC curves for the proposed method.

of the Markov chain is independently assigned a Dirichlet distribution as priors. The adaptive method cannot accommodate spatial variables, such as equivalence ratio and the percentage of fuel; therefore a bank of Markov chains has to be estimated for both stable, denoted as $\{\mathcal{M}_i^{(s)}\}_{i=1}^{N_s}$, and unstable, denoted as $\{\mathcal{M}_i^{(u)}\}_{i=1}^{N_u}$, cases. Given a quantized pressure sequence $y_{1:t}$, the adaptive method classifies it to be unstable if $\max_i p(y_{1:t}|\mathcal{M}_i^{(u)}) > \max_i p(y_{1:t}|\mathcal{M}_i^{(s)})$ and vice versa. It is noted that average instead of maximum can also be used, but they generate similar results.

To train the adaptive method, 1000 samples are selected from downsampled quantized pressure time series under each of 62 operating conditions, which are ~ 20 times of the training data used for the proposed nonparametric Bayes method. Two hyperparameters, namely, Markov chain order D_u for the unstable case and Markov chain order D_s for the stable case, need to be chosen before training. However, there is no coherent ex ante method to choose them; therefore an ex post method is employed by comparing the ROC curves generated under different combinations of D_u and D_s to select the best hyperparameters. In contrast, the proposed nonparametric Bayes method automatically chooses the proper orders, which saves a lot of efforts in the model selection step and is universally applicable. The best D_u and D_s for the adaptive method are 8 and 6, respectively, which coincide with the maximal order identified by the proposed nonparametric Bayes method. The results from four of these combinations are shown in Fig. 10 as examples.

To train the Markov chain models, 500 samples are selected from downsampled quantized pressure time series under one stable and one unstable condition, respectively. The orders for both are set to be 8 and other hyperparameters are set similarly as in the proposed method. Then, the two Markov chain models are used for sequential classification of the same 124 test cases. These models do not take the operating conditions (i.e., effects of exogenous variables) into consideration.

The results of the proposed method are compared with the best adaptive method [$D_u = 8$ and $D_s = 6$, as shown in Fig. 10(c)], and with Markov chain models. The ROC curves for these methods are shown in Fig. 11. It is seen that, when testing with the same data set, even though the training samples are only small fractions of those used in the

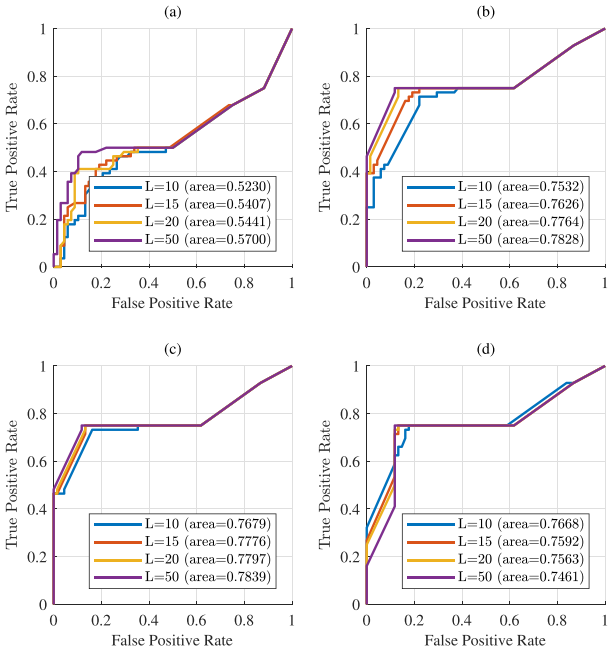


Fig. 10. ROC curves for classification using the symbolc dynamics-based adaptive method, with different D_u and D_s values. (a) $D_u = 1$ and $D_s = 1$. (b) $D_u = 4$ and $D_s = 3$. (c) $D_u = 8$ and $D_s = 6$. (d) $D_u = 9$ and $D_s = 9$.

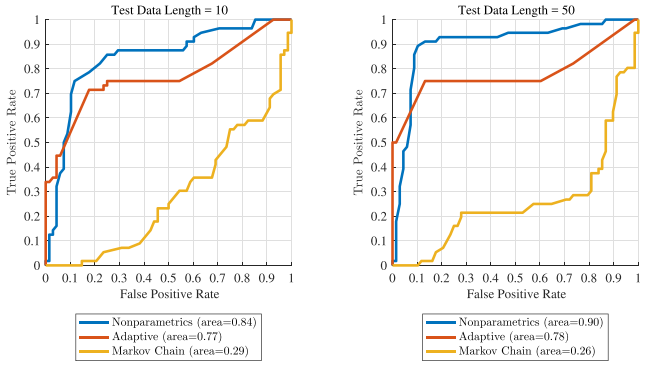


Fig. 11. ROC curves for performance comparison.

adaptive method, the nonparametric Bayes method not only yields superior classification performance but also saves the efforts for model selection than the adaptive method. The Markov chain models yielded significantly deteriorated performance as seen in Fig. 11, because they do not have information about different operation conditions.

VIII. Conclusion and Future Work

The proposed Bayesian nonparametric Bayes method provides a flexible and parsimonious model of categorical panel data, which yields superior finite-sample performance with guaranteed large-sample properties. It is demonstrated that the proposed method outperforms: 1) various standard machine learning methods, such as KNN, SVM, and MLPNN [26] for econometric public tata and 2) symbolic dynamics-based adaptive method of pattern classification [29], [30] and Markov chain method [9] for a real-time application with experimental data from a swirl-stabilized combustor apparatus.

While there are many directions in which the proposed nonparametric Bayes method can be extended, the authors suggest the following two topics for future research.

1) Development of a variational inference algorithm for the proposed nonparametric Bayes method.
2) Characterization of the underlying assumptions in the proposed method to achieve asymptotic properties.
3) Testing of the proposed method on n-gram models in natural language processing to further examine flexibility of the Pitman–Yor process [31].

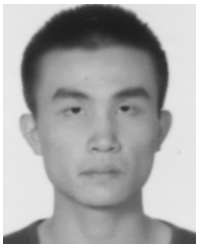
Acknowledgment

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.

References

[1] D. A. Kenny, *Cross-Lagged Panel Design*. Hoboken, NJ, USA: Wiley, 2005.
[2] P. D. Allison, *Fixed Effects Regression Models*, vol. 160. Thousand Oaks, CA, USA: SAGE, 2009.
[3] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.
[4] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Mar. 2001.
[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[6] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, 2006.
[7] M. DeYoreo and A. Kottas, "Bayesian nonparametric modeling for multivariate ordinal regression," *J. Comput. Graph. Stat.*, vol. 27, no. 6, pp. 1525–1538, Nov. 2017.
[8] Y. Yang and D. B. Dunson, "Bayesian conditional tensor factorizations for high-dimensional classification," *J. Amer. Stat. Assoc.*, vol. 111, no. 514, pp. 656–669, 2016.
[9] A. Sarkar and D. B. Dunson, "Bayesian nonparametric modeling of higher order Markov chains," *J. Amer. Stat. Assoc.*, vol. 111, no. 516, pp. 1791–1803, 2016.
[10] H. Akaike, "Factor analysis and AIC," *Psychometrika*, vol. 52, no. 3, pp. 317–332, 1987.
[11] P. Diaconis and D. Freedman, "On the consistency of Bayes estimates," *Ann. Stat.*, vol. 14, no. 1, pp. 1–26, 1986.
[12] F. Darema, "Dynamic data driven applications systems: New capabilities for application simulations and measurements," in *Proc. 5th Int. Conf. Comput. Sci. (ICCS)*, Atlanta, GA, USA, 2005, pp. 661–712.
[13] S. Sarkar, S. R. Chakravarthy, V. Ramanan, and A. Ray, "Dynamic data-driven prediction of instability in a swirl-stabilized combustor," *Int. J. Spray Combustion Dyn.*, vol. 8, no. 4, pp. 235–253, 2016.
[14] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Stat.*, vol. 1, no. 2, pp. 209–230, 1973.
[15] H. Ishwaran and L. F. James, "Gibbs sampling methods for stick-breaking priors," *J. Amer. Stat. Assoc.*, vol. 96, no. 453, pp. 161–173, 2011.
[16] S. Walker, "New approaches to Bayesian consistency," *Ann. Stat.*, vol. 32, no. 5, pp. 2028–2043, 2004.
[17] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.
[18] W. H. Wong, "Theory of partial likelihood," *Ann. Stat.*, vol. 14, no. 5, pp. 88–123, 1986.
[19] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.
[20] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.
[21] J. Pitman, "Exchangeable and partially exchangeable random partitions," *Probab. Theory Rel. Fields*, vol. 102, no. 2, pp. 145–158, 1995.
[22] J. W. Miller and M. T. Harrison. (Feb. 2015). "Mixture models with a prior on the number of components." [Online]. Available: https://arxiv.org/abs/1502.06241
[23] D. A. Van Dyk and T. Park, "Partially collapsed Gibbs samplers: Theory and methods," *J. Amer. Stat. Assoc.*, vol. 103, no. 482, pp. 790–796, 2008.

[24] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York, NY, USA: Springer-Verlag, 1994.
[25] R. E. Kass and A. E. Raftery, "Bayes factors," *J. Amer. Stat. Assoc.*, vol. 90, no. 430, pp. 773–795, 1995.
[26] C. Bishop, *Pattern Analysis and Machine Learning*. New York, NY, USA, Springer, 2006.
[27] V. Rajagopalan and A. Ray, "Symbolic time series analysis via wavelet-based partitioning," *Signal Process.*, vol. 86, no. 11, pp. 3309–3320, 2006.
[28] A. Ray, "Symbolic dynamic analysis of complex systems for anomaly detection," *Signal Process.*, vol. 84, no. 7, pp. 1115–1130, 2004.
[29] K. Mukherjee and A. Ray, "State splitting and merging in probabilistic finite state automata for signal representation and analysis," *Signal Process.*, vol. 104, pp. 105–119, Nov. 2014.
[30] Y. Wen, K. Mukherjee, and A. Ray, "Adaptive pattern classification for symbolic dynamic systems," *Signal Process.*, vol. 93, no. 1, pp. 252–260, 2013.
[31] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes," in *Proc. 21st Int. Conf. Comput. Linguistics 44th Annu. Meet. Assoc. Comput. Linguistics*, 2006, pp. 985–992.

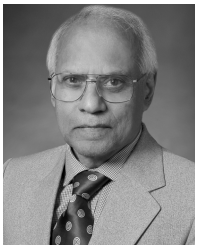


Sihan Xiong received the B.S. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2012, and the M.A. degree in mathematics from Pennsylvania State University, State College, PA, USA, in 2015, where he is currently pursuing the Ph.D. degree in mechanical engineering.

His current research interests include Bayesian nonparametrics, deep learning and reinforcement learning, and their applications to dynamical systems.

Yiwei Fu received the B.S. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014. He is currently pursuing the M.S. degree in electrical engineering, currently with the Ph.D. degree in mechanical engineering, with Pennsylvania State University, State College, PA, USA.

His current research interests include machine learning, time series analysis, and robotics.



Asok Ray (SM'83–F'02) received the graduate degrees in each discipline of electrical engineering, mathematics, and computer science, and the Ph.D. degree in mechanical engineering from Northeastern University, Boston, MA, USA.

He joined Pennsylvania State University, State College, PA, USA, in 1985, where he is currently a Distinguished Professor of mechanical engineering and mathematics, and a Graduate Faculty of electrical engineering, and nuclear engineering. He has authored or co-authored over 600 research publications, including over 300 scholarly articles in refereed journals and research monographs. He is a fellow of ASME and the World Innovative Foundation.