BAYESIAN NONPARAMETRIC MODELING OF CATEGORICAL DATA WITH APPLICATIONS TO DYNAMIC DATA-DRIVEN SYSTEMS

A Dissertation in
Mechanical Engineering
by
Sihan Xiong

© 2018 Sihan Xiong

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

May 2018
The dissertation of Sihan Xiong was reviewed and approved* by the following:

Asok Ray  
Distinguished Professor of Mechanical Engineering and Mathematics  
Dissertation Advisor, Chair of Committee

Shashi Phoha  
Professor of Electrical Engineering

Christopher D. Rahn  
Professor of Mechanical Engineering

Mark Levi  
Professor of Mathematics

Mary Frecker  
Professor of Mechanical Engineering  
Associate Department Head for Graduate Studies

*Signatures are on file in the Graduate School.
Abstract

This thesis proposes a Bayesian nonparametric model of categorical data, which can be used for casual inference of time series, information fusion of heterogeneous sources and sequential pattern classification of complex dynamic systems. The proposed method provides a flexible and parsimonious model that allows both time-independent spatial variables and time-dependent exogenous variables to be predictors. Two statistical inference algorithms have been developed for the proposed model: Gibbs sampling and stochastic variational inference. 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 proposed method has been tested by numerical simulation, validated on an econometric public dataset, and validated for detection of combustion instabilities with experimental data that have been generated in a laboratory environment.
# Table of Contents

List of Figures vii

List of Tables x

Acknowledgments xi

Chapter 1

   Introduction 1
   1.1 Motivation ................................................. 1
   1.2 Major Contributions of the Thesis ........................... 2
   1.3 Organization of the Thesis ................................. 3

Chapter 2

   Bayesian Nonparametric modeling 5
   2.1 Problem Formulation ........................................ 6
   2.2 Conditional Probability Tensor Factorization .............. 7
   2.3 Nonparametric Priors ...................................... 8
       2.3.1 Dirichlet Process .................................... 8
       2.3.2 Posterior and Predictive Distribution ............... 11
       2.3.3 Stick Breaking Representation ....................... 13
       2.3.4 Bayesian Nonparametric Mixture Model ............. 15
   2.4 Bayesian Network .......................................... 17
       2.4.1 Latent Allocation Variables .......................... 17
       2.4.2 Sparsity Inducing Priors ............................. 18
   2.5 Posterior Consistency ..................................... 20

Chapter 3

   Statistical Inference 25
3.1 Gibbs sampling .......................... 26
  3.1.1 Posterior Computation ....................... 26
  3.1.2 Bayesian Hypothesis Testing .............. 27
  3.1.3 Numerical example ....................... 27
  3.1.4 Validation on a Public Dataset .......... 34

3.2 Stochastic variational inference .......... 36
  3.2.1 Variational Inference .................... 36
  3.2.2 Stochastic Variational Inference ........ 40
  3.2.3 Stochastic Variational Inference for the Proposed Model .... 44
  3.2.4 Numerical Example ....................... 44

Chapter 4
  Causality Inference 49
    4.1 Granger Causality ....................... 50
    4.2 Causality Test .......................... 51
    4.3 Numerical Example ....................... 52
    4.4 Validation with economics Data .......... 58

Chapter 5
  Sequential Classification: Detection of Combustion Instabilities 62
    5.1 Sequential Classification .................. 64
    5.2 Experimental Data Collection ............... 65
      5.2.1 Experimental Apparatus ................ 65
      5.2.2 Data preprocessing ..................... 67
    5.3 Information Fusion ....................... 67
      5.3.1 Granger Causality ..................... 70
      5.3.2 Sequential Classification ............... 70
    5.4 Spatio-temporal Model ..................... 73

Chapter 6
  Conclusion and Research Vision 82
    6.1 Summary of Results ...................... 82
    6.2 Suggestions for Future Work ............... 83
      6.2.1 Reinforcement Learning for Active Control of Combustion Instabilities ........... 83
      6.2.2 Collaborative Filtering for Recommend Systems .......... 87

Appendix A
  Collapsed Conditional of \( k \) 90
Appendix B
  Sampling of Dirichlet Process Mixture Model  92

Bibliography  95
List of Figures

2.1 Pictorial representation of HOSVD for a rank 3 tenor $\mathbf{r}$ with core tensor $\mathbf{g}$ and weight matrices $\mathbf{v}_j$ for $j = 1, 2, 3$. ......................... 9
2.2 Empirical cumulative distribution of 10 draws form the Dirichlet Process. The base distribution is $\text{N}(0,2)$ for both cases, the concentration parameter $\alpha$ is 1 for the left panel while the concentration parameter of right panel is 10. .............................. 10
2.3 Chinese restaurant process: the integers represent customers or the allocation variable. Each table $k$ stands for a cluster, which serves one dish $\theta_k$ ................................. 13
2.4 Stick breaking construction: A stick of length 1 is recursively broke off where $\pi_k$ denotes the length of the piece after the $k^{th}$ break. .............................. 14
2.5 Graphical representation of the Dirichlet process mixture model. 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. .............................. 16
2.6 Graphical representation of the infinite mixture model. 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. .............................. 17
2.7 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. .............................. 20

3.1 Gibbs sampling results for the numerical example. The horizontal (solid red line) indicates the threshold $= 3$. ......................... 30
3.2 Transition probabilities for the numerical example. ......................... 32
3.3 $p(\theta_t | y_{1:t})$ using estimated and true models ....................... 33
6.5 Flow chart of the reinforcement learning for the active control of combustion instabilities. ........................................... 88

6.6 Graphical representation of the model for collaborative filtering. 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. ............................... 89

B.1 Graphical representation of the Dirichlet process mixture model. 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. ..................................... 92

B.2 Equivalent graphical representation of the Dirichlet process mixture model. 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. .............................. 93
List of Tables

3.1 Transition probabilities for $y_t$ ............................................. 29
3.2 Results for recursive Bayes filter .............................................. 33
3.3 Results for different machine learning methods ............................ 36
3.4 Variational distributions for the proposed model ............................ 44
3.5 Transition probabilities for $y_t$ ............................................. 46
4.1 Transition probabilities for $y_t$ ............................................. 53
4.2 Transition probabilities for $\theta_t$ ........................................... 54
4.3 Hypothesis test of Granger Causality ......................................... 56
4.4 Hypothesis test of Granger Causality ......................................... 59
5.1 Operating conditions ................................................................. 66
5.2 Hypothesis test of Granger Causality ......................................... 71
Acknowledgments

This thesis would not have been possible without my advisor Dr. Asok Ray, who kindly offered his guidance, encouragement, and mentorship for the last five years. I would also like to thank Dr. Mark Levi for the discussions and for being my adviser for M.A. degree in Mathematics. Moreover, I am deeply grateful to Dr. Christopher D. Rahn and Dr. Shashi Phoha for serving on my committee and providing valuable suggestions and time to improve the quality of my thesis in a unique way. I would like to thank Dr. Ji-Woong Lee for the mathematical training and for the help in improving my technical writing. I would like to thank Dr. Dominic Santavicca for kindly providing the experimental data on combustion that is used for validation in this thesis.

I would like to thank Mr. Jihang Li and Mr. Yue Xie for all the insightful discussions and valuable suggestions that moved forward my research. I would especially like to thank my colleague Mr. Yiwei Fu and Mr. Sudeepta Mondal for sharing vision and knowledge in our academic projects. I would like to thank Ms. Pritthi Chattopadhyay and Mr. Michael Hauser for their intelligent contributions in our joint publications. I would also like to thank Dr. Yue Li, Dr. Devesh Kumar Jha, Dr. Nurali Virani, Mr. Chandrachur Bhattacharya, Dr. Eric Keller, and all other members of Dr. Ray’s group for their professional association.

I would like to thank my parents for their love and support along the way. I am deeply indebted to them for their dedication, encouragement and inspiration without expecting anything in return.

Lastly, I would like to gratefully acknowledge the financial support I received from the U.S. Air Force Office of Scientific Research under Grant NO. FA9550-15-1-0400.
Dedication

I dedicate this dissertation to my beloved mom and dad who always believed in me.
Chapter 1

Introduction

1.1 Motivation

Intelligent systems, also known as cyber-physical systems, make decisions and actuate the physical plants based on the information and knowledge extracted from sensor data (especially time series data). A great many intelligent systems are being developed today and will grow in number and complexity over time for their efficiency and robustness. Typical examples of intelligent systems are autonomous vehicles, mobile robots, and monitoring systems of physical plants.

Data-driven models are indispensable for modern intelligent systems as physics-based models are often too complicated to derive or are infeasible to be implemented for real-time decision and control tasks. Therefore, a variety of machine learning methods are proposed to model the time series data for information extraction and decision-making, such as Gaussian processes [1], long short term memory neural networks [2], and Markov chain models [3]. Particularly, the Markov chain based symbolic time-series analysis [4] has been successfully implemented in several applications, such as battery health [5], and combustion stability [6], target detection with seismic sensors [7], fault detection in shipboard auxiliary systems [8], robot gait modeling and classification [9].

However, there are several unsolved problems in the symbolic time-series analysis, which are listed below.

1. How to capture the spatio-temporal characteristics of panel data that contain observations of multiple phenomena over multiple time periods for the same
set of units (e.g., experimental conditions), naturally arising from complex systems?

2. How to fuse information of heterogeneous sources in a coherent and interpretable framework?

3. How to identify the structure of the underlying Markov chain model based on hypothesis tests?

4. How to make causal inferences for multiple time series?

5. How to estimate the model parameters in a data efficient and real time implementable way?

6. Is there a guarantee for accuracy of the estimated parameters when the data length approaches infinity?

7. How to develop a sequential classification algorithm for streaming data in a unified Bayesian framework?

This thesis develops a novel Bayesian nonparametric model that addresses the above problems in a coherent and interpretable way [10]. The developed model is used to address several practical problems in learning and decision-making by developing theoretical formulations and solving them using concepts from Bayesian nonparametrics, probabilistic graphic model, Gibbs sampling and variational inference.

1.2 Major Contributions of the Thesis

The major contributions of this thesis are causal inference of categorical time series, information fusion of heterogeneous sources and sequential classification of complex dynamical systems are delineated as below.

- Development of a flexible model that consists of multiple time-independent spatial variables and time-dependent exogenous variables, which is capable of capturing spatio-temporal characteristics and fusing information from heterogeneous sources.
• Generation of a parsimonious model based on tensor factorization and non-parametric priors, which allows the model to borrow statistical strength from different predictors and enhances the data efficiency.

• Theoretically rigorous guarantee of posterior consistency for general (e.g., nonstationary) data-generating processes under regular conditions.

• Development of a Gibbs sampling algorithm for parameter estimation and hypothesis testing in a unified Bayesian framework, which not only provides uncertainty quantification of the estimated parameters but also identification of the underlying model structure.

• Development of a stochastic variational inference algorithm for fast inference and real time implementation.

• Extension of Granger causality to categorical time series in the setting of Bayes factor analysis, which is validated with both numerical data and real world economics data.

• Numerical and experimental validation of the proposed method in various applications, which shows superior performance compared to state of the art technologies.

1.3 Organization of the Thesis

This thesis is organized into six chapters, including the present one. The contents of each chapter are briefly given below.

1. Chapter 1 introduces and motivates various problems for Bayesian nonparametric modeling of categorical data.

2. Chapter 2 formulates a flexible and parsimonious Bayesian model of categorical data based on tensor factorization and nonparametric priors, and establishes its asymptotic properties.

3. Chapter 3 develops two statistical inference algorithms for the Bayesian nonparametric model developed in Chapter 2: Gibbs sampling and Stochastic variational inference. Gibbs sampling can handle varying dimension models
and thus provides a Bayesian approach for hypothesis tests on the significance of predictors. On the other hand, stochastic variational inference is significantly faster than Gibbs sampling in most cases, rendering it scalable to large datasets and feasible for real time implementation.

4. Chapter 4 extends the Granger causality to categorical time series based on Bayesian factor analysis, which is validated with both a numerical example and real world economics data.

5. Chapter 5 develops a sequential classification algorithm for detection of combustion instabilities that can fuse information from heterogeneous sources and incorporate varying operating conditions. The algorithm is validated with experimental data generated from a laboratory scale swirl stabilized combustor.

6. Chapter 6 summarizes the conclusions and provides several research directions for further investigation.
Chapter 2

Bayesian Nonparametric modeling

Statistical modeling of categorical panel data has evolved as a subdiscipline in pattern analysis with applications to anomaly detection, state estimation and sequential classification. 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 [11] facilitates investigation of causal relationships between two variables through regression on the lagged score of both variables, including fixed-effects model [12] and random-effects model [13]. 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 yield relatively poor performance for small-size data and hence restrict the underlying models to be of low order.

In the machine learning discipline, there exits a plethora of methods for pattern classification of high-dimension variables, such as support vector machines [14], deep neural networks [15] and extreme learning machines [16]. Even though these algorithms may achieve good accuracy in prediction tasks, it is often difficult for users to interpret the results and the uncertainty quantification is not provided. 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 [17,18]. However, these
methods are only suitable for independent and identically distributed (IID) data and may not be applicable to systematic hypothesis testing. To alleviate the above limitations, this paper proposes a Bayesian regression model of categorical panel data based on tensor factorization and nonparametric priors.

This Chapter is organized as follows. Section 2.1 describes the development of a regression model for panel data and introduces the concept of conditional probability tensor. Section 2.2 reviews tensor factorization techniques, especially the factorization of conditional probability tensors. Section 2.3 provides the background of the Dirichlet process and Bayesian nonparametric mixture model. In Section 2.4, a Bayesian network for estimation of conditional probability tensors is developed based on tensor factorization and Bayesian nonparametric priors. Finally, the posterior consistency of the proposed model is established in section 2.5.

2.1 Problem Formulation

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

\[
p(y_t^{(k)} \mid \mathcal{F}_{t-1}) = p(y_t^{(k)} \mid y_{t-1}^{(k)}, \ldots, y_{t-D}^{(k)}, \theta_t^{(k)}, \psi^{(k)})
\]  

(2.1)

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 \( \{\theta_t^{(k)}\}_{t=1}^{T_k} \). If there is no spatial variable, the proposed model reduces to a time series model that can be used to represent the non-stationary 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 \( \psi^{(k)} \).

**Remark 2.1.1.** Given \( \theta_t^{(k)}, \psi^{(k)} \), 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 than or equal to \( D \)) is the minimum order beyond which the lags are not important.

For simplicity of notations, predictors \( z_t \equiv (z_{1,t}, \ldots, z_{q,t}) \) are substituted for \( (y_{t-1}, \ldots, y_{t-D}, \theta_t, \psi) \), 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 \( j = 1, \ldots, q \). Since the response variables \( y \) and their respective time-lags have the same number of categories, it follows that \( C_0 = C_1 = \cdots = C_D \). The quantity \( p(y_t | z_t) \) is treated as a \((q + 1)\)th rank tensor in the \( C_0 \times C_1 \cdots \times C_q \) dimensional space, called the conditional probability tensor. The objective is to build a Bayesian model to estimate the conditional probability tensor in an efficient and statistically interpretable way, which is based on two techniques: tensor factorization (Section 2.2) and nonparametric priors (Section 2.3).

### 2.2 Conditional Probability Tensor Factorization

Tensor factorization has been studied extensively by many researchers over the last five decades for the purpose of approximation and information compression. The two most popular approaches are parallel factor analysis (PARAFAC) and higher order singular value decomposition (HOSVD). PARAFAC factorizes a \( C_1 \times \cdots \times C_q \) tensor \( r = \{r_{s_1,\ldots,s_q}\} \) into \( q + 1 \) low rank tensors as

\[
r_{s_1,\ldots,s_q} = \sum_{h=1}^{k} g_h \prod_{j=1}^{q} v_{h}^{(j)}(s_j) \tag{2.2}
\]

where \( g \) is rank one tensor and \( v^{(j)} \) are component specific weights.

In contrast, HOSVD generalizes PARAFAC by decomposing the tensor \( r \) as

\[
r_{s_1,\ldots,s_q} = \sum_{h_1=1}^{k_1} \cdots \sum_{h_q=1}^{k_q} g_{h_1,\ldots,h_q} \prod_{j=1}^{q} v_{h_j}^{(j)}(s_j) \tag{2.3}
\]
where \( g \), called a core tensor, is of rank \( q \) and captures interactions between different components. Fig. 2.1 shows a pictorial representation of HOSVD for a rank 3 tensor. HOSVD achieves better data compression and requires fewer components as compared to PARAFAC, which can be obtained as a special case of HOSVD with \( g \) being diagonal.

However, both PARAFAC and HOSVD are not suitable for modeling conditional probability tensors which require each entry is non-negative and the sum of each row is one. In 2015, Yang and Dunson [18] proposed the following HOSVD-type factorization for conditional probability tensors

\[
p(y_t | z_t) = \sum_{s_1=1}^{k_1} \cdots \sum_{s_q=1}^{k_q} \lambda_{s_1,\ldots,s_q}(y_t) \prod_{j=1}^{q} \omega_{s_j}^{(j)}(z_{j,t})
\]  \tag{2.4}

where \( 1 \leq k_j \leq C_j \) for \( j = 1, \ldots, q \). Moreover, the tensors \( \lambda_{s_1,\ldots,s_q}(y_t) \) and \( \omega_{s_j}^{(j)}(z_{j,t}) \) are all non-negative and satisfy the following constraints:

\[
\sum_{y_t=1}^{C_0} \lambda_{s_1,\ldots,s_q}(y_t) = 1, \quad \text{for each } (s_1, \ldots, s_q)
\]  \tag{2.5}

\[
\sum_{s_j=1}^{k_j} \omega_{s_j}^{(j)}(z_{j,t}) = 1, \quad \text{for each } (j, z_{j,t})
\]  \tag{2.6}

Because the factorization in Eq. (2.4) exists for every conditional probability tensor, the above constraints are not restrictive but it is ensured that \( \sum_{y_t=1}^{C_0} p(y_t | z_t) = 1 \).

### 2.3 Nonparametric Priors

#### 2.3.1 Dirichlet Process

In Bayesian statistics, unknown parameters are treated as random variables and are assigned prior distributions while the conditional distribution of all observable data given parameters are completely specified by the observational model [19]. The inference of a Bayesian model relies on computing the posterior distribution of parameters, which involves an inversion of the order of conditioning. If a Bayesian model contains at least one infinite dimensional parameter, then it is called a Bayesian nonparametric model [20]. Therefore, a Bayesian nonparametric model is
Figure 2.1. Pictorial representation of HOSVD for a rank 3 tenor \( r \) with core tensor \( g \) and weight matrices \( v_j \) for \( j = 1, 2, 3 \).

defined as a probability distribution (prior) on an infinite dimensional space, which is indeed a stochastic process. The Bayesian nonparametric models are increasingly replacing parametric models for the following two reasons:

- Bayesian nonparametric models can avoid the arbitrary and possibly unverifiable assumptions inherent in the parametric models.

- It can allow the complexity of model to grow with data size and thus mitigate the underfitting problem.

The Dirichlet process [21] is a stochastic process used as priors in Bayesian nonparametric models, particularly in infinite mixture models. It is a distribution of distributions, i.e., each draw from a Dirichlet process is itself a distribution. It is called a Dirichlet process because its finite dimensional marginal distributions are Dirichlet distributed, just as the Gaussian process [1], another popular stochastic process used for Bayesian nonparametric regression, has Gaussian distributed finite dimensional marginal distributions. The formal definition of the Dirichlet process is given as below.

**Definition 2.3.1.** A random distribution \( G \) on \( (\Omega, \mathcal{F}) \) is said to be Dirichlet process distributed with base distribution \( H \) and concentration parameter \( \alpha \), written...
Figure 2.2. Empirical cumulative distribution of 10 draws from the Dirichlet Process. The base distribution is $N(0, 2)$ for both cases, the concentration parameter $\alpha$ is 1 for the left panel while the concentration parameter of right panel is 10.

\[
G \sim DP(\alpha, H), \text{ if}
\]

\[
(G(A_1), \ldots, G(A_r)) \sim Dir(\alpha H(A_1), \ldots, \alpha H(A_r))
\]

(2.7)

for every finite measurable partition $A_1, \ldots, A_r$ of $\Omega$.

The base distribution $H$ and the parameter $\alpha$ in the definition of the Dirichlet process have intuitive meanings. The base distribution is basically the mean of the Dirichlet process: for any measurable set $A \subset \Omega$, we have $E[G(A)] = H(A)$. On the other hand, the concentration parameter can be understood as an inverse variance: $Var[G(A)] = H(A)(1 - H(A))/\alpha + 1$. The larger $\alpha$ is, the smaller the deviation of $G$ from $H$ is. The concentration parameter is also called the strength parameter, referring to the strength of the prior when using the DP as a nonparametric prior over distributions in a Bayesian nonparametric model, and the mass parameter, as this prior strength can be measured in units of sample size (or mass) of observations. Fig. 2.2 compares the empirical cumulative distributions drawn from two Dirichlet processes with different concentration parameters. The base distribution for both Dirichlet processes is Normal with 0 mean and 2 variance.

The existence of the Dirichlet process can be established by a number of approaches [22]. One approach is to check that Eq. 2.7 are consistent with each others, and invoke Kolmogorov’s consistency theorem to show that a distribution over functions from $F$ to $[0, 1]$ exists satisfying Eq. 2.7 for all finite measurable partitions.
2.3.2 Posterior and Predictive Distribution

Let $G \sim DP(\alpha, H)$. Since $G$ itself is a distribution, we can in turn draw samples from it. Let $\theta_1, \ldots, \theta_n$ be a sequence of independence draws from $G$. Because $G$ is a distribution over $\Omega$, the $\theta_i$'s take values in $\Omega$. We are interested in the posterior distribution of $G$ given observed values of $\theta_1, \ldots, \theta_n$. Let $A_1, \ldots, A_r$ be a finite measurable partition of $\Omega$, and let $n_k$ denotes the number of observed values in $A_k$. By Eq. 2.7 and the conjugacy between Dirichlet and the multinomial distribution, having

$$(G(A_1), \ldots, G(A_r)) | \theta_1, \ldots, \theta_n \sim \text{Dir}(\alpha H(A_1) + n_1, \ldots, \alpha H(A_r) + n_r)$$  \hspace{1cm} (2.8)

Since the above is true for all finite measurable partitions, the posterior distribution over $G$ must be a Dirichlet process as well. Because $n_k = \sum_i^n \delta_{\theta_i}(A_k)$ where $\delta_{\theta_i}$ is a point mass located at $\theta_i$, it follows that the posterior DP has updated concentration parameter $\alpha + n$ and base distribution $\frac{\alpha H + \sum_{i=1}^n \delta_{\theta_i}}{\alpha + n}$. In other words, the Dirichlet process provides a conjugate family of priors over distributions that is closed under posterior updates given observations. Rewriting the posterior DP, it follows that

$$G | \theta_1, \ldots, \theta_n \sim DP(\alpha + n, \frac{\alpha H + \sum_{i=1}^n \delta_{\theta_i}}{\alpha + n})$$  \hspace{1cm} (2.9)

It is noted that the posterior base distribution is a weighted average between the prior base distribution $H$ and the empirical distribution $\sum_{i=1}^n \frac{\delta_{\theta_i}}{n}$. The weight associated with the prior base distribution is proportional to $\alpha$, while the empirical distribution has weight proportional to the number of observations. Thus $\alpha$ can be interpreted as the strength or mass associated with the prior. As the amount of observations grows large, i.e. $n \to \infty$, the posterior will simply approach the empirical distribution which is in turn converges to the true underlying distribution uniformly. This establishes a consistency property of the DP: the posterior DP converges to the true underlying distribution as the amount of observations approaches infinity. On the other hand, as $\alpha \to 0$, the prior becomes non-informative in the sense that the predictive distribution is just determined by the empirical distribution.

Consider the predictive distribution for $\theta_{n+1}$, conditioned on $\theta_1, \ldots, \theta_n$ and with
$G$ marginalized out. Since $\theta_{n+1}|G, \theta_1, \ldots, \theta_n \sim G$, for a measurable $A \subset \Omega$, we have
\[
P(\theta_{n+1} \in A|\theta_1, \ldots, \theta_n) = E(G(A)|\theta_1, \ldots, \theta_n) = \frac{1}{\alpha + n}(\alpha H(A) + \sum_{i=1}^{n} \delta_{\theta_i}(A))
\] (2.10)

where the last step follows from the posterior base distribution of $G$ given the first $n$ observations. Thus with $G$ marginalized out:
\[
\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \alpha H(A) + \sum_{i=1}^{n} \delta_{\theta_i}
\] (2.11)

Therefore the posterior base distribution given $\theta_1, \ldots, \theta_n$ is also the predictive distribution of $\theta_{n+1}$. A salient property of the predictive distribution Eq. 2.11 is that it has point masses located at the previous draws $\theta_1, \ldots, \theta_n$. This implies that $G$ will take on previous values with positive probability regardless of the smoothness of $H$. Moreover, as the amount of observations grows, the value of any draw will be repeated by another draw, which shows that $G$ is indeed a discrete distribution. Therefore, the set of distribution over $\Omega$ with positive probability under DP is small. However, this set is also very large because any distribution over $\Omega$ can be approximated by a sequence of distributions from this set with arbitrary accuracy.

Another interesting feature can be seen from the predictive distribution Eq. 2.11 is the clustering property, which can be best illustrated by a metaphor. Suppose there is a Chinese restaurant with an infinite number of tables, each of which can serve one dish and sit an infinite many number of the customers. The first customer enters the restaurant and choose an arbitrary table to sit. The second customer enters and can either choose to sit with the first table or by himself at a new table. In general, the $n+1^{th}$ customer either joins a previous occupied table $k$ with the probability proportional to the number of customers sitting in table $k$, or sit at a new table with probability proportional to $\alpha$. Then the customers can be identified as the draws from the Dirichlet process with dishes representing the values they take and the tables representing the cluster. Fig. 2.3 illustrates the Chinese restaurant process. It should be noted that there is a rich gets richer phenomenon: the table with largest number of customers (or the most popular dish) expects to attract more customers than other tables. Moreover, it can be shown [23] that
Figure 2.3. Chinese restaurant process: the integers represent customers or the allocation variable. Each table \( k \) stands for a cluster, which serves one dish \( \theta_k \).

\[
E[m|n] \simeq a \log \left(1 + \frac{n}{\alpha}\right) \tag{2.12}
\]
\[
\text{Var}[m|n] \simeq a \log \left(1 + \frac{n}{\alpha}\right) \tag{2.13}
\]

where \( m \) is the number of clusters and \( n \) is the number of observations. It shows that the number of clusters grows only logarithmically in the number of observations.

2.3.3 Stick Breaking Representation

We have already observed that draws from a DP consist of a weighted sum of point mass. To make this precise, [24] provide a constructive definition of the DP as such, called the stick breaking representation, which is significantly more straightforward than the previous definition of the Dirichlet process. It is simply given as follows

\[
V_k \sim \text{Beta}(1, \alpha) \tag{2.14}
\]
\[
\pi_k = V_k \prod_{i=1}^{k-1} (1 - V_k) \tag{2.15}
\]
\[
\theta_k \sim H \tag{2.16}
\]
\[
G = \sum_{k=1}^{\infty} \pi_k \delta_k \tag{2.17}
\]

Then \( G \sim DP(\alpha, H) \). The construction of \( \pi \) can be understood metaphorically as follows. Starting with a stick of length 1, we break it at \( V_1 \), assigning \( \pi_1 \) to be the length of stick we just broke off. Now recursively break the other portion to obtain \( \pi_2, \pi_3 \) and so forth. Fig. 2.4 shows the procedure of stick breaking construction.
Figure 2.4. Stick breaking construction: A stick of length 1 is recursively broke off where $\pi_k$ denotes the length of the piece after the $k^{th}$ break.

The stick-breaking distribution over $\pi$ is sometimes written $\pi \sim GEM(\alpha)$, where the letters stand for Griffiths, Engen and McCloskey [25].

The above stick breaking representation can be generalized to construct the Pitman-Yor process [26] as follows

$$V_k \sim \text{Beta}(1 - b, a + kb)$$ (2.18)

$$\pi_k = V_k \prod_{l=1}^{k-1} (1 - V_l)$$ (2.19)

$$\theta_k \sim H$$ (2.20)

$$G = \sum_{k=1}^{\infty} \pi_k \delta_k$$ (2.21)

Then $G$ follows the Pitman-Yor process. Except the base distribution $H$, Pitman-Yor process has two parameters $a$ and $b$, where $a > -b$ and $0 \leq b < 1$. If we set $b = 0$, then the Pitman-Yor process reduces to the Dirichlet Process.

For a large range of real-world applications, the logarithmic growth described by Eq. 2.12 in the Dirichlet process is not a realistic assumption, since many important statistics are known to follow the power law [27]. Examples are the number of followers per user on Twitter, the size of cities, and the frequencies of words in the English language. It can be shown that the Pitman-Yor process can generate power law distributed clusters. Actually, for $b \neq 0$ we have

$$E[m|n] \simeq O(an^b)$$ (2.22)
where \( m \) is the number of clusters and \( n \) is the number of observations. And the predictive distribution of Pitman-Yor process is

\[
P(\theta_{n+1}|\theta_1, \ldots, \theta_n) = \frac{a + Knb}{n + a}H + \sum_{k=1}^{K_n} \frac{n_k - b}{n + a} \delta_{\theta_k}
\]

where \( n_k = \sum_{i=1}^{n} 1\{\theta_i = \theta_k\} \) and \( K_n \) is the number of unique \( \theta \).

### 2.3.4 Bayesian Nonparametric Mixture Model

Dirichlet Processes and Pitman-Yor processes are widely used in clustering problem for their clustering property. However, we have shown that the draw from the Dirichlet Process or Pitman-Yor process is a discrete distribution. Therefore, they are useless priors when we wish to deal with continuous distributions. This can be remedied by treating \( \theta \) as the parameter of the continuous variable \( x \)'s probability density function. Specifically, the following Dirichlet process mixture model can be used for clustering of continuous random variables (similar model can be developed for the Pitman-Yor Process)

\[
\pi \sim DP(\alpha, H) \tag{2.24}
\]

\[
\theta_i|\pi \sim \pi \tag{2.25}
\]

\[
x_i|\theta_i \sim F(\theta_i) \tag{2.26}
\]

where \( i = 1, \ldots, n \) and \( F \) represent the distribution of \( x \). The Fig. 2.5 shows the graph representation of the Dirichlet process mixture model.

This model is indeed an infinite mixture model, which is best illustrated by the following equivalent model

\[
V_k \sim Beta(1, \alpha) \tag{2.27}
\]

\[
\pi_k = V_k \prod_{l=1}^{k-1} V_l \tag{2.28}
\]

\[
\theta_k \sim H \tag{2.29}
\]

\[
z_i|\pi \sim Mult(\pi) \tag{2.30}
\]

\[
x_i|z_i, \theta \sim F(\theta_{z_i}) \tag{2.31}
\]
where $i = 1, \ldots, n$, $k = 1, \ldots, \infty$ and $z_i$ represent the allocation variable of $x_i$. 

Fig. 2.6 shows the dependence structure of this equivalent model. The infinite mixture models are particular useful when the number of clusters is not known a priori. The inference of model will return a posterior distribution over the number of clusters used to represent data and provide an alternative to model selection or averaging over finite mixture model. A typical example is the topics modeling problem, where the number of topics is not know a priori and will grow as the text corpus size increases.
Figure 2.6. Graphical representation of the infinite mixture model. 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.

2.4 Bayesian Network

2.4.1 Latent Allocation Variables

To construct a statistically interpretable and parsimonious (i.e., low-dimensional) model, the tensor factorization in Eq. (2.4) 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 dataset is arranged as $\{y_t, z_t\}_{t=1}^{T}$, where $t$ does not represent time but an index ranging from 1 to $T$. For later use, let us denote $y \equiv \{y_t\}_{t=1}^{T}$ and $z \equiv \{z_t\}_{t=1}^{T}$.

The conditional probability $p(y_t \mid z_t)$, factorized as in Eq. (2.4), is rewritten in the following form:

$$p(y_t \mid z_t) = \int_{x_{1,t}} \cdots \int_{x_{q,t}} p(y_t \mid x_t) \prod_{j=1}^{q} p(x_{j,t} \mid z_{j,t}) \tag{2.32}$$
where \( x_t \equiv (x_{1,t}, \ldots, x_{q,t}) \) denotes the latent class allocation variables, and \( x \equiv \{x_t\}_{t=1}^T \). For \( j = 1, \ldots, q \) and \( t = 1, \ldots, T \), it follows that

\[
x_{j,t} \mid \omega^{(j)}, z_{j,t} \sim \text{Mult}(\omega^{(j)}(z_{j,t})) \tag{2.33}
\]

\[
y_t \mid \tilde{\lambda}, x_t \sim \text{Mult}(\tilde{\lambda}_{x_t}) \tag{2.34}
\]

where \( \text{Mult} \) denotes multinomial distribution and \( \omega^{(j)} \equiv \{\omega^{(j)}(c)\}_{c=1}^{C_j} \) is the mixture probability matrix such that the \( c^{th} \) row \( \omega^{(j)}(c) \equiv \{\omega^{(j)}(c)\}_{s=1}^{S_j} \) is a probability vector. Moreover, \( \tilde{\lambda} \equiv \{\lambda_{s_1,\ldots, s_q}(s_1,\ldots, s_q)\} \) is a conditional probability tensor such that \( \lambda_{s_1,\ldots, s_q} \equiv \{\lambda_{s_1,\ldots, s_q}(c)\}_{c=1}^{C_0} \) 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 towards a low dimension.

- It follows from Eq. (2.33) that soft clustering is implemented for each predictor \( z_j \equiv \{z_{j,t}\}_{t=1}^T \) to inherit statistical strength across different categories.
- It follows from Eq. (2.34) that the distribution of \( y_t \) is determined by a reduced-order conditional probability tensor \( \tilde{\lambda} \).
- The clustering assignments \( x_j \equiv \{x_{j,t}\}_{t=1}^T \) 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 2.4.1.** It is important to distinguish between the number of clusters \( \tilde{k}_j \) formed by the latent allocation variables \( x_j \) and the dimension \( k_j \) of the mixing probability vector \( \omega^{(j)}(c) \). 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 inclusion of the predictor \( z_j \) in the model, because \( p(y_t \mid z_t) \) does not vary with \( z_{j,t} \) if \( z_j \) 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 3.1.2.

### 2.4.2 Sparsity Inducing Priors

In real-life applications, the tensor \( \tilde{\lambda} \) 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, $\tilde{\lambda}$ is clustered among different combinations of $(s_1, \ldots, s_q)$ in a nonparametric way by imposing Pitman-Yor process prior [22] on it. Thus, by employing the stick-breaking representation of Pitman-Yor Process [28], it follows that

$$\lambda_l | \gamma \sim \text{Dir}(\alpha), \quad \text{for } l = 1, \ldots, \infty$$

(2.35)

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

(2.36)

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

(2.37)

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

$$\phi_{s_1, \ldots, s_q} | \pi \sim \text{Mult}(\pi)$$

(2.38)

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

$$y_t | \lambda, \phi, x_t \sim \text{Mult}(\lambda_{\phi_{s_t}})$$

(2.39)

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

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

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

(2.40)

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

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

(2.41)

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

Remark 2.4.2. The exponential prior in Eq. (2.41) assigns increasing probabilities to smaller values of $k_j$ as the parameter $\mu_j$ becomes larger, and it is a uniform prior on $\{1, \ldots, C_j\}$ when $\mu_j$ 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 $\mu_j$ 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.

Combining Eqs. from (2.33) to (2.41) yields a Bayes network representation of the model; Figure 2.7 summarizes its dependency structure.

### 2.5 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 [29]:

Figure 2.7. 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.
• The information contents in priors are eventually dominated by those in the data.

• 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 $P(y_t \mid y_{t-1}, \ldots, y_{t-D}, \theta, \psi)$, a metric is defined on the space $\mathcal{P}$ as the following norm:

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

(2.42)

Let $P_0 \in \mathcal{P}$ be the true conditional probability tensor and let $p_0$ be the corresponding probability law of the true datagenerating process. Let $\Pi$ be the prior on $\mathcal{P}$ induced by the proposed model, and let $\Pi(\cdot \mid D_T)$ denote the corresponding posterior distribution given the observed dataset $D_T = \{y_t^{(k)}, y_{t-1}^{(k)}, \ldots, y_{t-D}^{(k)}, \theta_t^{(k)}, \psi_t^{(k)}\}_{t=1}^{T_k}$.

Having $T = (T_1, \ldots, T_K)$, we say $T \to \infty$ 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 literature [30] from the following two perspectives:

• The collected data are not independent identically distributed.

• 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 [31]. In this context, the following theorem [32] is relevant.

**Theorem 2.5.1.** Suppose $\mathcal{F}_1, \mathcal{F}_2, \ldots$ is a sequence of increasing $\sigma$-fields such that $R_t = \sum_{\tau=1}^{t} r_{\tau}$ is measurable with respect to $\mathcal{F}_t$ for every $t$. Let $i_t = E(r_t \mid \mathcal{F}_{t-1})$, $I_t = \sum_{\tau=1}^{t} i_{\tau}$ and $j_t = \text{Var}(r_t \mid \mathcal{F}_{t-1})$, $J_t = \sum_{\tau=1}^{t} j_{\tau}$. If there exists constants $\delta > 0$, $\gamma_t \to \infty$ such that

$$p(I_t / \gamma_t > \delta) \to 1$$

(2.43)
\[ J_t / \gamma_t^2 \overset{p}{\to} 0 \] (2.44)

then \( |R_t - I_t| \overset{p}{\to} 0 \).

To apply the above theorem in the present setting, the following notations are introduced for each \( k \):

\[
\begin{align*}
  r_t^{(k)}(P) &= \log \frac{p_0(y_t^{(k)} | \mathcal{F}_{t-1})}{p(y_t^{(k)} | \mathcal{F}_{t-1})}, \\
  R_t^{(k)}(P) &= \sum_{\tau=1}^{t} r_{\tau}^{(k)}(P) \\
  i_t^{(k)}(P) &= \text{E}(r_t^{(k)}(P) | \mathcal{F}_{t-1}), \\
  I_t^{(k)}(P) &= \sum_{\tau=1}^{t} i_{\tau}^{(k)}(P) \\
  j_t^{(k)}(P) &= \text{Var}(r_t^{(k)}(P) | \mathcal{F}_{t-1}), \\
  J_t^{(k)}(P) &= \sum_{\tau=1}^{t} j_{\tau}^{(k)}(P)
\end{align*}
\]

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 [33]. Based on these concepts, the following assumption is made.

**Assumption 2.5.1.** Three regular conditions of the assumption are stated below.

1. \( p(\delta_1^{(k)}(P) < i_t^{(k)}(P)/\gamma_{T_k}^{(k)} < \delta_2^{(k)}(P)) \rightarrow 1 \) uniformly on any compact subset of \( \mathcal{P} \setminus P_0 \) for any \( k \), where \( \delta_1^{(k)}(P) \) and \( \delta_2^{(k)}(P) \) are positive continuous functions of \( P \), and \( \gamma_{T_k}^{(k)} \rightarrow \infty \).

2. \( J_{i_{T_k}}^{(k)}(P)/\gamma_{T_k}^{(k)} \overset{p}{\to} 0 \) uniformly on any compact subset of \( \mathcal{P} \setminus P_0 \) for any \( k \).

3. Let \( B_\epsilon = \{ P \in \mathcal{P} : \|P - P_0\| < \epsilon \} \), \( U_\epsilon = \{ P \in \mathcal{P} : \delta_1^{(k)}(P) < \epsilon \text{ for any } k \} \) and \( L_\epsilon = \{ P \in \mathcal{P} : \delta_2^{(k)}(P) < \epsilon \text{ for any } k \} \). 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 2.5.1.** Condition 1 in Assumption 2.5.1 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 2.5.1 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 2.5.1 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 [32].

The following theorem establishes the posterior consistency of the proposed model under Assumption 2.5.1 by showing that $\Pi(\cdot \mid D_T)$ concentrates in arbitrarily small neighborhoods of $P_0$ as $T \to \infty$.

**Theorem 2.5.2.** If the true data generating process $p_0$ satisfies Eq. (2.1) and all three conditions in Assumption 2.5.1, then

$$\forall \epsilon > 0, \quad \Pi(B^c_\epsilon \mid D_T) \overset{p}{\to} 0 \text{ with respect to } p_0 \text{ as } T \to \infty$$

**Proof.** Based on the first two conditions in Assumption 2.5.1 and Theorem 2.5.1, it follows that

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

uniformly on any compact subset of $P \setminus P_0$. Because of the third condition in Assumption 2.5.1, 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,

$$\Pi(B^c_\epsilon \mid D_T) = \frac{\int_{B_\epsilon^c} \exp\left(-\sum_{k=1}^{K} R_{T_k}^{(k)}(P)\right) d\Pi(P)}{\int_{P} \exp\left(-\sum_{k=1}^{K} R_{T_k}^{(k)}(P)\right) d\Pi(P)}$$

$$= \frac{\int_{B_\epsilon^c} \exp\left(\eta_3 \gamma_T - \sum_{k=1}^{K} R_{T_k}^{(k)}(P)\right) d\Pi(P)}{\int_{P} \exp\left(\eta_3 \gamma_T - \sum_{k=1}^{K} R_{T_k}^{(k)}(P)\right) d\Pi(P)}$$

$$\leq \frac{\int_{L_{\eta_1}} \exp\left(\eta_3 \gamma_T - \sum_{k=1}^{K} R_{T_k}^{(k)}(P)\right) d\Pi(P)}{\int_{U_{\eta_2}} \exp\left(\eta_3 \gamma_T - \sum_{k=1}^{K} R_{T_k}^{(k)}(P)\right) d\Pi(P)}$$

$$\equiv \frac{N(D_T)}{D(D_T)}$$

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

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

Similarly for $D(D_T)$, because $\sum_{k=1}^{K} R_{T_k}^{(k)}(P) \xrightarrow{p} \sum_{k=1}^{K} I_{T_k}^{(k)}(P)$ pointwise on $U_{\eta_2}$, the limit inferior of the integrand of $D(D_T)$ is infinity. Because the prior $\Pi$ has full support on $P$, $\Pi(U_{\eta_2}) > 0$. By Fatou’s lemma, we obtain that $\liminf_{T \to \infty} D(D_T) = \infty$. Now, it is concluded that $\forall \epsilon > 0 \ \Pi(B^c_\epsilon \mid D_T) \xrightarrow{p} 0$ as $T \to \infty$. \hfill \square
Chapter 3

Statistical Inference

In Chapter 2, a Bayesian network has been developed for the estimation of conditional probability tensors, which frames all inference about the unknown quantities as a calculation of the posterior distribution. However, in the proposed model, the posterior distribution involves complicated integrals, the computation of which is intractable. We therefore need to resort to approximation methods for posterior computation. In this chapter, two approximation methods for this purpose have been developed: Gibbs sampling [34] and stochastic variational inference [35]. Gibbs sampling has the advantage of being asymptotically exact and is most suitable for small datasets with high accuracy requirement. Additionally, Gibbs sampling can handle varying dimension model and thus provide a Bayesian approach for hypothesis test on the significance of the predictors. On the other hand, stochastic variational inference has an explicit objective to maximize. Therefore, it is easier to monitor the convergence of the algorithm. Moreover, stochastic variational inference is significantly faster than Gibbs sampling in most cases, rendering it scalable to large datasets and feasible for real time implementation.

This chapter is organized as follows. Section 3.1 first presents the Gibbs sampling algorithm for posterior computation and hypothesis test of the proposed model. Then the algorithm is validated with a numerical example and a public available econometric dataset. In Section 3.2, a stochastic variation inference algorithm is developed for the inference of proposed model and compared with the Gibbs sample algorithm.
3.1 Gibbs sampling

3.1.1 Posterior Computation

Gibbs sampling is one of the Monte Carlo Markov Chain techniques for sampling from a desired probability distribution based on construction of a Markov chain that has the desired probability distribution as its stationary distribution. Because the posterior distribution of the Bayes network developed in Chapter 2 has no analytic form, a possible way for the statistical inference is to use Gibbs sampling to draw samples from the posterior distribution, which can then be used to calculate the quantities of interest such as the expectation and the percentiles.

Since the dimension of $\omega^{(j)}$ 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 (MCMC) [36], which performs the trans-dimensional exploration in the model space. The difficulties of trans-dimensional modeling can be circumvented by product partition model [37,38] that allows the construction of a stationary Markov chain on the clustering configuration space. In this thesis, the varying dimension $\omega^{(j)}$ is integrated out to sample $k_j$ directly from $p(k_j \mid x_j, z_j)$ (See Appendix A for details), which forms a partially collapsed Gibbs sampler [39] that alternates between the space with all variables and the space with all variables but $\omega = \{\omega^{(j)}\}_{j=1}^q$.

To compute the posteriors of the Pitman-Yor process (See Appendix B), the infinite dimensional $\pi$ and $\lambda$ at $L^{th}$ component are truncated, as presented in [28] to achieve the desired accuracy by choosing an appropriate $L$, which is chosen in this thesis to be 100 to satisfy the accuracy requirements. Posterior sampling of other variables are straightforward. The details are summarized in Algorithm 1, where $\xi$ collects the variables that are not explicitly mentioned.

To execute Algorithm 1, several hyperparameters need to be chosen. The implication and determination of $\mu_j$ 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 $b = 0$ suffices for all applications in this thesis, 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 $\alpha$ and $\beta_j$ are hyperparameters of Dirichlet distribution and serve as pseudo-counts. 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 thesis, they are chosen to be: $\alpha = 1$ and $\beta_j = 1/C_j$ for all applications.

### 3.1.2 Bayesian Hypothesis Testing

This subsection performs hypothesis testing on the importance of a particular predictor for interpreting the underlying model in many applications as demonstrated in Chapter 5. It is also a better utilization of computational resources for sequential classification by excluding the unimportant predictors. As mentioned earlier, a particular predictor $z_j$ is important if and only if the number of clusters $\tilde{k}_j$ formed by the corresponding latent class allocation variables $x_j$ is greater than 1. Therefore, to perform Bayesian tests for the hypothesis described above, one only needs to compute the Bayes factor [40] in favor of $H_1: \tilde{k}_j > 1$ against $H_0: \tilde{k}_j = 1$ given by

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

(3.1)

where $p(H_0|y, z)$ and $p(H_1|y, z)$ are equal to the proportions of samples in which the $\tilde{k}_j$’s 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.

$$p(\tilde{k}_j = 1) = \sum_{k=1}^{C_j} p(k_j = k) \sum_{l=1}^{k} p(x_j,l = l \forall t|k_j = k)$$

$$= \left( \prod_{r=1}^{C_j} \gamma_{j}^{(n_{j,c})} \right) \left( \sum_{k=1}^{C_j} \frac{p(k_j = k)k}{\prod_{k=1}^{C_j}(k \gamma_{j})^{(n_{j,c})}} \right)$$

where the pertinent parameters are defined in Algorithm 1.

### 3.1.3 Numerical example

This subsection evaluates finite-sample performance of the proposed method with simulated data generated from a non-stationary Markov model, which is a special case of panel data. The underlying Markov model for data generation is known.
Algorithm 1: Gibbs sampling for proposed model

Require: Dataset \{y_t, z_t\}_{t=1}^T, hyperparameters \alpha, \beta, \{\mu_j\}_{j=1}^q, \{\beta_j\}_{j=1}^q, truncating components \(L\), number of samples \(N\), and the initial sample \((\phi_0, \pi_0, \lambda_0, \omega_0, x_0, k_0)\)

Ensure: Posterior samples \{(\phi_n, \pi_n, \lambda_n, \omega_n, k_n)^N\}_{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
      \[
p(\phi_{s_1, \ldots, s_q} = l \mid \xi) \propto \pi_l \prod_{c=1}^{C_q} (\lambda_l(c))^{n_{s_1, \ldots, s_q}(c)}\]
      where \(n_{s_1, \ldots, s_q}(c) = \sum_{t=1}^{T} 1\{x_{s_1,t} = s_1, \ldots, x_{s_q,t} = s_q, y_t = c\}\).
   3: For \(l = 1, \ldots, L\), update \(\pi_l\) as follows
      \[
      V_l \mid \xi \sim \text{Beta}(1 - b + n_l, a + b + \sum_{k > l} n_k), \quad l < L
      \]
      \[
      V_L = 1, \quad \pi_l = \frac{V_l}{\sum_{k=1}^{l-1} (1 - V_k)}
      \]
      where \(n_l = \sum_{(s_1, \ldots, s_q)} 1\{\phi_{s_1, \ldots, s_q} = l\}\).
   4: For \(l = 1, \ldots, L\), sample \(\lambda_l\) from their Dirichlet full conditionals
      \[
      \lambda_l \mid \xi \sim \text{Dir}(\alpha + n_l(1), \ldots, \alpha + n_l(C_0))
      \]
      where \(n_l(c) = \sum_{(s_1, \ldots, s_q)} 1\{\phi_{s_1, \ldots, s_q} = l\} n_{s_1, \ldots, s_q}(c)\).
   5: For \(j = 1, \ldots, q\) and \(c = 1, \ldots, C_j\), sample
      \[
      \omega^{(j)}(c) \mid \xi \sim \text{Dir}(\beta_j + n_{j,c}(1), \ldots, \beta_j + n_{j,c}(k_j))
      \]
      where \(n_{j,c}(s_j) = \sum_{t=1}^{T} 1\{x_{j,t} = s_j, z_{j,t} = c\}\).
   6: For \(j = 1, \ldots, q\) and for \(t = 1, \ldots, T\), sample the \(x_{j,t}\) from their multinomial full conditionals
      \[
      p(x_{j,t} = s \mid \xi, x_{i,t} = s_i, i \neq j) \propto \omega_s^{(j)}(z_{j,t}) \lambda_{\phi_{s_1, \ldots, s_q}}(y_t)
      \]
   7: For \(j = 1, \ldots, q\), sample \(k_j\) using their multinomial full conditionals
      \[
      p(k_j = k \mid \xi) \propto \exp(-\mu_j k) \prod_{c=1}^{C_j} n_{j,c}^{-k_{j,c}}, \quad k_j = \max\{x_{j,t}, \ldots, C_j\}
      \]
      where \(n_{j,c} = \sum_{t=1}^{T} 1\{z_{j,t} = c\}\).
   8: end for

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 non-stationary Markov model \(p(y_t | F_{t-1}) = p(y_t | y_{t-1}, y_{t-2}, y_{t-5}, \theta_t)\), where only
the time-lags $y_{t-1}, y_{t-2}, y_{t-5}$ and the exogenous variable $\theta_t$ are important predictors. The exogenous variable $\theta_t$ itself is binary and is generated by a 1st 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 3.1. In this table, whenever $\theta_t = 0$ the probability of $y_t = 0$ is 0.5, which corresponds to a white noise.

<table>
<thead>
<tr>
<th>$y_{t-1}$</th>
<th>$y_{t-2}$</th>
<th>$y_{t-5}$</th>
<th>$\theta_t$</th>
<th>$p(y_t = 1)$</th>
<th>$p(y_t = 0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.25</td>
<td>0.75</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.15</td>
<td>0.85</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.15</td>
<td>0.85</td>
</tr>
</tbody>
</table>

To estimate $p(y_t|\mathcal{F}_{t-1})$, 506 samples of $\{y_t\}_{t=1}^{506}$ and 500 samples of $\{\theta_t\}_{t=1}^{506}$ are col-
Figure 3.1. Gibbs sampling results for the numerical example. The horizontal (solid red line) indicates the threshold = 3.
lected, 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 $z_t \equiv (y_{t-1}, y_{t-2}, \ldots, y_{t-6}, \theta_t)$. To compute posteriors using Algorithm 1, $j/2$ is assigned to $\mu_j$ for $j = 1, \ldots, 6$, and 0 is assigned to $\mu_7$.

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 5th sample to reduce the autocorrelation. Figure 3.1 summarizes the Gibbs sampling results based on the downsampled after-burn-in samples. Figure 3.1(a) exhibits the log-likelihood of the model for 10,000 iterations. Figure 3.1(b) illustrates 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 $\theta_t$. Figure 3.1(c) shows relative frequency of how many predictors are important. The proposed method also leads to parsimonious representations as shown in Figure 3.1(d) and Figure 3.1(e). As discussed in Subsection 2.4.2, 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. Figure 3.1(f) illustrates the Bayes factors as computed in Subsection 3.1.2 for different predictors. Bayes factor $BF_{10}$ in Eq. (4.2) can be interpreted as the evidence against $H_0$, and a threshold of $t = 3$ indicates that this evidence is positive [41], which is used in the sequel. Furthermore, $BF_{10} > 150$ indicates very strong evidence against $H_0$ [41]. If the inclusion proportions in Figure 3.1(b) equal to 1, then the corresponding Bayes factors in Figure 3.1(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 identifying the model structure, the proposed method estimates the transition probabilities. Figure 3.2 illustrates two cases from Table 3.1. When $y_{t-1} = 0, y_{t-2} = 1, y_{t-5} = 0, \theta_t = 1$, it follows from Table 3.1 that the true transition probability of $y_t = 0$ is 0.9. Similarly, when $y_{t-1} = 0, y_{t-2} = 1, y_{t-5} = 1, \theta_t = 0$, it also follows from Table 3.1 that the true transition probability of $y_t = 0$ is 0.5. Figure 3.2 exhibits the profiles of corresponding transition probabilities, where the estimated transition probability per sample are obtained with the running mean as well as the 95% and 5% percentiles. It is seen in
Figure 3.2 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 $p(y_t = 0 \mid y_{t-1} = 0, y_{t-2} = 1, y_{t-5} = 1, \theta_t = 0)$ 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 dataset. 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 $\theta_t$. This is accomplished by a recursive Bayes filter (RBF) to obtain $p(\theta_t \mid y_{1:t})$ for the estimated and true models, where the prediction and update equations of the RBF are presented as:

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

$$p(\theta_{t+1} \mid y_{1:t+1}) = \frac{p(y_{t+1} \mid y_{1:t}, \theta_{t+1})p(\theta_{t+1} \mid y_{1:t+1})}{\sum_{\theta_{t+1}} p(y_{t+1} \mid y_{1:t}, \theta_{t+1})p(\theta_{t+1} \mid y_{1:t+1})}$$
The single plate in Figure 3.3 compares the profiles of probabilities $p(\theta_t = 1 \mid y_{1:t})$ of the estimated model and the true model, respectively, for 500 samples. It is observed that the estimated model behaves very similarly to the true model after the initial transients are over. To classify $\theta_t$, let us consider a decision rule such that $\theta_t$ is classified to be 1 if the $p(\theta_t = 1 \mid y_{1:t}) > t$ for a given threshold $t$, where an optimal 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 3.2. 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.

### 3.1.4 Validation on a Public Dataset

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 dataset of 7,293 individuals for a varying number of observations for each individual over a period of 1 to 7 years. The raw data have been preprocessed before their usage for testing the proposed nonparametric model. The testing procedure is briefly described below.

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 10 predictors \(z \equiv (z_1, \ldots, z_{10})\) 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 dataset, with 0 representing male and 1 representing female; \(z_3\) is the \(MARRIED\) variable in the original dataset, 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 dataset, 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 dataset, 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 dataset, 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); \(z_9\) is corresponds to the sum of the doctor visits and hospital visits in the original dataset, with \(z_9 = i\) if \(i - 1 < (DOCVIS + HOSPVIS)/10 \leq i\); and \(z_{10}\) corresponds to the insurance status in the original dataset, with \(z_{10} = 2\) if \(PUBLIC = 1\), \(z_{10} = 1\) if \(ADDON = 1\) and \(z_{10} = 0\) otherwise (uninsured).

After data preprocessing, 2,000 samples are randomly chosen from the dataset

\(^1\)http://people.stern.nyu.edu/wgreene/Econometrics/PanelDataSets.htm
with 1,500 of them as training set and the remaining 500 as the test set. This procedure is repeated 5 times and the average test results are recorded. To compute posteriors using Algorithm 1, the hyperparameters are set as: $\mu_j = 1$ for $j = 1, \ldots, 10$.

**Figure 3.4.** Gibbs sampling results for German Health Care data

Gibbs sampling has been performed for 150,000 iterations, where the initial 100,000 data points are discarded as burn-in period and the after burn-in samples are downsamples by taking every 5th data point to reduce the autocorrelation. Figure 3.4 shows the results of Gibbs sampling for this dataset. 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 3.3 presents the classification results of the test set using the proposed nonparametric model versus various standard machine learning methods [42]. 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 (MAP) 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 support vector machine (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, 10 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.

<table>
<thead>
<tr>
<th>Table 3.3. Results for different machine learning methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
</tr>
<tr>
<td>Nonparametric model</td>
</tr>
<tr>
<td>Logistic Regression</td>
</tr>
<tr>
<td>KNN ($K = 5$)</td>
</tr>
<tr>
<td>SVM</td>
</tr>
<tr>
<td>Decision Tree</td>
</tr>
<tr>
<td>Random Forest</td>
</tr>
<tr>
<td>AdaBoost</td>
</tr>
<tr>
<td>MLPNN</td>
</tr>
</tbody>
</table>

### 3.2 Stochastic variational inference

#### 3.2.1 Variational Inference

Variational inference [43] is a method that converts intractable Bayesian inference problems into high-dimensional optimization problems, for which a variety of optimization techniques can be used such as coordinate ascent [44] and stochastic gradient descent algorithms [45]. In this subsection, we review the variational inference method for a class of Bayesian models, which contains the proposed the
Bayesian network when the dimensions of latent features $k_j$ are fixed.

The models of interest involves observations, local latent variables, global unknown parameters and fixed hyperparameters. The $n$ observations are denoted as $x = x_{1:n}$ and $z = z_{1:n}$ are local latent vectors, each of which is a collection of $m$ variables $z_i = z_{i,1:m}$. The global parameter vector is $\theta$ and the fixed hyperparameters are $\alpha$. (Note $\alpha$ can govern local random variables, but they are only allowed to be parameters of global variables in the sequel for the simplicity of notations.)

We assume that the joint distribution can factorize into a global terms and a product of local terms as follows

$$p(x, z, \theta|\alpha) = \prod_{i=1}^{n} p(x_i, z_i|\theta)$$ (3.2)

Fig. 3.5 illustrates the graphical model. The goal here is to approximate the posterior distribution of latent variables given the observations, i.e., $p(\theta, z|x)$. It follows that the $i^{th}$ observation $x_i$ and the $i^{th}$ local latent variables $z_i$ are conditionally independent of all other observations and local variables given the global parameters

$$p(x_i, z_i|x_{-i}, z_{-i}, \theta, \alpha) = p(x_i, z_i|\theta, \alpha)$$ (3.3)

where $x_{-i}$ and $z_{-i}$ denotes the set of local variables except the $i^{th}$.

After describing the assumptions of the dependence structure, we further assume that the global parameters and local latent variables are from conjugate exponential families [46]. Therefore, the complete conditionals of global parameters and the local contexts have the following form

$$p(\theta|x, z, \alpha) = h_g(\theta) \exp\{\eta_g(x, z, \alpha)^T t_g(\theta) - a_g(\eta_g(x, z, \alpha))\}$$ (3.4)

$$p(z_{i,j}|x_i, z_{i,-j}, \theta) = h_l(z_{i,j}) \exp\{\eta_l(x_i, z_{i,-j}, \theta)^T t_l(z_{i,j}) - a_l(\eta_l(x_i, z_{i,-j}, \theta))\}$$ (3.5)

The subscripts $g$ and $l$ are used to indicate that an object is associate with global or local terms. The scalar function $h(\cdot)$ and $a(\cdot)$ are respectively the base measure and and log-normalizer of the distribution in the exponential family. The vector function $\eta(\cdot)$ and $t(\cdot)$ represent the natural parameters and sufficient statistics of the distribution in the exponential family. Because these are conditional distributions, the natural parameter $\eta$ is a function of the variables being conditioned on.
Due to the conjugacy between the global parameters and the local contexts, $\eta_g$ must have the following form

$$\eta_g(x, z, \alpha) = (\alpha_1 + \sum_{i=1}^{n} t(x_i, z_i), \alpha_2 + n)$$  \hspace{1cm} (3.6)$$

where $t$ is the sufficient statistics of the local context, $\alpha_1$ is a vector of the same dimension as $\theta$ and $\alpha_2$ is a scalar.

This class of models with local latent variables and global unknown parameters, where the complete conditionals are in the exponential family, covers a wide range of useful Bayesian models such as Bayesian mixture models, latent Dirichlet allocation [47] and hidden Markov model [48]. The proposed Bayesian network belongs to this class because the global parameters and the local latent variables are from Dirichlet and Multinomial distribution, which are in the conjugate exponential family.
Variational inference solves the inference problem by introducing a family of distributions over the unknown variables that is indexed by a set of free parameters, and then optimize those parameters to find the member of the family that is closest to the posterior in the sense of Kullback-Leibler divergence \[49\] (KL divergence). The distribution used to approximate the posterior is called the variational distribution.

The simplest variational family is the mean field family, which assume that each unknown variable is independent and governed by its own parameters,

\[
q(z, \theta) = q(\theta | \lambda) \prod_{i=1}^{n} \prod_{j=1}^{m} q(z_{i,j} | \phi_{i,j}) \tag{3.7}
\]

where \(\lambda\) are the global variational parameters and \(\phi_i = \phi_{i,1:m}\) are the local variational parameters that govern the \(i^{th}\) local latent variable.

The KL divergence between the variational distribution and the posterior is

\[
\text{KL}(q(z, \theta) || p(z, \theta | x)) = \mathbb{E}_q[q(z, \theta)] - \mathbb{E}_q[p(z, \theta | x)] \\
= \mathbb{E}_q[q(z, \theta)] - \mathbb{E}_q[p(x, z, \theta)] + \log p(x) \tag{3.8}
\]

where \(\log p(x)\) is the log marginal probability and \(\mathcal{L}(q) = \mathbb{E}_q[q(z, \theta)] - \mathbb{E}_q[p(x, z, \theta)]\) is the evidence lower bound, which the variational inference needs to maximize.

There are tow implications of Eq. 3.8

- The evidence lower bound is a lower bound of log marginal probability because the KL divergence is always greater or equal to zero.
- The maximization of the evidence lower bound is equivalent to finding the variation distribution that is closest to the posterior in KL divergence because the log marginal probability does not depend on the variational distribution.

Under the assumption we have made about the model, it can be shown by the calculus of variation \[50\] that setting the derivative of the evidence lower bound with respect to variational distribution to be zero is equivalent to setting the variational distribution as follows

\[
q(\theta | \lambda) = h_g(\lambda) \exp\{\lambda^T t_g(\theta) - a_g(\lambda) \tag{3.9}\}
\]

\[
q(z_{i,j} | \phi_{i,j}) = h_l(z_{i,j}) \exp\{\phi_{i,j}^T t_l(z_{i,j}) - a_l(\phi_{i,j}) \tag{3.10}\}
\]
where the variational parameters $\lambda$ and $\phi_{i,j}$ satisfies the following equations

$$\lambda = \mathbb{E}_q[\eta_g(x, z, \alpha)] \quad (3.11)$$
$$\phi_{i,j} = \mathbb{E}_q[\eta_l(x_i, z_{i,-j}, \theta)] \quad (3.12)$$

By updating the variational parameters $\lambda$ and $\phi_{i,j}$ alternatively according to Eq. 6-3.12, we form the algorithm for coordinate ascent variational inference. The full algorithm is summarized in the Algorithm 2, which is guaranteed to find a local optimum of the evidence lower bound. Computing the expectations at each step is easy for Bayesian networks with tractable complete conditionals.

**Algorithm 2 Coordinate ascent variational inference**

1: Initialize the global variational parameter $\lambda$ randomly
2: while the evidence lower bound not converge do
3: for each local variational parameter $\phi_{i,j}$ do
4: Update $\phi_{i,j}$: $\phi_{i,j} = \mathbb{E}_q[\eta_l(x_i, z_{i,-j}, \theta)]$
5: end for
6: Update the global variational parameters: $\lambda = \mathbb{E}_q[\eta_g(x, z, \alpha)]$
7: end while

### 3.2.2 Stochastic Variational Inference

The coordinate ascent algorithm is not efficient for large datasets because the local variational parameters have to be optimized for each data point before re-estimating the global variational parameters. Stochastic variational inference [35] leverages stochastic optimization to mitigate this problem. Data are repeatedly subsampled to form noisy estimates of the natural gradient of the evidence lower bound, and these estimates are followed with a decreasing step size. In this subsection, we review the stochastic variational inference for the class of models discussed in Section 3.2.

The natural gradient of a function takes the information geometry of its parameter space into consideration, adjusting the direction of the traditional gradient with a Riemannian metric. Natural gradients for maximum-likelihood estimation is discussed in [51], which gives faster convergence than standard gradients. It can be shown that the natural gradient of the evidence lower bound in Eq. 3.8 has the
following simple form

\[
\nabla_\lambda L = \mathbb{E}_q[\eta_q(x, z, \alpha)] - \lambda \\
\n\nabla_{\phi_{i,j}} L = \mathbb{E}_q[\eta_l(x_i, z_i, -j, \theta)] - \phi_{i,j}
\]

Stochastic optimization algorithms uses noisy estimates of the gradient with a decreasing step size. Noisy estimates of a gradient are often cheaper to compute than the true gradient, and following such estimates can allow algorithms to escape shallow local optima of complicated objective functions. In variational inference of the global parameters, the gradient can be written as a sum of local terms (one for each data point) and a fast noisy approximation can be computed by subsampling the data. Under certain conditions on the step size schedule, these Stochastic optimization algorithms will converge to a local optimum [52]. An overview of stochastic optimization is given in [53]; Bottou [54] gives an overview of its role in machine learning.

Consider an objective function \( f(\lambda) \) and a random function \( B(\lambda) \) that has expectation equal to the gradient so that \( \mathbb{E}_q[\lambda] = \nabla_\lambda f(\lambda) \). Then the stochastic gradient algorithm, a kind of stochastic optimization algorithms, updates the parameter \( \lambda \) iteratively following the realizations of \( B(\lambda) \). Specifically, at iteration \( t \), we have

\[
\lambda^{(t)} = \lambda^{(t-1)} + \rho_t b_t(\lambda^{(t-1)})
\]

where \( b_t \) is an independent draw from the random function \( B(\cdot) \). If the sequence of step sizes \( \rho_t \) satisfies

\[
\sum \rho_t = \infty \quad \sum \rho_t^2 < \infty
\]

then \( \lambda^{(t)} \) will converge to a local optimal \( \lambda^* \). A commonly used step size schedule is specified as follows

\[
\rho_t = (t + \tau)^{-\kappa}
\]

where \( \kappa \in (0.5, 1] \) is the forgetting rate which controls how quickly old information is forgotten; the larger the forgetting rate is, the smaller the step size is. On the other hand, the delay \( \tau \geq 0 \) down-weights early iterations.
Stochastic variational inference optimizes the evidence lower bound $L(\lambda)$ by subsampling the data to form noisy estimates of the natural gradient. First, the evidence lower bound $L(\lambda)$ can be decomposed into a global term and a sum of local terms

$$L(\lambda) = \mathbb{E}_q[\log p(\theta)] - \mathbb{E}_q[\log q(\theta)] + \sum_{i=1}^n \max_{\phi_i} \left( \mathbb{E}_q[\log p(x_i, z_i|\theta)] - \mathbb{E}_q[\log q(z_i)] \right)$$

Then define the following random function of the variational parameters,

$$L_I(\lambda) = \mathbb{E}_q[\log p(\theta)] - \mathbb{E}_q[\log q(\theta)] + n \max_{\phi_I} \left( \mathbb{E}_q[\log p(x_I, z_I|\theta)] - \mathbb{E}_q[\log q(z_i)] \right)$$

where $I$ is a discrete uniform distribution on $\{1, \ldots, n\}$. It follows that the expectation of $L_I$ equals to the evidence lower bound $L(\lambda)$. Therefore, the natural gradient of $L_I$ with respect to the global variational parameter $\lambda$ is a noisy but unbiased estimate of the natural gradient of $L$. Suppose the realization of $I$ is $i$, then the natural gradient of $L_I$ with respect to the global variational parameter $\lambda$ can be computed based on Eq. 3.13

$$\nabla_{\lambda} L_i = \mathbb{E}_q[\eta_g(x^{(n)}_i, z^{(n)}_i, \alpha)] - \lambda$$

where $\{x^{(n)}_i, z^{(n)}_i\}$ are a dataset formed by $n$ copies of observation $x_i$ and the corresponding local latent variable $z_i$.

Finally, the noisy natural gradients in Eq. 3.19 are used to optimize the evidence lower bound. Define the intermediate global variational parameters $\tilde{\lambda}_t$ to be $\mathbb{E}_q[\eta_g(x^{(n)}_i, z^{(n)}_i, \alpha)]$, then the update rule in Eq. 3.15 can be written as follows

$$\lambda^{(t)} = \lambda^{(t-1)} + \rho_t (\tilde{\lambda}_t - \lambda^{(t-1)})$$

To summarize, the stochastic variational inference is performed by repeating the following steps

1. Sample a data point from the dataset and optimize its corresponding local variational parameters.

2. Compute the intermediate global variational parameters using the same
formula in coordinate ascent variational inference as if the sampled data point
were repeated N times.

3. Update the global variational parameters to be a weighted average of the
intermediate parameters and their current values with appropriately chosen
weights.

Alg. 3 illustrates the details of the stochastic variational inference algorithm.

Algorithm 3 Stochastic variational inference
1: Initialize the global variational parameter $\lambda$ randomly
2: Set the step size $\rho_t$ appropriately
3: while the evidence lower bound not converge do
4: Sample a data point $x_i$ uniformly from the dataset
5: Compute its local variational parameter
   $\phi = \mathbb{E}_q[\eta_l(x_i^{(n)}, z_i^{(n)})]$
6: Compute the intermediate global parameters as if $x_i$ is replicated $n$ times
   $\tilde{\lambda}_t = \mathbb{E}_q[\eta_g(x_i^{(n)}, z_i^{(n)})]$
7: Update the global variational parameters
   $\lambda^{(t-1)} = (1 - \rho_t) \lambda^{(t-1)} + \rho_t \tilde{\lambda}_t$
8: end while

The stochastic variational inference can be extended to the minibatch mode
by using multiple samples. Suppose $B$ samples are used at each iteration, then
the update rule for the global variational parameters in Eq. 3.20 is replaced by the
following equation

$$\lambda^{(t)} = (1 - \rho_t) \lambda^{(t-1)} + \rho_t \frac{1}{B} \sum_{i} \tilde{\lambda}_i$$  (3.21)

Because the noisy natural gradients associated with each data point have expected
value equal to the gradient, the average of these noisy natural gradients has the
same expectation and the Alg. 3 remains valid. There are two reasons to use
minibatches

- The update of local variational parameters for multiple data points can be
  computed in parallel, which significantly accelerates the algorithm.

- Averaging over minibatches yields more accurate intermediate global estimates
  and helps the algorithm to escape the shallow local optima.
3.2.3 Stochastic Variational Inference for the Proposed Model

The stochastic variational inference presented above can be directly applied to the Bayesian network proposed in Chapter 2 when the number of latent features $k_j$ is fixed. Compared with Gibbs sampling algorithm (Alg. 3), the stochastic variational inference lacks the ability to identify the significant predictors and the structure of the model. However, the stochastic variational inference is much faster than Gibbs sampling algorithm in most cases. Therefore, it is most useful in the scenario when large dataset needs to be processed in a short time. A typical example is the real time detection of abnormality.

To perform the variational inference, we need to introduce a variational distribution for each variable. Table 3.4 collects the information about the type, variational distribution and variational parameters of each variable. Alg. 4 summarizes the procedures of the stochastic variational inference for the proposed model.

### Table 3.4. Variational distributions for the proposed model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Variational Distribution</th>
<th>Variational Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_l$</td>
<td>Global</td>
<td>Dirichlet</td>
<td>$\gamma_l$</td>
</tr>
<tr>
<td>$V_l$</td>
<td>Global</td>
<td>Beta</td>
<td>$\nu_l, \mu_l$</td>
</tr>
<tr>
<td>$\phi_{s_1,\ldots,s_q}$</td>
<td>Global</td>
<td>Multinomial</td>
<td>$\theta_{s_1,\ldots,s_q}$</td>
</tr>
<tr>
<td>$\omega^{(j)}(c)$</td>
<td>Global</td>
<td>Dirichlet</td>
<td>$\eta_{j,c}$</td>
</tr>
<tr>
<td>$x_{j,t}$</td>
<td>Local</td>
<td>Multinomial</td>
<td>$\rho_{j,t}$</td>
</tr>
</tbody>
</table>

3.2.4 Numerical Example

This subsection validates the stochastic variational inference with simulated data generated from a non-stationary 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 non-stationary Markov model $p(y_t|F_{t-1}) = p(y_t|y_{t-1}, y_{t-2}, y_{t-5}, \theta_t)$, where only the time-lags $y_{t-1}, y_{t-2}, y_{t-5}$ and the exogenous variable $\theta_t$ are important predictors.
Algorithm 4 Stochastic variational inference for the proposed model

1: Initialize the global variational parameter $\gamma$, $\nu$, $\mu$, $\theta$, $\eta$ randomly
2: Set the step size $\rho_t$ appropriately
3: while the evidence lower bound not converge do
4: Sample a data point $\{z_t, y_t\}$ uniformly from the dataset
5: while not converge do
6: For each $j$ and $s$, update the local variational parameters as follows
   \[
   \rho_{j,t,s} \propto \exp\left\{ \sum_{s_1, \ldots, s_{j-1}, s_{j+1}, \ldots, s_q} \left( \sum_{t=1}^L (\psi(\gamma_{l,y_t}) - \psi(C_0)\gamma_{l,c}) \theta_{(s_1, \ldots, s_q, l = 1)} \prod_{j' \neq j} \rho_{j',t,s,j'} \right) \right. \\
   \left. + \psi(\eta_{j,z_t,s}) - \psi(\sum_{r} \eta_{j,z_t,r}) \right\}
   \]
   where $\psi(\cdot)$ is the digamma function.
7: end while
8: For each $(s_1, \ldots, s_q)$ and $l$, compute the intermediate $\tilde{\theta}_{(s_1, \ldots, s_q), l}$ as follows
   \[
   \tilde{\theta}_{(s_1, \ldots, s_q), l} \propto \exp\left\{ \psi(\nu_l) - \psi(n_l + \mu_l) + \sum_{k=1}^{L-1} \psi(\mu_k) - \psi(n_k + \mu_k) \right\} \\
   + \sum_{c=1}^{C_0} n_{(s_1, \ldots, s_q), c} \left[ \psi(\gamma_{l,c}) - \psi(\sum_{r} \gamma_{l,r}) \right]
   \]
   where $n_{(s_1, \ldots, s_q), c} = T \prod_{j=1}^{Q} \rho_{j,t,s} \mathbb{1}\{y_t = c\}$.
9: For each $l$ and $c$, compute the intermediate $\tilde{\gamma}_{l,c}$ as follows
   \[
   \tilde{\gamma}_{l,c} = \alpha + n_{l,c}
   \]
   where $n_{l,c} = \sum_{s_1, \ldots, s_q} \theta_{(s_1, \ldots, s_q), l} n_{(s_1, \ldots, s_q), c}$.
10: For each $l$, compute the intermediate $\tilde{\nu}_l$ and $\tilde{\mu}_l$ as follows
    \[
    \tilde{\nu}_l = 1 - b + n_l \\
    \tilde{\mu}_l = a + bl + \sum_{k > l} n_k
    \]
    where $n_l = \sum_{s_1, \ldots, s_q} \theta_{(s_1, \ldots, s_q), l}$.
11: For each $j$, $c$ and $s$, compute the intermediate $\tilde{\eta}_{j,c,s}$ as follows
    \[
    \tilde{\eta}_{j,c,s} = \beta_j + n_{j,c,s}
    \]
    where $n_{j,c,s} = T \prod_{j'=1}^{Q} \rho_{j',t,s} \mathbb{1}\{z_{j,t} = c\}$.
12: Update the global variational parameters
    \[
    \xi^{(t-1)} = (1 - \rho_t) \xi^{(t-1)} + \rho_t \tilde{\xi}_t
    \]
    where $\xi$ represents $\gamma$, $\nu$, $\mu$, $\theta$, $\eta$.
13: end while
The exogenous variable $\theta_t$ itself is binary and is generated by a 1st order Markov chain with the following transition probability matrix:

$$
\begin{bmatrix}
0.8 & 0.20 \\
0.1 & 0.90
\end{bmatrix}
$$

The transition probabilities for $y_t$ are listed in Table 3.5.

<table>
<thead>
<tr>
<th>$y_{t-1}$</th>
<th>$y_{t-2}$</th>
<th>$y_{t-5}$</th>
<th>$\theta_t$</th>
<th>$p(y_t = 1)$</th>
<th>$p(y_t = 0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.25</td>
<td>0.75</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.35</td>
<td>0.65</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.25</td>
<td>0.75</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.15</td>
<td>0.85</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.15</td>
<td>0.85</td>
</tr>
</tbody>
</table>

To perform the stochastic variational inference, 4000 samples are collected for training and 1000 samples for monitoring. We evaluate how well a model fits the data with log predictive probability, which is basically the logarithm of the
conditional probability of monitoring data given the training data. The higher log predictive probability is, the better a model fits the data.

Fig. 3.6 shows the log predictive probability as a function of time for both the stochastic variational inference algorithm and Gibbs sampling algorithm. Because the stochastic variational inference algorithm can not automatically identifies the significant predictors, we assume that we know the underlying structure and set $k = [2, 2, 1, 1, 2, 1, 2]$ in both the stochastic variational inference algorithm and the Gibbs sampling algorithm for fair comparison. Mover, the batch size is set to be 10 and the forgetting rate is set to be 0.6 in the stochastic variational inference algorithm. It is seen that the stochastic variational inference algorithm is orders of magnitude faster than the Gibbs sampling algorithm.

![Figure 3.6. Comparison of Stochastic variational inference to Gibbs sampling](image)

The stochastic variational inference algorithm is sensitive to the forgetting rate and the batch size. In Fig. 3.7, the batch size is fixed at 1 and the log predictive probability profiles are presented with different forgetting rates. It is seen that setting the forgetting rate to be 0.6 yields the best performance.

In Fig. 3.8, the forgetting rate is fixed at 0.6 and the log predictive probability profiles are presented with different batch sizes. It is seen that setting the batch size to be 10 achieves the best performance.
**Figure 3.7.** Log predictive probability with different decay rates

**Figure 3.8.** Log predictive probability with different batch size
Chapter 4

Causality Inference

Modeling and decision-making in complex dynamical systems (e.g., human brain [55], macro-economy [56] and distributed physical processes [57]) 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 causal inference and accurate prediction.

In many dynamical systems, the characteristic time of physical process 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 [58]. Although several methods have been proposed for prediction based on fusion of heterogeneous time series (e.g., [59–61]), they lack a coherent probabilistic interpretation and may not 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., [62–64]) have been proposed for causal inference-making; among these techniques, the concept of causality introduced by Granger [65] hereafter called Granger causality, is ap-
parently one of the most widely used in time series analysis. 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 chapter is to use the Bayesian nonparametric model developed
in Chapter 2 for the causal inference of categorical time series and the fusion of
correlated information from heterogeneous sources (e.g., sensors of possibly different
modalities). From this perspective, major contributions of the paper are delineated
as follows:

- Extension of Granger causality [65] to categorical time series in the setting of
  Bayes factor analysis [41].
- Testing of the underlying algorithm by numerical simulation as well as with
  public economics data to infer the causal relationship between two categorical
time series.

This chapter is organized as follows. Section 4.1 introduces the concept of
Granger causality and develops the model. Section 4.2 presents how the Bayes
factor analysis can be used for causal inference. The underlying algorithms are
then tested with simulation data and public economics data in Section 4.3 and in
Section 4.4 respectively.

4.1 Granger Causality

This section introduces the concept of Granger causality and the corresponding
regression model.

Definition 4.1.1. (Granger Causality) Let \( \{y_t\}_{t=1}^T \) and \( \{\theta_t\}_{t=1}^T \) be two (statistically)
stationary categorical time series. Then, the variable \( \theta \) Granger-causes the variable
\( y \) if the past values of \( \theta \) contain statistically significant information for predictions
of \( y \) besides those contained in the past values of \( y \). Similarly, \( y \) Granger-causes \( \theta \)
if the past values of \( y \) contain statistically significant information for predictions of
\( \theta \) besides those contained in the past values of \( \theta \).
Remark 4.1.1. There are the following four types of Granger causality relationship between $\theta$ and $y$:

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

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

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

(4.1)

where the first $D_y$ predictors represent the time-lags of variable $y$; and the next $D_\theta$ predictors represent the time-lags of variable $\theta$. In the sequel, for simplicity of notations, predictors $z_t \equiv (z_{1,t}, \ldots, z_{q,t})$ are substituted for $(y_{t-1}, \ldots, y_{t-D_y}, \theta_{t-1}, \ldots, \theta_{t-D_\theta})$.

Remark 4.1.2. If the explanatory power of $\theta_{t-1}, \ldots, \theta_{t-D_\theta}$ to the regression is significant, then the null hypothesis (that $\theta$ does not Granger-cause $y$) is rejected and the alternative hypothesis (that $\theta$ Granger-causes $y$) is accepted. Hypothesis tests on the significance of time-lags are elaborated later in Eq. (4.2).

Remark 4.1.3. If $y$ and $\theta$ 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.

4.2 Causality Test

This section describes hypothesis testing on the significance of a set of predictors to the regression in Eq. (4.1). It can be used for making the causal inference to provide a better utilization of computational resources for sequential classification by excluding the unimportant predictors. As mentioned earlier, a particular predictor $z_j$ is important if and only if the number of clusters $\tilde{k}_j$ formed by the 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 Bayesian tests for the hypothesis described above, one only needs to compute the Bayes factor 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

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

where $y \equiv \{y_t\}_{t=1}^T$ and $z \equiv \{z_t\}_{t=1}^T$; and $p(H_0 | y, z)$ and $p(H_1 | y, z)$ are numerically computed as the proportions of samples in which the $\tilde{k}_j$'s 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:

$$p(\tilde{k}_j = 1) = \sum_{k=1}^{C_j} p(k_j = k) \sum_{l=1}^{k} p(x_{j,t} = l \forall t | k_j = k)$$

$$= \left( \prod_{r=1}^{C_j} \gamma_j^{(n_{j,c})} \right) \left( \sum_{k=1}^{C_j} \frac{p(k_j = k)k}{\prod_{l=1}^{C_j} (k^{(n_{j,c})})} \right)$$

Specifically, to test whether $\theta$ Granger-causes $y$, it is only necessary to choose $\Lambda = \{D_1 + 1, \ldots, q\}$.

### 4.3 Numerical Example

This section tests the ability of the proposed method to infer the casual relationships between two categorical time series. Here the underlying model for data generation is assumed to be known and is used as the ground truth for performance evaluation. The data generation process is described next.

In this numerical experiment, there are two sequences of binary symbols $y_t$ and $\theta_t$. Symbol sequences $y_t$ are generated from a Markov model $p(y_t | y_{t-1}, y_{t-3}, y_{t-4})$, where only the time-lags $y_{t-1}, y_{t-2}, y_{t-5}$ are important predictors. Symbol sequences $\theta_t$ are generated from another Markov model $p(\theta_t | \theta_{t-1}, \theta_{t-2}, y_{t-1}, y_{t-3})$, 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 $\theta$ but not the other way around because $y$ only depends on its own past. Table 4.1 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 4.2 lists the transition probabilities for $\theta_t$, where the predictors are $y_{t-1}, y_{t-3}, \theta_{t-1},$ and $\theta_{t-2}$ only.

To estimate the regression model in Eq. (4.1) with the parameter $T = 1,005$,
samples of \( \{ y_t \}_{t=1}^{1005} \) and \( \{ \theta_t \}_{t=1}^{1005} \) are collected simultaneously. Based on the assertion that \( y_t-D \) and \( \theta_t-D \) are not important for predicting \( y_t \) and \( \theta_t \) if \( D \) is greater than 5, the predictors for both \( y_t \) and \( \theta_t \) are set as follows:

\[
\mathbf{z}_t \equiv (y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}, \theta_{t-1}, \theta_{t-2}, \theta_{t-3}, \theta_{t-4}, \theta_{t-5})
\] (4.3)

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

To compute posteriors using Algorithm 1, \( \mu_j \) is assigned to be 1 for \( j = 1, \ldots, 10 \). After the initial 200,000 samples are discarded as a burn-in period, the remaining 50,000 after-burn-in samples are downsampled by taking every 5\(^{th} \) sample to reduce the autocorrelation. Figure 4.1 summarizes the results, in which Fig. 4.1(a) exhibits the log-likelihood of the model for 10,000 iterations and Figure 4.1(b) illustrates the ability of the proposed method to identify the important predictors. In this case, the key predictors are 1, 3 and 4, and the resulting prediction coincides with the ground truth: \( y_{t-1}, y_{t-3} \) and \( y_{t-4} \). Figure 4.1(c) shows relative frequency of how many predictors are important. The proposed method also leads to parsimonious representations as shown in Figure 4.1(d) and Figure 4.1(e). As discussed in Subsection 2.4.2, 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. Referring to [41], Figure 4.1(f) illustrates the Bayes factors.
Table 4.2. Transition probabilities for $\theta_t$

<table>
<thead>
<tr>
<th>$y_{t-1}$</th>
<th>$y_{t-3}$</th>
<th>$\theta_{t-1}$</th>
<th>$\theta_{t-2}$</th>
<th>$p(\theta_t = 1)$</th>
<th>$p(\theta_t = 0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.40</td>
<td>0.60</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.65</td>
<td>0.35</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.70</td>
<td>0.30</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.40</td>
<td>0.60</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.47</td>
<td>0.53</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.69</td>
<td>0.31</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.45</td>
<td>0.55</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.75</td>
<td>0.25</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.30</td>
<td>0.70</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.75</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.66</td>
<td>0.34</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.65</td>
<td>0.35</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.20</td>
<td>0.80</td>
</tr>
</tbody>
</table>

as computed in Subsection 4.2 for different predictors. For example, Bayes factor $BF_{10}$ in Eq. (4.2) can be interpreted as the evidence against $H_0$, and a threshold of $t = 20$ indicates that this evidence is strong. Furthermore, having $BF_{10} > 150$ indicates very strong evidence against $H_0$ [41]. If the inclusion proportion in Figure 4.1(b) is equal to 1, then the corresponding Bayes factor in Figure 4.1(f) tends to infinity, as it is the case for the predictors 1, 3 and 4.

Similarly, Figure 4.2(a) exhibits the log-likelihood of the model for 10,000 iterations and Figure 4.2(b) illustrates the ability of the proposed method to identify the important predictors. In this case, the important predictors are 1, 3, 6 and 7, and the resulting predictions coincide with the ground truth $y_{t-1}, y_{t-3}, \theta_{t-1}, \theta_{t-2}$. Figure 4.2(c) shows the relative frequency of how many predictors are important.
Figure 4.1. Gibbs sampling results of numerical example for prediction of $y$
The proposed method also leads to parsimonious representations as shown in Figures 4.2(d) and 4.2(e). As discussed in Subsection ??, the tensor $\lambda_{s_1...s_q}(y_t)$ has much more components than required but they can be clustered nonparametrically to significantly reduce the needed number of components. Figure 4.2(f) illustrates the Bayes factors as computed in Subsection 3.1.2 for different predictors. Bayes factor $BF_{10}$ in Eq. (4.2) can be interpreted as the evidence against $H_0$, and a threshold of $t = 20$ indicates that this evidence is strong [41]. Furthermore, $BF_{10} > 150$ indicates very strong evidence against $H_0$ [41]. If the inclusion proportions in Figure 4.2(b) equal to 1, then the corresponding Bayes factors in Figure 4.2(f) tend to infinity, as it is the case for the predictors 1, 3, 6 and 7.

In addition to correctly identifying the model structure, the proposed method estimates the transition probabilities. Figure 4.3 illustrates two cases from Table 4.1 and Table 4.2. Setting $y_{t-1} = 0$, $y_{t-3} = 1$, and $y_{t-4} = 0$, it follows from Table 4.1 that the true transition probability of $(y_t = 1)$ is 0.70. Similarly, setting $y_{t-1} = 1$, $y_{t-3} = 0$, $\theta_{t-1} = 1$, and $\theta_{t-2} = 0$, it follows from Table 4.2 that the true transition probability of $(y_t = 1)$ is 0.50. In Figure 4.3, the estimated transition probability per sample is obtained with the running mean as well as for 95% and 5% percentiles. It is seen in both plots of Figure 4.3 that the running mean is very close to the true transition probability. With limited data, the proposed method not only estimates the transition probability but also yields the uncertainty quantification in terms of the quantiles.

To identify the causal relationship between $y$ and $\theta$, the Bayes factor analysis is performed as described in Subsection 4.2. The results are summarized in Table 4.3, which show that $y$ Granger-causes $\theta$ but not the other way, which is in line with the ground truth.

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>Bayes Factor $BF_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$ does not Granger-cause $y$</td>
<td>0.43</td>
</tr>
<tr>
<td>$y$ does not Granger-cause $\theta$</td>
<td>Infinity</td>
</tr>
</tbody>
</table>

Table 4.3. Hypothesis test of Granger Causality
Figure 4.2. Gibbs sampling results of numerical example for prediction of $\theta$
4.4 Validation with economics Data

This section validates the nonparametric regression model with (publicly available) real-world economics data. Specifically, monthly data of the consumer price index (CPI) for the United States 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 [66] indicates that low interest rates can cause high inflation, and empirical research [67] 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 [68], 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 $y_t$. Similarly, U.S. LIBOR raw data are used to calculate the monthly difference, and then this difference is converted to a categorical variable by discretizing to quintiles, denoted as $\theta_t$. The entire dataset is used for training the proposed algorithm.

To estimate the regression model in Eq. (4.1), based on the assertion that $y_{t-D_y}$ and $\theta_{t-D_\theta}$ are not important for predicting $y_t$ and $\theta_t$ if both $D_y$ and $D_\theta$ are greater than 6 (i.e., 6 months for both CPI and LIBOR), the predictor for $y_t$ and $\theta_t$ is set as follows:

$$z_t \equiv (y_{t-1}, y_{t-2}, \ldots, y_{t-6}, \theta_{t-1}, \theta_{t-2}, \ldots, \theta_{t-6})$$  \hspace{1cm} (4.4)

To compute the posteriors using Algorithm 1, $\mu_j$ is assigned as $j/2$ for $j = 1, \ldots, 6$ and $(j-6)/6$ for $j = 7, \ldots, 12$. After the initial 100,000 samples are discarded as a burn-in period, the remaining 50,000 after-burn-in samples are downsampled by taking every 5th sample to reduce the autocorrelation. Figures 4.4 and 4.5 respectively summarize the results for $y_t$ and $\theta_t$. These figures have similar characteristics to their counterparts in the numerical example in Section 4.3. The results show that, for $y_t$ or CPI, the important lags are $y_{t-1}, y_{t-2}, y_{t-3}$ and $\theta_{t-1}$. Similarly, for $\theta_t$ 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 the vice versa. This conclusion is further confirmed by Bayes factor analysis as summarized in Table 4.4.

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>Bayes Factor $BF_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>US CPI does not Granger-cause LIBOR</td>
<td>7.29</td>
</tr>
<tr>
<td>LIBOR does not Granger-cause US CPI</td>
<td>Infinity</td>
</tr>
</tbody>
</table>
Figure 4.4. Results of Gibbs sampling of economics dataset for prediction of $y$
Figure 4.5. Results of Gibbs sampling of economics dataset for prediction of \( \theta \)
Sequential Classification: Detection of Combustion Instabilities

In previous chapters, we have developed a Bayesian nonparametric model for the estimation of conditional probability tensors and its statistical inference methods. In this chapter, we propose a sequential classification algorithm for complex dynamical systems and its application in the detection of combustion instabilities.

Combustion instabilities accrue from nonlinear interactions between the unsteady heat release and acoustics in the confined chamber of a combustor [69] [70] [71]. Consequently, if the self-sustained pressure oscillations have high amplitudes, the performance and operational life of machineries (e.g., gas turbine engines), which use such combustors, could be adversely affected. Therefore, it is imperative to appropriately design and operate combustors to ensure timely detection and mitigation of combustion instabilities, which may require model-based and/or dynamic data-driven analysis involving analyses of time series of pressure oscillations and flame images.

The difficulties in handling the complex nonlinear dynamics, as encountered in the analysis of combustion instabilities, often limit the applications of physics-based modeling tools for anomaly detection and decision & control of combustion dynamics at different operating conditions. To this end, there has been much interest in early detection of combustion instabilities from the perspectives of dynamic data-driven application systems (DDDAS) [72,73]. For example, Nair et al. [74] have used the local flow test and 0-1 chaos test to study the chaotic
structure of pressure time series, and concluded that the pressure measurements
during stable operation in combustors undergo deterministic chaos and gradually
relax their chaotic behavior when the system approaches an unstable condition. It
is also claimed that low-amplitude irregular pressure fluctuations therein possess
multifractal structures, which may contain valuable prognostic information for
early detection of combustion instabilities. Similarly, Gianni et al. [75] have used
a topologically invariant index to recognize the transition mechanism leading to
combustion instabilities, which is shown to be an early precursor. Other researchers
believe that time series acquired in combustors can be modeled as a random process,
such as a Markov chain, whose model structure may reflect the physical nature
of the combustion process. Much research efforts have been expended on Markov
chain modeling, and several popular strategies are listed in [76] [26] [77] [78].

Implementations of DDDAS [72] include fault detection and estimation in
thermal pulse combustors [79] and flame lean-blowout predictions [80] in laboratory-
scale combustors using symbolic time series analysis. Especially Sarkar et al. [6]
reported an information-theoretic state-splitting and state-merging algorithm to
model flame lean-blowout phenomena in the framework of a special class of prob-
abilistic finite state automata (PFSA), called $D$-Markov machines [4, 81], whose
entropy rate may also be used to detect the combustion instabilities. However, this
tree-based algorithm lacks a coherent probabilistic interpretation and may not able
to accommodate more general temporal interactions and take operating conditions
into consideration. Moreover, the task of parameter estimation becomes difficult
even for moderately high depth (e.g., $D > 1$) in $D$-Markov machines due to the
paucity of time series data. As for the detection procedure, this method is based
on an empirical threshold rather than the statistical detection theory. This method
may not be sequentially implementable, thus possibly limiting its use in real-time
detection of thermo-acoustic instabilities.

Compared with exiting data-driven methods for detection of combustion instabil-
ities, the proposed Bayesian nonparametric approach has the following advantages.

- Different information sources can be modeled in a coherent way and operating
  conditions are taken into consideration.

- Significant predictors can be automatically identified base on the varying
dimension mechanism, which saves the efforts of model selection.
The proposed model is parsimonious and shares the statistical strength rendering it very data efficient and suitable for real time implementation.

From the perspectives of active control of combustion instabilities (e.g., [58]), the proposed method may serve as a statistical filter to predict the system state with a high level of confidence, and thus potentially improve the performance of active controllers by reducing the (potentially destabilizing) delay in the feedback loop.

This chapter is organized as follows. Section 5.1 presents the sequential classification algorithm for complex dynamical systems based on the Bayesian nonparametric model. Section 5.2 describes the test apparatus and the procedures of data preprocessing. In Section 5.3, pressure data and chemiluminescence data are fused for sequential classification of combustion instabilities at a fixed operating condition. In Section 5.4, a spatio-temporal model is developed for classification of combustion instabilities under varying operating conditions.

5.1 Sequential Classification

This section presents the sequential classification algorithm for complex dynamical systems based on the Bayesian nonparametric model developed in Chapter 2. It is 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 $(i)D_T = \{(i)y_t, (i)z_t\}_{t=1}^{T_i}$ 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 $(i)D_T$ is used to compute posterior samples $\{(i)\phi, (i)\lambda, (i)\omega\}_{n=1}^{N}$ for each class $i$, as described in Algorithm 1. In the testing phase, a test data set $D_T$ needs to be classified to belong to one of the $C$ classes. For this purpose, the conditional probability $p(D_T \mid (i)D_T)$ is computed by the following equations:

$$p(D_T \mid (i)D_T) = \prod_{t=1}^{T} p(y_t \mid z_t; (i)D_T) \quad (5.1)$$

$$p(y_t \mid z_t; (i)D_T) \approx \frac{1}{N} \sum_{n=1}^{N} \left( \sum_{s_1=1}^{k_1} \ldots \sum_{s_q=1}^{k_q} \prod_{j=1}^{q} \omega_{(i)j, s_1, \ldots, s_q}(y_t) \prod_{j=1}^{q} \omega^{(i)j}(z_{j,t}) \right) \quad (5.2)$$

64
Based on the conditional probability $p(D_T \mid (^iD_T_i))$, the posterior probability of the observed data $D_T$ belonging to the class $i$ is denoted as $p(C_i \mid D_T)$ and is given as:

$$p(C_i \mid D_T) = \frac{p(D_T \mid (^iD_T_i)p(C_i)}{\sum_{r=1}^C p(D_T \mid (^rD_T_r)p(C_r)}$$

(5.3)

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

$$D_{\text{class}} = \arg \max_i p(C_i \mid D_T)$$

(5.4)

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 $p(D_T \mid (^iD_T_i)$ could be evaluated sequentially as shown in Eq. (5.1). For fast implementation, the values of $p(y_t \mid z_t; (^iD_T_i)$ in Eq. (5.2) can be precomputed and stored for different values of $(y_t, z_t)$.

For binary classification, the likelihood ratio test [82] is constructed as follows:

$$\frac{p(D_T \mid (^1D_T_1))}{p(D_T \mid (^0D_T_0))} \geq \Theta$$

(5.5)

where $\Theta$ is the threshold. One criterion to choose the threshold $\Theta$ is the receiver operating characteristic (ROC). The ROC curve is obtained by varying $\Theta$, and provides a trade-off between the probability of detection $p_D = p(\text{decide } 1 \mid 1 \text{ is true})$ and the probability of false alarms $p_F = p(\text{decide } 1 \mid 0 \text{ is true})$. 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 $\Theta$.

5.2 Experimental Data Collection

5.2.1 Experimental Apparatus

The test apparatus is built upon a swirl-stabilized, lean-premixed, laboratory-scale combustor that has been used to perform the experimental investigation. Figure 5.1 presents a schematic diagram of the combustor apparatus [83] 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 5.1. Schematic diagram of the combustor apparatus

<table>
<thead>
<tr>
<th>Table 5.1. Operating conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
</tr>
<tr>
<td>Variables</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Fixed Conditions</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

High pressure air is delivered from a compressor system after passing through filters that remove any spurious liquid or solid particles. The air supply pressure is set to 1.34 MPa (∼ 180 psig) using a dome pressure regulator. The air is pre-heated to a maximum temperature of 250°C by an 88 KW electric heater. The fuel for this study is natural gas (approximately 95% methane), which is supplied to the system at a pressure of 1.48 MPa (∼ 200 psig). The flow rates of the air and natural gas are measured by thermal mass flow meters. The inlet velocity is set at 40 m/s and the combustor length is fixed at 0.10 m (∼ 25 inches). The desired equivalence ratio and the percentage of pilot fuel are set by adjusting these flow rates with needle valves. By varying equivalence ratio and the percentage of pilot fuel as
listed in Table 5.1, there are total 62 operating conditions. Under each operating condition, 8 seconds of pressure and chemiluminescence measurements have been collected at the sampling rate of 8192 Hz. Stable and/or unstable operating modes are recorded along with each time series data.

5.2.2 Data preprocessing

Oversampling is often preferred for data collection from engineering perspectives and may mask the true nature of system dynamics when real-valued time series data are symbolized. To avoid such a problem, the pressure measurements from combustors are first downsampled by a factor that is representative of a typical lag, which is the first minimum of the average mutual information plot [84], as demonstrated in Fig. 5.2. Then the time series data are symbolized through maximum entropy partitioning [85] with a ternary alphabet \( \Sigma = \{1, 2, 3\} \) for both stable and unstable cases, where the ground truth is decided such that the root mean square value of pressure greater than 0.483 kPa (0.07 psi) indicates an unstable situation.

![Figure 5.2. Profile of the average mutual information](image)

5.3 Information Fusion

In this section, pressure data and chemiluminescence data are fused for sequential classification of combustion instabilities at a fixed operating condition base on the Bayesian nonparametric model developed in Chapter 2. To train the Bayesian
nonparametric model, 500 samples have been used from the downsampled quantized pressure time series under both stable and unstable conditions. The memory $D$ of each of $y_t$ and $\theta_t$ in the experimental data is observed to be largely limited to 5 for both stable and unstable time-series data. Hence, the predictors of $y_t$ or $\theta_t$ are $z_t \equiv (y_{t-1}, y_{t-2}, \ldots, y_{t-5}, \theta_{t-1}, \theta_{t-2}, \ldots, \theta_{t-5})$ and the corresponding regression model is hereafter referred to as 'full order model'. Since $y_t$ and $\theta_t$ has three categories, it follows that $C_y = C_\theta = 3$.

To compute posteriors, as in Algorithm 1, the values $[1.0, 1.5, 2.0, 2.5, 3.0]$ are assigned to $\mu_j$ for $j = 1, \ldots, 10$. After discarding 200,000 data points during the initial burn-in period, the remaining 50,000 samples are downsampled by taking every $5^{th}$ data point to reduce the autocorrelation. Gibbs sampling results of pressure data are summarized as $p(y_t \mid y_{t-1}, \ldots, y_{t-5}, \theta_{t-1}, \ldots, \theta_{t-5})$ in Figures 5.3(a)(b) for a stable mode and in Figures 5.3(c)(d) for an unstable mode. Similarly, Gibbs sampling results of chemiluminescence data are summarized as $p(\theta_t \mid y_{t-1}, \ldots, y_{t-5}, \theta_{t-1}, \ldots, \theta_{t-5})$ in Figures 5.4(a)(b) for a stable mode and in Figures 5.4(c)(d) for an unstable mode.

Figures 5.3(a)(c) and Figures 5.4(a)(c) show the profiles of log likelihood with different iterations for pressure and chemiluminescence data under stable and unstable conditions, respectively; Similarly, Fig. 5.3(b)(d), Fig. 5.4(b)(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

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

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

$$y_{t-1}, y_{t-3}, y_{t-4}, y_{t-5}, \theta_{t-1}, \theta_{t-2}, \theta_{t-3}, \theta_{t-4} \text{ and } \theta_{t-5}$$
Figure 5.3. Gibbs sampling of pressure data

while those for unstable chemiluminescence data are identified as

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

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 \( z_t \equiv (y_{t-1}, y_{t-2}, \ldots, y_{t-5}) \). The stable and unstable cases are
shown in Figures 5.5(a)(b) and Figures 5.5(c)(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.
5.3.1 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 4.2. The results are summarized in Table 5.2, 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 interactions between pressure and chemiluminescence measurements could be caused by a third unknown physical quantity, the exploration of which is a topic of future research.

5.3.2 Sequential Classification

To evaluate the performance of sequential classification for identification of thermoacoustic instabilities, 100 instances of 50 samples, which are different from training samples, have been selected from downsampled quantized pressure measurements.
Figure 5.5. Gibbs sampling of reduced order model

Table 5.2. Hypothesis test of Granger Causality

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>Operating Condition</th>
<th>$BF_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$ does not Granger-cause $y$</td>
<td>Stable</td>
<td>Infinity</td>
</tr>
<tr>
<td>$y$ does not Granger-cause $\theta$</td>
<td>Stable</td>
<td>Infinity</td>
</tr>
<tr>
<td>$\theta$ does not Granger-cause $y$</td>
<td>Unstable</td>
<td>Infinity</td>
</tr>
<tr>
<td>$y$ does not Granger-cause $\theta$</td>
<td>Unstable</td>
<td>Infinity</td>
</tr>
</tbody>
</table>

from both stable and unstable modes. Figure 5.6 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., Fig. 5.6(a)) uses the full order model, and the bottom plate (i.e., Fig. 5.6(b)) uses the reduced-order model for the same test data sequence. While the observed test sequence is correctly classified by both models, the reduced-order model is slower than the full-order model that contains more information.
Figure 5.7 shows a family of receiver operating characteristic (ROC) curves for the proposed detection algorithm with varying lengths of 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 5.6. Posterior probabilities using different models
5.4 Spatio-temporal Model

In this section, a spatio-temporal model is developed to accommodate the effect of operating conditions. To train the Bayesian nonparametric model, 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 \( \mathbf{z}_t \equiv (y_{t-1}, y_{t-2}, \ldots, y_{t-8}, \psi_1, \psi_2) \), where \( \psi_1 \) and \( \psi_2 \) 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., \( \psi_1 \) and \( \psi_2 \)) model the spatial part, which form a spatio-temporal model of combustion data. Since \( y_t \) has 3 categories, it follows that \( C_0 = C_1 = \cdots = C_8 = 3 \). Similarly, \( \psi_1 \) can take 4 different values and \( \psi_2 \) can take 19 values, which implies \( C_9 = 4 \) and \( C_{10} = 19 \) respectively. To compute posteriors, we assign \( j/2 \) to \( \mu_j \) for \( j = 1, \ldots, 8 \) and 0 to \( \mu_j \) 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 5\textsuperscript{th} data point to reduce the autocorrelation.
Figure 5.8 and Figure 5.9 summarize the Gibbs sampling results for stable and unstable cases, respectively. Figure 5.8(a) and Figure 5.9(a) illustrate the log likelihood of the model with different iterations. Figure 5.8(b) and Figure 5.9(b) illustrate the proposed method’s ability to identify the important predictors of the regression model. Figure 5.8(c) and Figure 5.9(c) shows relative frequency of how many predictors are important. The proposed method also leads to parsimonious representations as shown in Figure 5.8(d)(e) and Figure 5.9(d)(e). By comparing the Figure 5.8(b) and Figure 5.9(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 [86]. 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. Figure 5.8(f) and Figure 5.9(f) illustrate the Bayes factors of the different predictors of the stable and unstable case.

To evaluate the performance of sequential classification for combustion instability, two instances of 50 samples (different from training samples) are selected from downsampling quantized pressure measurements under each operating condition (124 test cases in total for 62 operating conditions). Figure 5.10 shows the posterior probability of each class as a function of the length of the observed data. It is seen in Figure 5.10(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 Figure 5.10(b) the observed sequence is correctly classified as unstable. Figure 5.11 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 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 [87] (which is called an adaptive method in this paper) and with a Markov chain method [88]. In the adaptive method, each row of the transition probability matrix 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 \( \mathcal{M}_i^{(s)} \), and unstable, denoted as \( \mathcal{M}_i^{(u)} \), 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 \( \sim 20 \) 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 is 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 Figure 5.12 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 Figure 5.12(c)), and with Markov chain models. The ROC curves for these methods are plotted in Figure 5.13. It is seen that, when testing with the same dataset, even though the training samples are only small fractions of those used in the 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 Figure 5.13, because they do not have information about different operation conditions.
Figure 5.8. Gibbs sampling results for stable pressure data
Figure 5.9. Gibbs sampling results for unstable pressure data
Figure 5.10. Posterior probabilities for stable and unstable cases. (Class 0 represents stability and Class 1 represents instability.)
Figure 5.11. ROC curves for the proposed method
Figure 5.12. ROC curves for classification using the symbolic dynamics-based adaptive method, with different $D_u$ and $D_s$.
Figure 5.13. ROC curves for performance comparison
Conclusion and Research Vision

This chapter concludes the thesis with summary of results and suggestions for future work.

6.1 Summary of Results

This dissertation on Bayesian nonparametric modeling of categorical data has made contributions to causal inference of categorical time series, information fusion of heterogeneous sources and sequential classification of complex dynamical systems. The merits of the work in this dissertation are presented below.

- Development of a flexible model, containing multiple time-independent spatial variables and time-dependent exogenous variables, which is capable of capturing spatio-temporal characteristics and fusing information from heterogeneous sources. For example, the model can take operating conditions into consideration and infuse the information from both pressure sensors and chemiluminescence sensors for the real time detection of combustion instabilities.

- Generation of a parsimonious model that shirks towards low dimension structures and inherits its statistical strength from different predictors, which enhances the data efficiency and improves estimation accuracy especially for limited data.
• Provision of large-sample properties for general (non-IID) data generating processes under regular conditions.

• Development of a Gibbs sampling algorithm for parameter estimation and hypothesis testing in a unified Bayesian framework, which not only provide the uncertainty quantification of the estimated parameters but also the identification of the underlying model structure.

• Development of a stochastic variational inference algorithm for fast inference and real time implementation. In an numerical example, the stochastic variational inference algorithm is shown to be orders of magnitude faster than the Gibbs sampling algorithm.

• Extension of Granger causality to categorical time series in the setting of Bayes factor analysis, which is validated with both numerical and econometric data.

• It is demonstrated that the proposed method outperforms: (i) various standard machine learning methods such as KNN, SVM, and MLPNN [42] for econometric public data in terms of accuracy and R2 score and (ii) symbolic dynamics-based adaptive method of pattern classification [87,89] and Markov chain method [88] for a real-time application with experimental data from a swirl-stabilized combustor apparatus in terms of area under curve (AUC).

### 6.2 Suggestions for Future Work

#### 6.2.1 Reinforcement Learning for Active Control of Combustion Instabilities

One future research direction is to apply the proposed Bayesian nonparametric model to reinforcement Learning for Active Control of Combustion Instabilities. Passive controllers of combustion instabilities, which are largely restricted to acoustic dampers (e.g. Helmholtz resonators) and variable combustor geometry, have been extensively studied. However, there are significant limitations of these techniques, because of complex designs of actual combustors, geometry modifications are
unreliable in practice. Addition of extra hardware like dampers can significantly increase weight of combustors and thus negatively affect the vehicle performance.

The above facts have led to the study of active control systems that dynamically operate an actuator in order to suppress the combustion instabilities. It has been shown that the pilot flame (i.e., secondary flame) can be used to suppress pressure oscillations with a small fraction of the primary flame energy and low actuation frequency, which may provide a promising solution to active control of combustion instabilities.

The combustion dynamics in such applications are described by coupled nonlinear partial differential equations, which have significant challenges in modeling and thus render model-based controllers unreliable. Therefore, it is proposed to use policy search reinforcement learning algorithms to directly learn an optimal policy for control of the secondary flame from data. Before applying these policy search algorithms to actual combustors, it is logical to first evaluate their performance with simulated and experimental data from a electrically heated Rijke tube. The Rijke tube has two electric heaters to emulate primary flame and secondary flame respectively, which can capture the main acoustic and heat characteristics of actual combustors. A schematic plot of the Rijke tube is shown in Fig. 6.1.

![Figure 6.1. Schematic plot of the electrically heated Rijke tube.](image)

The dynamics of Rijke tube are described by the following partial differential
equations.

\[ \gamma M \frac{\partial u'}{\partial t} + \frac{\partial p'}{\partial x} = 0 \]  \hspace{1cm} (6.1)

\[ \frac{\partial p'}{\partial t} + \gamma M \frac{\partial u'}{\partial x} = 0 \]  \hspace{1cm} (6.2)

\[ \dot{Q}' = \frac{2L_w(T_w - T)}{S\sqrt{3}} \sqrt{\pi \lambda C_v \bar{\rho} d_w} \left[ \sqrt{\frac{u_0}{3} + u'_f(t - \tau)} - \sqrt{\frac{u_0}{3}} \right] \delta(x - x_f) \]  \hspace{1cm} (6.3)

where \( u' \), \( p' \) and \( \dot{Q}' \) are the fluctuation of velocity, pressure and heat release rate respectively. To build the simulator, we approximate the above equations using Galerkin expansion, which renders the following reduced order model.

\[ \ddot{\eta} + 2\zeta_j \omega_j \dot{\eta} + k^2_j \eta_j = -\frac{2K_1}{\gamma M} j\pi \left[ \sqrt{\frac{1}{3} + u'_f(t - \tau_1)} - \sqrt{\frac{1}{3}} \right] \sin(j\pi x_{f_1}) \]  \hspace{1cm} (6.4)

\[ -\frac{2K_2}{\gamma M} j\pi \left[ \sqrt{\frac{1}{3} + u'_f(t - \tau_2)} - \sqrt{\frac{1}{3}} \right] \sin(j\pi x_{f_2}) \]

Reinforcement learning concerns with how agents ought to take action based on the observations received from environment so that the cumulative reward is maximized. In our case, the controller is the agent and the combustor is the environment. The schematic plot of reinforcement learning for active control of combustion instabilities using secondary heater is shown in the Fig 6.2.

The reinforcement learning algorithm we used is trust region policy optimization \([90]\), which is a model free policy search algorithm. It directly models the control policy with a neural network and tries to maximize the lower bound of the cumulative reward for each iteration. In our application, the controller is represented by a gated recurrent neural network. The input of the controller is the pressure time series observed between the current and the previous action. The output of the controller can take 3 values, which corresponds to not change the secondary heater temperature, decrease the secondary heater temperature by 10 K, and increase secondary heater temperature by 10 K respectively. The structure of the controller is shown in Fig. 6.3.

The reward function consists of three terms, penalizing high amplitude of pressure oscillation, frequent switching of actions and high secondary heater temperature respectively.
As we can see from the Fig. 6.4, the average return gradually increases and the entropy of the controller declines steadily, which means the controller becomes more and more confident in the actions it takes.

Fig. 6.5 illustrates the performance of the learned control policy. The first two plots compares the pressure profiles of the combustor with and without control. It can be seen that the learned control policy can decrease the amplitude from 1000 Pa to 20 Pa within a short time. The last two plots show the secondary heater
temperature profile and the course of actions the controller takes. It is clear that the energy needed and the actuation frequency is relatively low and is achievable in practice.

In the future, the features extracted from the observations by the Bayesian nonparametric model developed in this thesis can be used as the input of the reinforcement learning algorithm. This will potentially reduce the computation cost and enhance the performance.

### 6.2.2 Collaborative Filtering for Recommend Systems

Collaborative filtering is a class of machine learning methods for predicting a user’s preference or rating of an item, based on his/her previous preferences or ratings and decisions made by similar users. Generally speaking, there are two types of collaborative filtering: memory-based collaborative filtering and model-based collaborative filtering. Memory-based collaborative filtering provides recommendations based on the similarities between users or items while model-based collaborative filtering learns a model from historical data and then uses the model to predict preferences of users. Because of the poor performance of memory-based collaborative filtering on real-life large-scale and sparse data, model-based collaborative filtering has attracted increasing attention along with the recent boom of e-commerce and social
Figure 6.5. Flow chart of the reinforcement learning for the active control of combustion instabilities.

network systems. A variety of machine learning methods have been used for the development of model-based collaborative filtering, such as Bayesian nonparametrics [91], restricted Boltzmann machine [92], and Neural Autoregressive Distribution Estimator [93].

The proposed Bayesian nonparametric model in this thesis can be directly applied to collaborative filtering with users and items being the predictors and ratings being the response variables. However, the assumption that each row of the mixture probability matrix \( \omega^{(j)} \) is independent as described in Eq. 2.40 is not appropriate because the number of users and items are often very large, and may yield poor performance due to the paucity of data. To alleviate this issue, the the rows of \( \omega^{(j)} \) can be clustered by the Bayesian nonparametric priors the same as the tensor \( \lambda \). The resulting Bayesian network is shown in Fig. 6.6. In the future, the corresponding statistical inference algorithm should be developed and validated with benchmark datasets like MovieLens 1M, MovieLens 10M, and Netflix datasets.
Figure 6.6. Graphical representation of the model for collaborative filtering. 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.
Appendix A

Collapsed Conditional of $k$

This appendix provides the technical details of deriving the collapsed conditional $p(k_j|x_j, z_j)$ in Section 3.1. We use the generic $p_0$ to denote the prior of $k$. By integrating out $\omega^{(j)}$, we have

$$p(x_j|z_j, k_j) = \int p(x_j|z_j, \omega^{(j)}, k_j)p(\omega^{(j)})d\omega^{(j)}$$

$$= \prod_{r=1}^{C_j} \left\{ \frac{\Gamma(k_j \beta_j) \prod_{l=1}^{k_j} \Gamma(\beta_j + n_{j,r}(l))}{\Gamma(\beta_j) k_j^{k_j} \Gamma(\beta_j + k_j + n_{j,j,r})} \right\}$$

$$= \prod_{r=1}^{C_j} \left\{ \frac{1}{(k_j \beta_j)_{n_{j,r}}} \prod_{l=1}^{\max x_{j,l,r}} \beta_j^{n_{j,r}(l)} \right\}$$

$$= \left\{ \prod_{r=1}^{C_j} \prod_{l=1}^{\max x_{j,l,r}} \beta_j^{n_{j,r}(l)} \right\} \left\{ \prod_{r=1}^{C_j} 1 \right\}$$

(A.1)

where $a^{[b]} = a(a + 1) \ldots (a + b - 1)$ with $a^{[0]} = 1$, $x_{j,r} = \{x_{j,t} : z_{j,t} = r\}$, $n_{j,r} = \sum_t \mathbf{1}\{z_{j,t} = r\}$ represents the count of the $r^{th}$ category of the $j^{th}$ predictor $z_j$ and $n_{j,r}(l) = \sum_t \mathbf{1}\{x_{j,t} = l, z_{j,t} = r\}$ denotes the number of latent allocation variables that associated with the $r^{th}$ category of the $j^{th}$ predictor and instantiated at $l$.

Because $p(k_j|z_j) = p_{0,j}(k_j)$, we have

$$p(x_j|z_j) = \sum_{k_j=\max(x_{j,r})} p(x_j|z_j, k_j)p_{0,j}(k_j)$$

$$= \left\{ \prod_{r=1}^{C_j} \prod_{l=1}^{\max x_{j,l,r}} \beta_j^{n_{j,r}(l)} \right\} U_{n_{j,1}, \ldots, n_{j,C_j}} (\max \mathbf{x}_j)$$

(A.2)
where $U_{n_{j,1},...,n_{j,C_j}}(x) = \sum_{k_j=x}^{C_j} p_{0,j}(k_j) \prod_{r=1}^{C_j} \{\Gamma(k_j \beta_j)/\Gamma(k_j \beta_j + n_{j,r})\}$. This yields a closed form of $p(k_j|x_j, z_j)$ as follows

$$
p(k_j|x_j, z_j) = \frac{p(k_j|z_j)p(x_j|z_j, k_j)}{p(x_j|z_j)} = \frac{p_{0,j}(k_j) \prod_{r=1}^{C_j} \frac{\Gamma(k_j \beta_j)}{\Gamma(k_j \beta_j + n_{j,r})}}{U_{n_{j,1},...,n_{j,C_j}}(\text{max } x_j)} \quad (A.3)
$$

where $k_j = \text{max } x_j, \ldots, C_j$. Because of Stirling’s approximation $\Gamma(n+\alpha)/\Gamma(n) \approx n^\alpha$, for moderately large values of $n_{j,r}$, we have

$$
p(k_j|x_j, z_j) \approx \frac{1}{\sum_{l=\text{max } x_j}^{C_j} p_{0,l}(l) \prod_{r=1}^{C_j} n_{j,r}^{(k_j-l)\beta_j}} \quad (A.4)
$$

This completes the derivation of the collapsed conditional of $k$. 

91
This appendix presents three sampling algorithms for the Dirichlet Process Mixture Model. The first sampling algorithm is designed based on the graphical model described by the Fig. B.1. By integrating out $\pi$ and using the predictive distribution

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig_B1.png}
\caption{Graphical representation of the Dirichlet process mixture model. 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.}
\end{figure}
of the Dirichlet process, we can directly sample $\theta_i$ as follows

$$p(\theta_i|\theta_{-i}, x_i) \propto p(\theta_i|\theta_{-i}) p(x_i|\theta_i)$$

$$= \left[ \frac{1}{n-1+\alpha} \sum_{j \neq i} \delta_{\theta_j}(\theta_i) + \frac{\alpha}{n-1+\alpha} H(\theta_i) \right] p(x_i|\theta_i) \quad (B.1)$$

Although the above algorithm is a valid sampler, it has extremely slow mixing behavior because data points have to be moved from the old cluster to the new cluster one by one when changing the parameter of a cluster or creating a new cluster. The issues of the above sampler can be addressed easily by grouping data points by cluster and generating updates of the cluster parameters given the entire data in the cluster (See Fig. B.2). The resulting algorithm is presented as follows

$$p(z_i = c|z_{-i}, \theta, x_i) \propto \begin{cases} \frac{n_{-i,c}}{n-1+\alpha} p(x_i|\theta_{z_i}) & \exists j \neq i, c = z_j \\ \frac{\alpha}{n-1+\alpha} \int p(x_i|\theta) H(\theta) d\theta & \forall j \neq i, c \neq z_j \end{cases} \quad (B.2)$$

$$p(\theta_k|z, \theta_{-k}, x) \propto \prod_{i: \theta_i = \theta_k} p(x_i|\theta_k) H(\theta_k) \quad (B.3)$$

where $n_{-i,c}$ represents the number of data points belong to the cluster $c$ when the $i^{th}$ data point is excluded.

![Figure B.2.](image)

**Figure B.2.** Equivalent graphical representation of the Dirichlet process mixture model. 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.

The last sampling algorithm is based on the stick breaking representation of
the Dirichlet process, which truncates the Dirichlet process at the $L^{th}$ component

$$V_l | \mathbf{z} \sim \text{Beta}(1 + n_l, \alpha + \sum_{k > l} n_k) \quad \text{for } l < L$$  \hspace{1cm} (B.4)

$$V_L = 1$$  \hspace{1cm} (B.5)

$$\pi_k = V_k \prod_{l=1}^{k-1} V_l$$  \hspace{1cm} (B.6)

$$p(z_i = j | \mathbf{z}, \mathbf{\theta}, x_i) \propto \pi_j p(x_i | \theta_j)$$  \hspace{1cm} (B.7)

$$p(\theta_k | \mathbf{z}, \mathbf{\theta}_{-k}, \mathbf{x}) \propto \prod_{i: \theta_{z_i} = \theta_k} p(x_i | \theta_k) H(\theta_k)$$  \hspace{1cm} (B.8)
Bibliography


[38] M I L L E R , J . W . and M . T . H A R R I S O N ( 2 0 1 5 ) “ M i x t u r e m o d e l s w i t h a p r i o r

[39] V A N D Y K , D . A . and T . P A R K ( 2 0 0 8 ) “ P a r t i a l l y c o l l a s p e d G i b b s s a m p l e r s: T h e o r y a n d methods,” Journal of the American Statistical Association,
103(482), pp. 790–796.

[40] A K A I K E , H . ( 1 9 8 7 ) “ F a c t o r a n a l y s i s a n d A I C ,” Psychometrika, 5 2 ( 3 ), pp.
317–332.


[42] B I S H O P , C . ( 2 0 0 6 ) P a t t e r n A n a l y s i s a n d M a c h i n e L e a r n i n g , S p r i n g e r , N e w
York, NY, USA.

 exponential families, and variational inference,” Foundations and Trends® in
Machine Learning, 1(1–2), pp. 1–305.

[44] W R I G H T , S . J . ( 2 0 1 5 ) “ C o o r d i n a t e d e s c e n t a l g o r i t h m s ,” Mathematical Programming,
151(1), pp. 3–34.

[45] C A U W E N B E R G H S , G . ( 1 9 9 3 ) “ A f a s t s t o c h a s t i c e r r o r-d e s c e n t a l g o r i t h m f o r
supervised learning and optimization,” in Advances in neural information
processing systems, pp. 244–251.

[46] H E N S M A N , J . , M . R A T T R A Y , a n d N . D . L A W R E N C E ( 2 0 1 2 ) “ F a s t v a r i a t i o n a l
inference in the conjugate exponential family,” in Advances in Neural

[47] B L E I , D . M . , A . Y . N G , a n d M . I . J O R D A N ( 2 0 0 3 ) “ L a t e n t d i r i c h l e t a l l o c a t i o n ,”
Journal of machine Learning research, 3(Jan), pp. 993–1022.

[48] E D D Y , S . R . ( 1 9 9 6 ) “ H i d d e n m a r k o v m o d e l s ,” Current opinion in structural

[49] J O Y C E , J . M . ( 2 0 1 1 ) “ K u l l b a c k-l e i b l e r d i v e r g e n c e ,” in International Encyclopedia

inference: A review for statisticians,” Journal of the American Statistical
Association, (just-accepted).

[51] A M A R I , S . - I . ( 1 9 9 8 ) “ N a t u r a l g r a d i e n t w o r k s e f f i c i e n t l y i n l e a r n i n g ,” Neural
computation, 10(2), pp. 251–276.


[82] Poor, H. V. (1994) An introduction to signal detection and estimation, Springer-Verlag, New York, NY, USA.


Vita

Sihan Xiong

Sihan Xiong received the bachelor’s degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2012. He has obtained the master’s degree in Mathematics in 2015 and is currently pursuing the Ph.D. degree in mechanical engineering with Pennsylvania State University, State college, the United States. His current research focuses on Bayesian nonparametrics, deep learning and reinforcement learning as well as their applications in dynamical systems.