ABSTRACT

TANG, YONGQIANG. Dirichlet Process Mixture Models for Markov processes.
(Under the direction of Dr. Subhashis Ghosal.)

Prediction of the future observations is an important practical issue for statisticians. When the data can be viewed as exchangeable, de Finneti's theorem concludes that, conditionally, the data can be modeled as independent and identically distributed (i.i.d.). The predictive distribution of the future observations given the present data is then given by the posterior expectation of the underlying density function given the observations. The Dirichlet process mixture of normal densities has been successfully used as a prior in the Bayesian density estimation problem. However, when the data arise over time, exchangeability, and therefore the conditional i.i.d. structure in the data is questionable. A conditional Markov model may be thought of as a more general, yet having sufficiently rich structure suitable for handling such data. The predictive density of the future observation is then given by the posterior expectation of the transition density given the observations. We propose a Dirichlet process mixture prior for the problem of Bayesian estimation of transition density. Appropriate Markov chain Monte Carlo (MCMC) algorithm for the computation of posterior expectation will be discussed. Because of an inherent non-conjugacy in the model, usual Gibbs sampling procedure used for the density estimation problem is hard to implement. We propose using the recently proposed "no-gaps algorithm" to overcome the difficulty. When the Markov model holds good, we show the consistency of the Bayes procedures in appropriate topologies by constructing appropriate uniformly exponentially consistent tests and extending the idea of Schwartz (1965) to Markov processes. Numerical examples show excellent agreement between asymptotic theory and the finite sample behavior of the posterior distribution.

KEY WORDS: Dirichlet mixture, Markov process, no-gaps algorithm, Poisson equation, posterior consistency, sup-L_1 distance, time series, uniformly exponentially consistent tests.

DIRICHLET PROCESS MIXTURE MODELS FOR MARKOV PROCESSES

by

Yongqiang Tang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

STATISTICS

Raleigh

2003

APPROVED BY:

Professor Subhashis Ghosal Chair of Advisory Committee	Professor Bibhuti Bhattacharyya
Professor Peter Bloomfield	Professor Anastasios Tsiatis

To my parents

Biography

Yongqiang Tang was born in Jiangsu, China on February 20, 1977. Yongqiang won championship in the National Mathematical Competition in his high school. He was exempt from the college entrance examination and entered the University of Science and Technology of China in September 1994. He got his Bachelor of Science degree in Biological Sciences in July 1999. He continued to pursue a Ph.D. degree in Statistics at North Carolina State University (NCSU) since August 1999. His interest at NCSU has focused on Biostatistics, Statistical Genetics and Genomics, and time series. He has interned at the Bioinformatics Research Center, the Center for Applied Aquatic Ecology at NCSU and Duke University Medical Center during his study at NCSU.

Acknowledgments

I would like to express my deepest gratitude and appreciation to my academic advisor, Dr. Subhashis Ghosal, for his ceaseless inspiration and encouragement. Besides one of the leading researchers in the area of Bayesian Nonparametrics, he is also one of the most respectable people one could ever associate with. Without his great guidance and patience, it would have been more difficult for me to complete my doctoral degree.

I would also thank to Dr. Sastry Pantula for his continuous support of my graduate study, Dr. Cavell Brownie for her guidance on my work at the Center for Applied Aquatic Ecology at NCSU, my Ph.D. committee members for their valuable comments on my work and the draft. I also thanks to Terry Byron for his aid in computing. My thanks also go to other faculty, staff and students in our department. The congenial and conducive environment they provide has made my study such a pleasant and fruitful experience.

Finally and foremost, I would like to thank my parents and brother for their love and support all the way in my life.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

There has been much work on the estimation of the density function of a population based on a sample of independent and identically distributed (i.i.d.) observations. However, the i.i.d. assumption does not hold in many situations. Often, the observations show evidence of dependence on the past. Assuming that only the immediate past matters, we obtain a Markov process. The goal of this thesis is to propose a nonparametric Bayesian method for the estimation of the transition density of a Markov process model, and to study its asymptotic behavior.

In the conditional density estimation problem, the sample is a list of real-valued random variables $\{X_{-m+1}, X_{-m+2}, \ldots, X_0, \ldots, X_n\}$ * drawn from a discrete time series model. The observation X_i depends on only its m lagged values given all its past trajectory, that is, given $X_{-m+1}, \ldots, X_{i-1}$, the distribution of X_i is $f(\cdot|X_{i-1}, \ldots, X_{i-m})$. Our goal is to estimate the transition density f. Note that when $m = 1$, the time series model is a Markov process with the state space $\mathbb{R}$ and that when $m > 1$, the time series model could be represented by a Markov process in the product state

*we assume that the first m observations $X_{-m+1}, \ldots, X_0$ are fixed or have known distribution.

space $\mathbb{R}^m$. Without loss of generality, we refer to the general time series models as Markov processes.

Classical parametric approaches for the estimation of the transition density rely on the assumption that the functional form of the transition density f is completely known except for certain parameters, which have to be estimated from the sample. However, the assumption of a parametric model is often only a convenient mathematical artifice. To avoid unnatural model assumptions, non-parametric or semiparametric techniques for inference have been developed. Prakasa Rao (1978) proposed a kernel-type density estimation method to estimate the transition density and its invariant distribution function for a stationary uniformly ergodic Markov process (refer to Appendix A for the definition) under the assumption that both the transition density and the density for the invariant distribution are uniformly continuous. Mixture models are flexible nonparametric models and have interesting applications in Bayesian density estimation. Muller, West and MacEachern (1997) studied a Bayesian locally weighted finite mixture model for nonlinear autoregressive time series in which the size and terms of the mixing proportions of the mixture models have a random prior from the Dirichlet process. Ferguson (1983) and Lo (1984) considered the Dirichlet mixture of normal (DMN) model for the estimation of the density of an i.i.d. population. In this thesis, we propose using the Dirichlet process mixture (DPM) model to estimate the transition density of a Markov process. Our DPM model is a natural extension of Ferguson's and Lo's DMN model. Our model is different from Muller, West and MacEachern's in that the parameter in our model could be viewed as dynamically varying with time while in the latter, the parameter is static. Our models will be formally introduced in the next section.

Distributional symmetry is a key natural assumption in a statistical model if no information, other than the data, is available. The symmetry is represented probabilistically by exchangeability (refer to Section 2.1). A fundamental representation

theorem of de Finetti (1937) shows that an exchangeable probability measure is a mixture of the products of identical measures. This forms the fundamental pillar of Bayesian statistics. Thus exchangeable observations are conditionally i.i.d. This introduces the "parameter" and the "prior" in Bayesian inference. To predict the next observation, one finds the conditional distribution of the next given the available observations by integrating out the unknown distribution (parameter) with respect to the posterior distribution. The posterior is obtained from the conditional i.i.d. structure by the Bayes theorem. However, particularly, for observations arising over time (as in a time series analysis), the assumption of this form of distributional symmetry seems to be unconvincing. Therefore, the observations are not conditionally i.i.d., but the future observations should be conditionally dependent on the past given the parameter. A Markov process is one of the most natural extensions of the i.i.d. structure that can possibly capture this dependence, yet has a rich structure. For a Markov process, one assumes that given the present, the future will not depend on the past any further. However, the dependence will "propagate" and all the involved variables will be dependent. It is also easily possible to incorporate this "one step immediate dependence" to m steps by simply looking at an m-tuple of m successive observations. Therefore it seems to be reasonable to model the observations as a conditional Markov process. A prior is then put on the transition density of the process, and prediction can be done by integrating out the transition density with respect to the posterior distribution. Clearly, the conditional Markov process model contains the exchangeable model as a special case (with 0 terms). Including more terms is likely to enhance prediction power up to some stage, before it starts overfitting the data.

In this thesis, we shall demonstrate our models numerically in three ways. First, we will simulate data from known Markov models when $m = 1$, and compare the Bayes estimate of the transition density with the true one. In the literature, a widely

used nonparametric Bayesian model for estimating the density function of an i.i.d. sample is the DMN model, which is a special case of our model. We will compare the prediction power of our method under the Markov process assumption when the state space is $\mathbb{R}$ with that of the DMN model under the i.i.d. assumption. To see how our general Markov model captures all the dependence structure in practice, we shall study its prediction performance with real data.

The nonparametric Bayesian method is useful when little information is available. However, inference based on the posterior is reliable only if the posterior shows reasonable large sample frequentist properties. This thesis shows posterior consistency of our model when the state space of the Markov process is the real line $\mathbb{R}$. To the best of our knowledge, this thesis presents one of the first theoretical examinations of consistency issues of nonparametric Bayesian methods for dependent data. Our result is presented in chapter 4.

1.2 The Nonparametric Bayesian Model

It has been long known that the mixtures of a standard distribution could be used to approximate many densities. Diaconis and Ylvisaker (1985) observed that discrete mixtures of beta densities provide a dense class of models for densities on $[0, 1]$. Similarly, the discrete mixtures of gamma densities provide a dense class of models for densities on $\mathbb{R}^+$. Mixture models have been flexible nonparametric models. Dirichlet process mixture (DPM) models follow a Bayesian framework and have particularly been useful in this context in which the mixing distribution is unknown and automatically determined by the data.

In the DPM model proposed in this thesis, the data are modeled as a switching-regression model with its density mixed with respect to a distribution on the parameter. We seek to enrich the class of the regression models through modeling

the uncertainties about the functional form of the mixing distribution. We assume that the mixing distribution could be any distribution in the relevant space with a Dirichlet process prior. The unknown mixing distribution could have any shape such as degenerate, discrete, multimodal, skewed, fat-tailed and could be automatically learned from the data. In a Bayesian framework, the data automatically determine the amount of smoothing given the prior.

The Dirichlet mixture model could be viewed as a two-layer hierarchical Bayesian model. At the first layer, we assume that each observation of the variables under study comes from the following switching regression model,

$$X_i = H(Z_i, \tau) + \varepsilon_i, \tag{1.2.1}$$

where $\{\varepsilon_i, i \geq 1\}$ are i.i.d. from $N(0, \sigma^2)$, the normal distribution with mean 0 and variance σ^2, $\tau = (\beta_0, \beta_1, \gamma^T)^T$, $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_m)^T$, $Z_i = (1, X_{i-m}, \ldots, X_{i-1})^T$,

$$H(z, \tau) = \beta_0 + \beta_1 g(\gamma^T z). \tag{1.2.2}$$

and g is a known link function. The logistic function

$$g(t) = \frac{1}{1 + \exp(-t)} \tag{1.2.3}$$

seems to be a particularly interesting choice because of its boundedness, monotonicity, smoothness and ability to approximate a linear function; See Figure 1.1 and the discussion below. We shall work with this choice. However, other choices of g with similar properties are possible and the treatment will be very similar. Note that if $\beta_0^* = \beta_0 + \beta_1$, $\beta_1^* = -\beta_1$ and $\gamma^* = -\gamma$, then

$$H(z, \theta) = \beta_0 + \beta_1 g(\gamma^T z) = \beta_0^* + \beta_1^* g(-\gamma^{*T} z) = H(z, \theta^*).$$

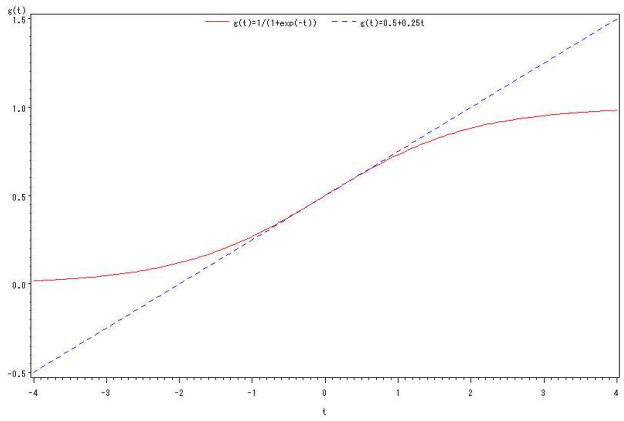
To avoid the identitiability issue of the parameter, we assume that $\gamma_m > 0$. The equation (1.2.1) could be interpreted as follows. When the linear combination of

the m lagged values $\gamma^T Z_i \ll 0$, the conditional mean of X_{i+1} is equal to β_0, while it changes gradually to $\beta_0 + \beta_1$ as $\gamma^T Z_i$ increases. The switching regression model could be used to approximate any autoregressive AR(m) model arbitrarily well. From Figure 1.1, we see that the logistic regression function (1.2.3) is approximately linear with slope 0.25 as t is near 0. Given any compact set $K \subset \mathbb{R}^m$, there exists a γ with sufficiently small $\|\gamma\|$ such that $\gamma^T Z_i$ will be close to 0 enough and hence

$$X_{i+1} \approx \beta_0 + 0.5 + 0.25\beta_1(\gamma^T Z_i) + \varepsilon_{i+1}.$$

when $Z_i \in K$. So for any given K, we could always adjust the parameter γ, β_0, β_1 in the switching regression model (1.2.1) and make it approximate any AR(m) model arbitrarily closely for $Z_i \in K$.

Figure 1.1: Logistic function *vs* linear function



In the hierarchical Bayesian model, we assume that the parameter $\theta = (\tau, \sigma)$ varies with time, that is, each observation is from

$$X_i = H(Z_i, \tau_i) + \varepsilon_i, \tag{1.2.4}$$

where $\varepsilon_i \sim N(0, \sigma_i^2)$ and θ_i's are i.i.d. from some distribution P. We need to specify the prior P for θ_i's. Often, the prior is chosen based on the user's prior belief. If we

have a strong belief that θ_i's are i.i.d. from some parametric distribution G_0, we could set the prior of θ_i as $P = G_0$. But usually the functional form of P is unavailable in practice. It is well known that the inference is often sensitive to the specification of such an explicit parametric functional form of P. To avoid this drawback, we instead assume that the prior P itself is random and could be any distribution function on the space Θ, where Θ denotes the space of θ_i and is a subset of Euclidean space $\mathbb{R}^{m+2} \times \mathbb{R}^+ \times \mathbb{R}^+$; note that $\gamma_m > 0$ and $\sigma > 0$.

At the second layer of the model, we need to specify a hyperprior for the random distribution P. Let $M(\Theta)$ denote the space of all probability measures on Θ. Conceptually, the specification of prior for P is different from that in a parametric Bayesian model. In parametric models, usually P is a distribution $G_0(\theta|\lambda)$ which is known except for the parameter λ. The unknown parameter λ is a real-valued scalar or a finite dimensional vector. One may use additional prior information or may choose a diffuse prior for λ. Unlike the parametric case, the space $M(\Theta)$ is infinite-dimensional. The distribution function P, viewed as an unknown element of $M(\Theta)$, is assumed to be random. The prior for P is hence a stochastic process on the infinite-dimensional space $M(\Theta)$. We will choose a Dirichlet process $D_{\alpha G_0}$ as the prior of P. The Dirichlet process prior is reviewed in Chapter 2. In a Dirichlet process, α is a known positive scalar, G_0 is a fixed distribution function in $M(\Theta)$. As noted by its constructive definition given in equation (2.1.1), the random realization P from the Dirichlet prior is almost surely discrete. In spite of the discreteness property of the Dirichlet process, its support for the weak-star topology is large enough and includes any probability measure with support belonging to the support of the center measure G_0. For example, if the support of G_0 is $\mathbb{R}^{m+2} \times \mathbb{R}^+ \times \mathbb{R}^+$, any P satisfying $P(\gamma_m > 0, \sigma > 0) = 1$ will be in the weak-star support of $D_{\alpha G_0}$. The Dirichlet process prior can be interpreted as follows. We are uncertain about the prior P of θ_i and

expect it to be close to G_0. The parameter α controls the extent of the closeness between P and G_0. A large α leads to a step function P that is likely to be close to G_0. It is similar to putting a prior G_0 directly on θ_i's for a given sample size. For a small α, the discrete P is likely to put most of its mass on just few atoms. Given a sample size n and a sufficiently small α, the model is very similar to the one that all θ_i's are clustered at a random point from G_0. However, the above interpretation indicates the prior information only. In a Dirichlet mixture model, the data will update the shape of the prior P properly with sufficiently many samples for any α and a center measure G_0 with large support.

The DPM model could be summarized in the following equations:

$$\begin{aligned} P &\sim D_{\alpha G_0} \\ X_i|P, Z_i &\sim f_P(X_i|Z_i) \text{ for } i \geq 1 \end{aligned} \tag{1.2.5}$$

where $Z_i = (1, X_{i-m}, \ldots, X_{i-1})$,

$$f_P(y|z) = \int_\Theta \phi_\sigma \left(y - H(z, \tau)\right) dP(\theta), \tag{1.2.6}$$

and

$$\phi_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{x^2}{2\sigma^2}\right]$$

is the density of the normal distribution $N(0, \sigma^2)$. In the model, P is the unknown parameter. Given P, $\{X_i,\, i \geq 0\}$ follows a time series model with the conditional density f_P.

The DPM is eqivalent to the two-layer hierarchical Bayesian model:

$$\begin{aligned} &P \sim D_{\alpha G_0}, \\ &\theta_1, \ldots, \theta_n | P \overset{\text{i.i.d.}}{\sim} P, \\ &X_i | Z_i, \theta_i \sim N(H(Z_i, \tau_i), \sigma_i^2) \text{ for } i \geq 1, \end{aligned} \tag{1.2.7}$$

where $\theta_i = (\tau_i, \sigma_i)$ for $i \geq 1$. The hierarchical model could be viewed as a dynamical time series model in that the parameters of the switching-regression model are varying

with time and are i.i.d. from P. The case of the static parameters is clearly included by taking P to be degenerate.

A special case of our model is the DMN model, in which the samples are assumed to be conditionally independent if $G_0(\beta_1 = 0) = 1$, that is $\beta_1 \equiv 0$ in the equation (1.2.4). Another special case of our model is the one which assumes that τ has a random mixing distribution P from the Dirichlet process prior and σ is a static parameter with a prior μ. This model is less flexible than the DPM model. We will study its asymptotic property of the posterior only. Essentially similar ideas, except for some notational complications, work for the less flexible model.

Let $\pounds_m$ denote the class of transition densities $f(\cdot|z)$ of interest. We specify a prior Π indirectly on the space $\pounds_m$ through the specification of a prior on P. The data series $\{X_n, n \geq 1-m\}$, conditioning on $f \in \pounds_m$ or equivalently on P, is a time series model with transition density f given by equation (1.2.6). Assume the first m points $X_{-m+1}, \ldots, X_0$ are fixed or have a known distribution. Then $\{\Pi, f\}$ together define the joint distribution of $\{X_n, n \geq 1-m\}$ and f. Formally, the posterior is a probability measure uniquely defined by the Bayes theorem as follows:

$$\Pi(df|X_0, \ldots, X_n) = \frac{\prod_{i=1}^n f(X_i|Z_i)\Pi(df)}{\int_{\pounds_m} \prod_{i=1}^n f(X_i|Z_i)\Pi(df)}. \tag{1.2.8}$$

Generally, we are interested in estimating the transition density function $f(y|z)$ and predicting the future value X_{n+1} given the whole history. In a Bayesian framework, the "decision rule" is the one that minimizes the expected loss calculated under the posterior. Because of the intractable analytical expression of the posterior, we have to estimate the quantity of interest numerically through Markov Chain Monte Carlo (MCMC) methods. As the random element in the Dirichlet mixture model is a random probability distribution P, it is difficult to use Monte Carlo method to simulate the random P exactly from the posterior. However, we could integrate out the random P and work with $\theta_1, \ldots, \theta_n$ directly. Chapter 3 describes an appropriate

MCMC algorithm.

1.3 Outline

The thesis is organized as follows. Chapter 2 gives a literature review. Chapter 3 presents the MCMC algorithm for simulation from the posterior, and shows how to compute the quantities in which we are interested. Chapter 4 presents our theoretical results on the posterior consistency for Markov processes with state space $\mathbb{R}$ under different topologies. Chapter 5 illustrates the DPM model with simulation studies and real data examples. For reference, Appendix A gives a brief review of the general state Markov Process and Appendix B gives a brief review of Markov Chain Monte Carlo simulation techniques.

Chapter 2

Literature Review

This chapter reviews and discusses the literature. The first section reviews the construction of the prior distribution on the space of probability measures and discusses the popular Dirichlet process mixture models. The second section reviews some theoretical results concerning the frequentist performance of Bayesian procedures. The proofs of posterior consistency of our nonparametric Bayesian modeling of Markov processes are based on these pioneering works.

2.1 Priors on the Space of Probability Measures

Conceptually, Bayesian nonparametric models are different from the parametric cases. In nonparametric modeling, the unknown parameter is in some class of functions. These functions might be the cumulative distribution or density functions, the regression function in regression models, the conditional cumulative or conditional density functions in time series models, the cumulative hazard or hazard function in survival analysis models and so on. The dimension of these function space are infinite. The Bayesian approach requires an infinite dimensional stochastic specification of the prior over the function space. However, in the past thirty years, there has been an

enormous literature on the specification of such priors.

Let Θ be a complete separable metric space with $\mathcal{B}(\Theta)$ being the corresponding Borel σ-algebra on Θ, and denote by $M(\Theta)$ the space of all probability measures on $(\Theta, \mathcal{B}(\Theta))$. The simplest Bayesian framework consists of a prior Π on $M(\Theta)$, and $\theta_1, \theta_2, \ldots$ is a sequence of i.i.d. random variables from $P \sim \Pi$. De Finetti's representation theorem (de Finetti, 1937) shows that a minimal judgement of exchangeability of the observation sequence leads to the above Bayesian formulation. The sequence of Θ-valued random variables $\{\theta_i, i \geq 1\}$ is said to be exchangeable if for each n and for every permutation g of $\{1, 2, \cdots, n\}$, the distribution of $\theta_1, \cdots, \theta_n$ is the same as that of $\theta_{g(1)}, \ldots, \theta_{g(n)}$. Let μ be a probability measure on Θ^∞. De Finetti's theorem shows that $\{\theta_i, i \geq 1\}$ is exchangeable if and only if there is unique probability measure Π on $M(\Theta)$ such that

$$\mu\{\theta_1 \in B_1, \cdots, \theta_n \in B_n\} = \int \prod_{i=1}^{n} P(B_i) d\Pi(P).$$

DeFinetti's theorem could be a guidance for constructing a prior distribution on $M(\theta)$. Among those priors on $M(\Theta)$, the Dirichlet process is the most widely used.

The Dirichlet Processes

The Dirichlet process, developed by Ferguson (1973), is the most widely used prior on $M(\Theta)$. Let G_0 be a fixed probability measure on $(\Theta, \mathcal{B}(\Theta))$ and α be a positive number. A random probability measure $P \in M(\Theta)$ is said to follow a Dirichlet process $D_{\alpha G_0}$ if for any finite measurable partition $B_1, \cdots, B_k$ of Θ, $(P(B_1), \cdots, P(B_k))$ has the Dirichlet distribution $D(\alpha G_0(B_1), \cdots, \alpha G_0(B_k))$ (if $G(B_i) = 0$, then $P(B_i) = 0$ with probability 1). The measure $G = \alpha G_0$ is referred to as the base measure of the Dirichlet process. Since for any $B \in \mathcal{B}(\Theta)$,

$$\mathrm{E}(P(B)) = G_0(B) \text{ and } \mathrm{Var}(P(B)) = \frac{G_0(B) G_0(B^c)}{1 + \alpha},$$

G_0 is viewed as the center of the process while α can be loosely interpreted as a precision parameter *. The larger α is, the closer we expect a realization from the process to be G_0. On the other hand, if $\alpha \to 0$, the random variable $P(B)$ has the maximum variance. Thus when $D_{\alpha G_0}$ is interpreted as a prior distribution for P in Bayesian inference, $D_{\alpha G_0}$ may be viewed as noninformative as $\alpha \to 0$.

A constructive definition of the Dirichlet process is given by Sethuraman (1994). A realization P from $D_{\alpha G_0}$ is almost surely of the form

$$P = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}, \tag{2.1.1}$$

where $\theta_1, \theta_2, \ldots$ are a sequence of i.i.d. random variables distributed according to G_0, $\beta_1, \beta_2, \ldots$ are a sequence of i.i.d. random variables from Beta$(1, \alpha)$, and $p_1 = \beta_1$, and for $n \geq 2$, $p_n = \beta_n \prod_{i=1}^{n-1}(1 - \beta_i)$. Note that it follows $\sum_{i=1}^{\infty} p_i = 1$ a.s.. Clearly the Dirichlet process places all its mass on the subset of all discrete probability measures on Θ. This fact was earlier noted by Ferguson (1973) and Blackwell and McQueen (1973).

In spite of the discreteness property of the Dirichlet process, its support for the weak star topology is quite large. Let Θ^* be the support of the center measure G_0. Any probability measure with support belonging to Θ^* will lie in the weak support of $D_{\alpha G_0}$. The Dirichlet process has the congjugacy property. Suppose that $\theta_1, \ldots, \theta_n$ are i.i.d. from P and P has a Dirichlet process prior $D_{\alpha G_0}$. Then posterior distribution of P is again a Dirichlet process:

$$P|\theta_1, \ldots, \theta_n \sim D_{\alpha^* G_0^*},$$

where $\alpha^* = \alpha + n$ and $G_0^* = (\alpha+n)^{-1}(\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i})$. Thus the predictive distribution

*Sethuraman and Tiwari (1982) shows that this interpretation may sometimes be misleading because $D_{\alpha G_0}$ converge weakly to a random degenerated measure δ_{θ^*} where $\theta^* \sim G_0$ as $\alpha \to 0$. In fact, as seen in equation (2.1.1), another role of α is to control probability of ties of observations generated from random P following the Dirichlet process. When α is small, the random realization P tends to concentrate on few points which are i.i.d. from G_0.

of $\theta_{n+1}|\theta_1, \ldots, \theta_n$ is G_0^*. This leads to the Polya urn sampling scheme as described by Blackwell and MacQueen (1973):

$$\begin{aligned} &\theta_1 \sim G_0 \\ &\theta_{i+1}|\theta_1, \cdots, \theta_i \sim G_{0i}^* = \frac{\alpha G_0 + \sum_{j=1}^{i} \delta_{\theta_j}}{\alpha + i} \text{ for } \quad i \geq 1. \end{aligned} \tag{2.1.2}$$

Dirichlet Mixture Processes

One disturbing aspect of the Dirichlet process is that it puts all its mass on the subset of all discrete distributions. In order to constrain the prior support to have distribution with some smoothness properties, Lo (1984) and Ferguson (1983) developed a useful construction of priors on densities through Dirichlet Mixture processes in which the sample $X_1, \ldots, X_n$ is from the Mixtures of Kernels $\psi(\cdot, P) = \int K(\cdot, \beta, \theta) dP(\theta)$, where $K(\cdot, \beta, \theta)$ is a density function given β and θ, while $P \sim D_{\alpha G_0}$. In Lo's and Ferguson's construction, it is assumed that $X_1, \ldots, X_n$ are conditionally independent given the parameter (β, P). Their construction could be naturally extended to the dependent data structure by taking the kernel K to be the conditional density of a discrete time series model. Obviously, our DPM model follows this structure.

In a Dirichlet mixture process, it is convenient to view the observations X_i as arising from $K(\cdot, \beta, \theta_i)$ for $i = 1, \ldots, n$ where $\theta_1, \ldots, \theta_n$ are i.i.d. from $P \sim D_{\alpha G_0}$. The random P could be integrated out, and the joint distribution of $\theta_1, \ldots, \theta_n$ could be obtained from the Polya urn presentation in equation (2.1.2). The posterior distribution of θ_i given all other parameters is

$$\theta_i | \left(\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots \theta_n, \beta, X_1, \cdots, X_n\right) \propto q_0 G_{-i}^*(d\theta_i) + \sum_{j \neq i} q_j \delta_{\theta_j}(d\theta_i), \tag{2.1.3}$$

where

$$\begin{aligned} q_0 &= \int K(X_i, \beta, \theta) dG_0(\theta), \\ q_j &= K(X_i, \beta, \theta_j), \end{aligned}$$

$$G^*_{-i}(d\theta_i) \propto K(X_i, \beta, \theta) dG_0(\theta).$$

In designing the MCMC sampling scheme from the posterior, the θ_i's could be directly drawn with the Gibbs sampler without having to draw the random P. In the case when the kernel $K(\cdot, \beta, \theta)$ is conjugate with respect to G_0, it is easy to compute q_0 and sample the θ_i's sequentially from the posterior with the Gibbs sampler. Such MCMC computation schemes were studied by Escobar (1994), Escobar and West (1995), MacEachern (1994), Bush and MacEachern (1996), and Escobar and West (1998). In the absence of conjugacy, MacEachern and Muller (1998) developed the "no gaps" algorithm in which the parameter space is cleverly augmented, and Neal (1998) developed algorithms based on Metropolis-Hasting sampling scheme. In our numerical demonstration, we will use Muller and MacEachern's "no gap" sampling scheme and appropriately adapt it to our situation.

Mixtures of Dirichlet processes considered by Antoniak (1974) allow the base measure of the Dirichlet process itself to be random. In the literature, generally the mixture of Dirichlet processes and Dirichlet mixture are used together to develop hierarchical Bayesian models. Another popular prior on $M(\Theta)$ is the Polya tree distribution, which generalizes the Dirichlet processes, and can be chosen to assign probability 1 to the space of continuous, or even absolutely continuous distributions. Ferguson (1974), Maudlin, Sudderth and Williams (1992), and Lavine (1992, 1994) and many others give a good introduction on the Polya tree priors. The tailfree process, considered by Freedman (1963, 1965), is a very general class of priors that includes the Polya tree distribution and the Dirichlet process. The neutral to the right (NTR) prior [Doksum (1974), Ferguson and Phadia (1979)] is a special case of the tailfree process and used widely in survival models. Various NTR priors in literature include gamma process, extended gamma process, beta process and beta-Stacy process. These processes are closely connected with the Levy process. Gelfand (1997)

gave a good review on these topics. Among various priors on $M(\Theta)$, the Dirichlet process is the easiest to specify and handle mathematically.

2.2 Theoretical Results

This section provides a brief review of some large sample frequentist properties of Bayesian procedures. We focus mainly on posterior consistency, but we touch on a few other issues as well.

Consider a sequence of experiments indexed by a parameter θ taking values in the space Θ. The space Θ need not be a subset of Euclidean space, so that the nonparametric and semiparametric problems are also included. The observation at the nth stage is denoted by $X^{(n)}$. The law of $X^{(n)}$ is a probability $P_\theta^{(n)}$ controlled by the parameter θ. In this section, we assume that $\{X_i, i \geq 1\}$ are i.i.d. P_θ where the probability measure P_θ has density p_θ with respect to a σ-finite measure μ, unless explicitly mentioned otherwise. Further we denote the prior on Θ by Π and the posterior distribution $\Pi(\theta|X_1, \ldots, X_n)$ by Π_n. There may exist different versions of the posterior. However, when the family $\{P_\theta^n, \theta \in \Theta\}$ is dominated, the posterior is essentially unique and is given by the Bayes theorem.

The posterior Π_n is said to be consistent at θ_0 if $\Pi_n(U) \to 1$ a.s. under the law determined by θ_0 for every neighborhood U of θ_0, that is, the posterior should concentrate around the true parameter value as more data comes in. Obviously this definition of posterior consistency depends on the topology on the parameter space Θ and the version of the posterior. By the Portmanteau theorem, the consistency of the posterior at θ_0 is equivalent to requiring that $\Pi_n \overset{\text{weakly}}{\longrightarrow} \delta_{\theta_0}$ a.s. the law determined by θ_0, where δ_{θ_0} is the degenerate measure at θ_0.

Bayesian methods without reasonable consistency properties may cause serious problems in that the posterior may mislead to a wrong value or wander indefinitely

around the parameter space. Suppose that an experimenter generates observations from a known (to the experimenter) distribution. It would be embarrassing if a Bayesian fails to come close to finding the mechanism used by the experimenter even with an infinite amount of data. Consistency is important also to subjective Bayesians, who do not believe in such true models. Consistency is equivalent to "intersubjective agreement", which means that two Bayesians, with different priors, presented with the same data, will ultimately have very close predictive distributions. Blackwell and Dubins (1962) showed that if two priors are mutually absolutely continuous, the two corresponding predictive distributions of the future $(X_{n+1}, X_{n+2}, \ldots)$ given the past will merge[†] in total variation distance almost surely. Diaconis and Freedman (1986) studied the mergence of the predictive distributions under weak convergence topology. For parametric problems, Ghosh, Ghosal and Samamta (1994) showed that under certain condition the posterior distributions merge in the total variation distance.

The frequentist property is well understood in finite dimensional (parametric) Bayesian models. The posterior will be consistent under mild regularity conditions and the i.i.d. set up. Moreover, the Bernstein-von Mises theorem [‡] asserts that if the densities form a smooth class, then the Bayes estimate and the maximum likelihood estimate will be close and that the posterior distribution of the parameter vector around the posterior mean is close to the distribution of the maximum likelihood estimate around the truth: Both are asymptotically normal with mean 0 and the same covariance matrix. For this reason, the Bernstein-von Mises theorem is sometimes called the "Bayesian Central Limit Theorem". Borwanker, Kallianpur

[†]Two sequences of distributions $\{P_n\}$ and $\{Q_n\}$ are said to merge in some topology if $d(P_n, Q_n) \to 0$ a.s. as $n \to \infty$, where d is the corresponding metric.

[‡]the phenomenon of the Berstein-von Mises theorem was first observed by Laplace (1774) and rediscovered by Bernstein (1917) and von-Mises (1931); Le Cam (1953, 1958) gave the first rigorous proof under i.i.d. assumption

and Prakasa Rao (1971) gave a version of Bernstein-von Mises theorem for general discrete time stochastic processes. Ghosal, Ghosh and Samanta (1995) gave a version of the Bernstein-von Mises theorem for general parametric Bayesian models which requires neither the i.i.d. structure nor the smoothness conditions.

However, in infinite-dimensional nonparametric problems, the situation becomes much more complex. Freedman (1963) constructed a classical example in which the posterior is not consistent. The question of interest is to estimate an unknown probability mass function θ on the set of positive integers in the infinite multinomial problem. Let θ_0 stand for the geometric distribution with parameter $\frac{1}{4}$. Freedman (1963) constructed a prior which gives positive mass to every weak neighborhood of θ_0 but the posterior concentrates in the weak neighborhoods of a geometric distribution with parameter $\frac{3}{4}$. Freedman (1963) continued to show that in his counterexample, most priors are troublesome in a topological sense in that they form the complement of a first category meagre set. Freedman's (1963) result shows that in infinite-dimensional problems, consistency is the exception, not the rule, which led to some criticism of Bayesian methods. Some more classical counterexamples of posterior inconsistency are listed as follows. In the symmetric location problem, the data are modeled as $X_i = \theta + \epsilon_i$ in which ϵ_i are i.i.d. from a symmetric distribution F. Diaconis and Freedman (1986) showed that the Bayes estimates of the location parameter θ in the symmetric location problem can be inconsistent if the true distribution of ϵ_i is continuous while F has a (symmetrized) Dirichlet process prior. See Doss (1985) for a variation of this problem where the Dirichlet process is not symmetric and the median is of interest. Some counterexamples of posterior inconsistency in survival models were given by Kim and Lee (2001). In spite of Freedman's (1963) conclusion, usually plenty of priors can be constructed to match one's prior belief arbitrarily close yet achieving consistency. Many recent works have focused on studying the posterior consistency in nonparametric Bayesian fields.

An early result on consistency is due to Doob (1948). His assumption is essentially the minimal. The posterior is consistent at every θ except possibly on a null set of Π measure 0 for any prior Π in the i.i.d. case. Doob's theorem holds for general Bayesian models if the parameter is consistently estimable. Many Bayesians are satisfied with Doob's theorem. However the theorem fails to tell at which point the posterior is consistent. The null set, depending on the prior, may be quite large. For example, consider $\{X_i\}$ are i.i.d. from $N(\theta, 1)$ and the prior is a point mass at 0. Then the posterior consistency fails at all points except 0.

Schwartz (1965) established a theory which enables us to study posterior consistency at a particular point θ_0 under the i.i.d. assumption. Schwartz's theory and its various extensions have been the main tool for establishing the posterior consistency, especially in nonparametric problems. To state her theory in detail, we recall the following definition.

Definition 2.2.1. A sequence of tests $\phi_n(X_1, \ldots, X_n)$ is "uniformly consistent" for

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta \in U^c$$

if both the type I and type II errors converge to 0, that is,

$$\mathrm{E}_0(\phi_n) \to 0 \text{ and } \sup_{\theta \in U^c} \mathrm{E}_\theta(1 - \phi_n) \to 0;$$

The sequence of test is " exponentially consistent" if both the type I and II error are exponentially small, that is, there exist $C > 0, \beta > 0$ such that

$$\mathrm{E}_0(\phi_n) \leq Ce^{-n\beta} \text{ and } \sup_{\theta \in U^c} \mathrm{E}_\theta(1 - \phi_n) \leq Ce^{-n\beta}.$$

Under the i.i.d. assumption, the existence of a uniformly exponentially consistent sequence of tests is equivalent to the existence of a uniformly consistent sequence of tests by Hoeffding's inequality (refer to Theorem C.0.2). Schwartz (1965) showed that if

(1) the prior puts positive mass in any Kullback-Leibler neighborhood (refer to Section 3.1) of θ_0,

(2) there exists a uniformly exponentially consistent sequence of tests for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in U^c$ for any neighborhood U of θ_0,

then the posterior is consistent at θ_0.

The Condition (2) ensures that for some $\beta_0 > 0$ and any neighborhood of U,

$$\liminf_{n\to\infty} e^{n\beta_0} \int_{U^c} \prod_{i=1}^{n} \frac{p_\theta(X_i)}{p_{\theta_0}(X_i)} \Pi(d\theta) = 0 \text{ a.s. } P_{\theta_0}^{\infty}, \tag{2.2.1}$$

and the condition (1) implies that for any $\beta > 0$,

$$\liminf_{n\to\infty} e^{n\beta} \int_{\Theta} \prod_{i=1}^{n} \frac{p_\theta(X_i)}{p_{\theta_0}(X_i)} \Pi(d\theta) = \infty \text{ a.s. } P_{\theta_0}^{\infty}. \tag{2.2.2}$$

Hence the ratio $\Pi_n(U^c)$ converges to 0 almost surely by taking $\beta = \beta_0$.

For parametric Bayesian models, Schwartz's conditions are of essentially weaker nature than the condition for the consistency of maximum likelihood estimator (MLE). However, Schwartz (1965) cleverly constructed an example in which Wald's condition holds, and hence the MLE is consistent, but the posterior is inconsistent for some prior which shrinks too fast around the true value of the parameter and thus fails to put positive mass over a sufficiently small Kullback-Leibler neighborhood. We should note that Schwartz's two conditions are sufficient and not necessary, as shown by the following example from Ghosh and Ramamoorthi (2003). Let $\{X_i, i \geq 1\}$ be i.i.d. from the uniform distribution on $(0, \theta)$ and $\theta \in \Theta = (0, 1]$. Suppose that $p_0 = U(0, 1)$ and $\Pi = U[0, 1]$. It is easy to see that the posterior is consistent but it does not satisfy Schwartz's condition (2) at p_0.

Schwartz-type theorem are of tremendous usefulness for nonparametric Bayesian models. Let U denote any neighborhood of p_0 under the weak star topology; then there exists an exponentially consistent sequence of tests for $p = p_0$ versus $p \in U^c$.

Thus if the prior puts positive mass to all Kullback-Leibler neighborhood of p_0, then the posterior is consistent at p_0 under the topology of weak star convergence by Schwartz's theorem. Under condition (1), the strong law of large numbers is a key result to prove the equation (2.2.2). Schwartz (1965) pointed out that her result could possibly be extended to the dependent data problem in which the strong law of large numbers holds, such as the discrete uniformly ergodic Markov process. However, the notion of a Kullback-Leilber neighborhood may need some modification. Most importantly, finding an exponentially consistent sequence of tests to satisfy Schwartz's condition (2) is usually the most challenging problem.

Weak star neighborhoods are large. It would be better if the posterior could be shown to concentrate in stronger neighborhoods. However, Le Cam (1973) and Barron (1989) showed that there does not exist a uniformly exponentially consistent sequence of tests for $p = p_0$ versus $p \in U^c$ if p_0 is nonatomic and U is a total variation neighborhood of p_0. Barron (1988) extended Schwartz's theory. Barron's result shows that if p_0 is in the Kullback-Leibler support of the prior, then for establishing posterior consistency at p_0, it is enough to show that there is an exponentially consistent sequence of tests for p_0 versus V_n where $U^c \cap V_n^c$ has exponentially small prior probability, that is, $\Pi(U^c \cap V_n^c) \leq c_2 e^{-n\beta_2}$ for some positive c_2 and β_2. Barron (1988) also observed that the condition that the type I errors are exponentially small could be replaced by $\Pr_0(\phi_n > 0 \quad \text{infinitely often}) = 0$ if all other conditions hold. For the estimation of the density function with i.i.d. observations, Barron, Schervish, and Wasserman (1999) developed sufficient conditions using bracketing metric entropy (refer to Appendix C) for showing posterior consistency under total variation (or equivalent the Hellinger) metric based on Barron's (1988) results. However, their conditions are stronger than needed. Ghosal, Ghosh and Ramamoorthi (1999a) obtained a consistency result that uses L_1- metric entropy [§] without bracketing, which

[§]the same as the the total variation metric except a factor of 2

is weaker than the bracketing entropy condition. The result of Ghosal *et al.* (1999a) is described below. Suppose that p_0 is in the Kullback-Leibler neighborhood of the prior, and for every $\epsilon > 0$ there exists a $\delta < \epsilon, c_1, c_2 > 0, \beta < \epsilon^2/2$ and $\mathcal{F}_n$ such that

(a) $J(\delta, \mathcal{F}_n) < n\beta$ where $J(\delta, \mathcal{F}_n)$ is the logarithm of the minimal number of balls of radius δ in total variation metric needed to cover the $\mathcal{F}_n$,

(b) $\Pi(\mathcal{F}_n^c) < c_1 e^{-nc_2}$;

then Hellinger (and total variation) consistency obtains.

The first condition of the theorem ensures that there exists an exponentially consistent sequence of tests if the densities are restricted to the "sieve" $\mathcal{F}_n$. The second condition ensures that the complement of the sieve in $M(\Theta)$ barely receives prior mass. Similarly to the Schwartz theory, the prior needs to put positive mass around any Kullback-Leibler neighborhood of p_0. The results established by Ghosal *et al.* (1999a) and Barron *et al.* (1999) could be used to show the posterior consistency in Dirichlet mixture of normal model, Polya trees and infinite dimensional exponential families in total variation (Hellinger) metric. Applying Schwartz's theorem, Ghosal *et al.* (1999b) showed that the posterior for the location parameter is consistent in the location problem if an appropriate Polya tree prior or a Dirichlet mixture of normal is used in place of the Dirichlet process. Amewou-Atisso, Ghosal, Ghosh and Ramamoorthi (2003) extended their result to semiparametric regression models $Y_i = \alpha + \beta x_i + \epsilon_i$ where the error term ϵ_i are i.i.d from an unknown density p symmetric around 0. They also gave a non-identical version of Schwartz's theorem. Kim and Lee (2001) studied sufficient conditions for posterior consistency in survival models.

The consistency of the posterior implies the consistency of the Bayes estimate of the density in nonparametric Bayesian problem. If the posterior is consistent at p_0 in the weak-star topology (respectively, the total variation metric), then the Bayes estimate of the density function under squared error loss converges to f_0 weakly (respectively, in the total variation distance).

Ghosal, Ghosh and van der Vaart (2000) studied the convergence rates of the posterior distributions and applied the results to several examples including priors on finite sieves, log-spline models, Dirichlet processes and interval censoring. One of their fundamental results is as follows.

Suppose that for a sequence $\{\epsilon_n\}$ with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$, a constant $C > 0$ set sets $\mathcal{P}_n \subset \mathcal{P}$ where $\mathcal{P}$ denotes the parameter space such that

(a) $\log D(\epsilon_n, \mathcal{P}_n, d) \leq n\epsilon_n^2$, where d is the metric on $\mathcal{P}$ and $D(\epsilon, \mathcal{P}, d)$ is the covering number (refer to Appendix C).

(b) $\Pi(\mathcal{P} - \mathcal{P}_n) \leq e^{-n\epsilon_n^2(C+4)}$,

(c) $\Pi\left\{P : -P_0(\log \frac{p}{p_0}) \leq \epsilon_n^2, P_0(\log \frac{p}{p_0})^2 \leq \epsilon_n^2\right\} \geq e^{-n\epsilon_n^2 C}$;

then $\Pi(P : d(P, P_0) \geq M\epsilon_n | X_1, \ldots, X_n) \to 0$ in probability for a sufficiently large M.

In other words, ϵ_n is the convergence rate of the posterior distribution. Condition (a) ensures that there exist certain highly powerful tests for testing against the complement of a neighborhood shrinking at a certain rate while condition (b) helps to effectively reduce the size of the parameter space $\mathcal{P}$. Condition (c) requires that the prior measures put a sufficient amount of mass near the true measure P_0. A similar result under stronger conditions involving bracketing entropy integrals was obtained by Shen and Wasserman (2001).

In this thesis, we extend Schwartz's (1965) theorem and Ghosal, Ghosh and Ramamoorthi's (1999a) theorem to find sufficient conditions for posterior consistency in our Dirichlet process mixture model of the Markov process. We have not attempted to obtain convergence rates of the posterior distribution for Markov process. We expect that this could be done by using some of the ideas of Ghosal and van der Vaart (2001), who studied the convergence rate of the posterior for the Dirichlet mixture of normal model with i.i.d. observations. We expect to address this issue in future research.

Chapter 3

Monte Carlo Simulation from the Posterior

This chapter describes a Markov Chain Monte Carlo (MCMC) algorithm that could be used to sample from the posterior distribution. Recall that we are interested in the predictive distribution of X_{n+1} and the predictor of the one-step future value X_{n+1}. In a Bayesian framework, the Bayes estimate is the one which minimizes the posterior expected error loss. Under the squared error loss, the Bayes estimate is just the posterior mean. In general, analytical expression of the Bayesian estimate, involving the posterior, is complex. However, the posterior for the Dirichlet mixture model is amenable to MCMC methods. Suppose that we want to estimate $\int f(\beta)\Pi(\beta|X_0,\ldots,X_n)$ where $\Pi(\beta|X_0,\ldots,X_n)$ is the posterior distribution. In MCMC, an irreducible and aperiodic Markov process is designed such that it has the stationary distribution $\Pi(\beta|X_0,\ldots,X_n)$ and it is easy to sample from the chain. The Bayes estimate $\int f(\beta)\Pi(\beta|X_0,\ldots,X_n)$ could be approximated by $m^{-1}\sum_{i=1}^{m} f(\beta_i)$ numerically by the strong law of large number (refer theorem A.1.3). Appendix 1 and 2 give an introduction on Markov process and MCMC simulation techniques.

Unlike the parametric Bayesian models, one main difficulty in developing MCMC algorithms for Dirichlet mixture model is that the model consists of a random mixing

distribution P that is difficult to sample precisely. However, the Dirichlet mixture model could be represented by a parametric model equivalently in a finite sample size problem in which the random P is integrated out. Reference for various MCMC methods for Dirichlet mixture type model is given in Section 2.1. In our numerical demonstration, we shall use Muller and MacEachern's (1998) "no gaps" sampling scheme. Theoretically the "no gaps" algorithm is proved to converge almost surely under a very mild sufficient condition.

3.1 More on Dirichlet Process and Some Notations

The Dirichlet process is reviewed in Section 2.1. Let Θ be a complete separable metric space with $\mathcal{B}(\Theta)$ the corresponding Borel σ-algebra on Θ. Let $M(\Theta)$ denote the space of all probability measures on $(\Theta, \mathcal{B}(\Theta))$. Let G_0 be a fixed probability measure on $(\Theta, \mathcal{B}(\Theta))$ and α be a positive number. The Dirichlet process is a probability measure on the space $M(\Theta)$. A random probability measure $P \in M(\Theta)$ is said to follow a Dirichlet process $D_{\alpha G_0}$ with parameter G_0 and α if for any finite measurable partition $B_1, \cdots, B_k$ of Θ, $(P(B_1), \ldots, P(B_k))$ has the Dirichlet distribution $D(\alpha G_0(B_1), \ldots, \alpha G_0(B_k))$ *. This section presents some concepts and results related to the Dirichlet process, which are the key in the development of MCMC algorithm for the Dirichlet mixture model.

In a model related to the Dirichlet process, the random P is infinite dimensional and is not easy to work with directly. Suppose that $\{\theta_i, i \geq 1\}$ are i.i.d. according to $P \sim D_{\alpha G_0}$. The following theorem, abstracted from Ferguson (1973), shows that we could integrate out the random component P and work with the random variables

*if $G(B_i) = 0$, then $P(B_i) = 0$ with probability 1

$\{\theta_i, i \geq 1\}$ directly. Result (c) in this theorem gives the conditional distribution of θ_{n+1} given $\theta_1, \ldots, \theta_n$, which implies the generalized Polya urn sampling scheme given in equation (2.1.2).

Theorem 3.1.1 (Ferguson, 1973). *Suppose that* $\{\theta_i, i \geq 1\}$ *are i.i.d. from* $P \sim D_{\alpha G_0}$. *Let* g *be a measurable function such that* $\int |g| dG_0 < \infty$.

(a) *The posterior distribution of* P *given* $\theta_1, \ldots, \theta_n$ *is a Dirichlet process with base measure*

$$\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i}, \tag{3.1.1}$$

where δ_θ *denotes the measure giving mass* 1 *to the point* θ.

(b) *With probability* 1, $\int |g| dP < \infty$, *and*

$$\mathrm{E}\left(\int g dP\right) = \int g dG_0.$$

In particular, if $g(\theta) = I(B)$, *we have* $\mathrm{E}(P(B)) = G_0(B)$. *Hence the marginal distribution of* θ_i*'s is* G_0.

(c) *The predictive distribution of* θ_{n+1} *given* $\theta_1, \ldots, \theta_n$ *is*

$$\theta_{n+1} | \theta_1, \ldots, \theta_n \sim G_{0n}^* = (\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i})/(\alpha + n). \tag{3.1.2}$$

The joint distribution of $\theta_1, \ldots, \theta_n$ *is*

$$\frac{\prod_{i=1}^{n} (\alpha G_0 + \sum_{j=1}^{i} \delta_{\theta_j})}{\alpha^{[n]}}$$

where $\alpha^{[k]} = \alpha(\alpha + 1) \ldots (\alpha + k - 1)$.

(d) *The Bayes estimate* $g(\theta)$ *given* $\theta_1, \cdots, \theta_n$ *under the squared error loss is given by*

$$\mathrm{E}\left(\int g dP \,\middle|\, \theta_1, \cdots, \theta_n\right) = \int g dG_{0n}^* = \frac{\alpha \int g dG_0 + \sum_{i=1}^{n} g(\theta_i)}{\alpha + n}.$$

The constructive definition of the Dirichlet process given in equation (2.1.1) implies that the random realization P from the Dirichlet process is almost surely discrete. So

there may be ties among $\{\theta_i, i \geq 1\}$. If G_0 contains a discrete component, some θ_i's may be equal not because of the inherent discreteness property of Dirichlet process, but also because the draws from G_0 happen to be equal. We shall assume that the base measure G is "nonatomic", that is, for any $\theta \in \Theta$, $G(\{\theta\}) = 0$. Then the ties among θ_i's are caused only by the inherent discreteness property of the Dirichlet process. This fact could also be seen from the posterior distribution of θ_{n+1} given $\theta_1, \ldots, \theta_n$ given in equation (3.1.2).

Let $\phi = (\phi_1, \ldots, \phi_k)$ denote the set of distinct θ_i's, where k is the number of distinct elements of $\theta_1, \ldots, \theta_n$. Let $s = (s_1, \ldots, s_n)$ denote the vector of configuration indicators defined by $s_i = j$ if and only if $\theta_i = \phi_j$, $i = 1, \ldots, n$. We will use the term "cluster" with notation $g = (I_1, \ldots, I_k)$ to refer to the set of all observations X_i's with identical configuration indicators s_i. The jth cluster is $I_j = \{i : s_i = j\}$. Obviously $\{I_1, \ldots, I_k\}$ are a disjointed partition of $I = \{1, \ldots, n\}$. We assume that no order is put on the clusters $I_1, \ldots, I_k$. Given the configuration vector s and ϕ, θ and g is uniquely determined. But if we allow arbitrary permutations of the ϕ indexed by $j = 1, \ldots, k$, any given θ corresponds to $k!$ pairs (ϕ, s). In this situation, we will assign equal probabilities to each of the $k!$ permutations. For example, when $n = 3$ and $\theta_1 = \theta_3 \neq \theta_2$, there are two distinct observations θ_1, θ_2. If $\phi = (\theta_1, \theta_2)$, then $s = (1, 2, 1)$. Otherwise $\phi = (\theta_2, \theta_1)$, then $s = (2, 1, 2)$. The k distinct values $\phi_1, \ldots, \phi_k$ could also be assumed to be naturally picked from the vector θ in the following way:

$$\begin{aligned} &\phi_1 = \theta_1, \\ &\text{when } j \geq 2,\ \phi_j = \theta_i, \text{ where } i = \min\{m : \theta_m \neq \phi_1, \ldots, \theta_m \neq \phi_{j-1}\}. \end{aligned} \tag{3.1.3}$$

Then (s, ϕ) would be uniquely defined given θ. For example, when $n = 3$, $\theta_1 = \theta_3 \neq \theta_2$, $s = (1, 2, 1)$ and $\phi = (\theta_1, \theta_2)$ while $s = (2, 1, 2)$, $\phi = (\theta_2, \theta_1)$ does not exist. As no order is put on g, it is uniquely defined. For this example, $g = (\{1, 3\}, \{2\})$. In either

situation, given k, the distinct random values in the vector ϕ will be i.i.d. with the distribution G_0. However, the conditional distribution of s given the particular ϕ will be different in the two notations as given in the following theorem.

Theorem 3.1.2 (Antoniak (1974) and Korwar and Hollander (1973)). *Suppose that $\theta_1, \ldots, \theta_n$ are i.i.d. from $P \sim D_{\alpha G_0}$ where G_0 is nonatomic. Let the k distinct values among $\theta_1, \ldots, \theta_n$ be denoted by $\phi = (\phi_1, \ldots, \phi_k)$, the vector of configuration indictors denoted by s. Let $g = (I_1, \ldots, I_k)$ denote the clusters with sizes $n_1, \ldots, n_k$. Then the following assertions hold.*

(i) *The distribution of k is given by*

$$\Pr(k) = \frac{\alpha^k c(n,k)}{\alpha^{[n]}},$$

where $c(n,k) = \sum_g [\prod_{j=1}^k (n_j - 1)!]$, the sum is over all possible g. Furthermore,

$$\mathrm{E}(k) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} \cong \alpha \log\left(\frac{n+\alpha}{\alpha}\right).$$

(ii) *The distribution of g is given by*

$$\Pr(g = (I_1, \ldots, I_k)) = \Pr(g,k) = \frac{\alpha^k \prod_{j=1}^k (n_j - 1)!}{\alpha^{[n]}}, \tag{3.1.4}$$

which depends only on the number and sizes of clusters.

(iii) *Suppose that $\phi_1, \ldots, \phi_k$ are taken naturally from the vector θ as in equation* (3.1.3)*. The distribution of s is given by*

$$\Pr(s = (s_1, \ldots, s_n)) = \Pr(g) = \frac{\alpha^k \prod_{j=1}^k (n_j - 1)!}{\alpha^{[n]}}.$$

If we allow arbitrary permutations of ϕ indexed by $j = 1, \ldots, k$, and assign equal probabilities to each of the $k!$ permutations, then the distribution of s is given by

$$\Pr(s = (s_1, \ldots, s_n)) = \frac{\alpha^k \prod_{j=1}^k (n_j - 1)!}{\alpha^{[n]} k!}. \tag{3.1.5}$$

(iv) *Given k, $\phi_1, \ldots, \phi_k$ are independent from the law G_0.*

We will illustrate the use of Theorem 3.1.2 when the sample size $n = 3$. The joint distribution of $\theta_1, \ldots, \theta_n$ given by Theorem 3.4.1 is quite complicated. Theorem 3.1.2 also tells us how to simplify this joint distribution function when G_0 is nonatomic.

Example 3.1.1. If $\theta_1 = \theta_3 \neq \theta_2$, then $g = (\{1, 3\}, \{2\})$. By equation (3.1.4),

$$\Pr(g) = \Pr(\theta_1 = \theta_3 \neq \theta_2) = \frac{\alpha^2}{\alpha(\alpha+1)(\alpha+2)}.$$

If $\phi = (\theta_1, \theta_2)$, $s = (1, 2, 1)$, and if $\phi = (\theta_2, \theta_1)$, $s = (2, 1, 2)$. If we assign equal probabilities to the two configuration of s, then by equation (3.1.5),

$$\Pr(s = (2, 1, 2)) = \Pr(s = (1, 2, 1)) = \frac{\Pr(\theta_1 = \theta_3 \neq \theta_2)}{2} = \frac{\alpha^2}{2\alpha(\alpha+1)(\alpha+2)}.$$

If the order in ϕ is defined as in equation (3.1.3), $\phi = (\theta_1, \theta_2)$ and $s = (1, 2, 1)$,

$$\Pr(s = (1, 2, 1)) = \frac{\alpha^2}{\alpha(\alpha+1)(\alpha+2)},$$

and the configuration $s = (2, 1, 2)$ does not exist.

When $n = 3$, more results are given in following equations. By equation (3.1.4),

$$\Pr(\theta_1 = \theta_3 \neq \theta_2) = \Pr(\theta_1 = \theta_2 \neq \theta_3) = \Pr(\theta_2 = \theta_3 \neq \theta_1) = \frac{\alpha^2}{\alpha(\alpha+1)(\alpha+2)},$$

$$\Pr(\theta_1 \neq \theta_2 \neq \theta_3) = \frac{\alpha^3}{\alpha(\alpha+1)(\alpha+2)},$$

$$\Pr(\theta_1 = \theta_2 = \theta_3) = \frac{2\alpha}{\alpha(\alpha+1)(\alpha+2)}.$$

When $\theta_1 = \theta_3 \neq \theta_2$, by (iii) of Theorem 3.1.2, θ_1 and θ_2 are i.i.d. from G_0. Then we have

$$\begin{aligned}
&\Pr(\theta_1 \in A, \theta_2 \in B, \theta_3 \in C, \theta_1 = \theta_3 \neq \theta_2) \\
=&\Pr(\theta_1 \in A, \theta_2 \in B, \theta_3 \in C | \theta_1 = \theta_3 \neq \theta_2) \Pr(\theta_1 = \theta_3 \neq \theta_2) \\
=&G_0(A \cap C) G_0(C) \frac{2\alpha}{\alpha(\alpha+1)(\alpha+2)}.
\end{aligned}$$

Similarly, we could get

$$\Pr(\theta_1 \in A, \theta_2 \in B, \theta_3 \in C, \theta_1 \neq \theta_2 \neq \theta_3) = \frac{\alpha^3 G_0(A)G_0(B)G_0(C)}{\alpha(\alpha+1)(\alpha+2)}.$$

$$\Pr(\theta_1 \in A, \theta_2 \in B, \theta_3 \in C, \theta_1 = \theta_2 = \theta_3) = \frac{2\alpha G_0(A \cap B \cap C)}{\alpha(\alpha+1)(\alpha+2)}.$$

Summing over all possible g, we could get the joint distribution of $(\theta_1, \theta_2, \theta_3)$ as

$$\begin{aligned}
&\Pr(\theta_1 \in A, \theta_2 \in B, \theta_3 \in C) \\
=&\Pr(\theta_1 \in A, \theta_2 \in B, \theta_3 \in C, \theta_1 = \theta_2 = \theta_3) + \Pr(\theta_1 \in A, \theta_2 \in B, \theta_3 \in C, \theta_1 \neq \theta_2 \neq \theta_3) \\
&+ \Pr(\theta_1 \in A, \theta_2 \in B, \theta_3 \in C, \theta_1 = \theta_2 \neq \theta_3) + \Pr(\theta_1 \in A, \theta_2 \in B, \theta_3 \in C, \theta_3 = \theta_2 \neq \theta_1) \\
&+ \Pr(\theta_1 \in A, \theta_2 \in B, \theta_3 \in C, \theta_1 = \theta_3 \neq \theta_2) \\
=&\left[\alpha(\alpha+1)(\alpha+2)\right]^{-1} \left\{\alpha^3 G_0(A)G_0(B)G_0(C) + 2\alpha G_0(A \cap B \cap C)\right. \\
&\left. +\alpha^2 \left[G_0(A \cap B)G_0(C) + G_0(A \cap C)G_0(B) + G_0(B \cap C)G_0(A)\right]\right\}.
\end{aligned}$$

We shall see that even when $n = 3$, the joint distribution of $\theta_1, \ldots, \theta_n$ is quite complex.

In the "no gaps" algorithm, we shall assume that the index of the distinct elements are allowed to arbitrarily permute. We add the subscript "$-i$" to all notations in the case the observation i is removed. For example, $\theta_{-i} = \{\theta_j : 1 \leq j \leq n, j \neq i\}$, k_{-i} refers to the number of clusters formed by θ_{-i}, ϕ_{-i} denote the set of distinct observations among θ_{-i},and $n_{-i,j}$ represents the number of elements in cluster j when the i-th observation is removed. Furthermore, we will add the superscript "(l)" to all notations to denote the posterior sample at the ith step of the "no gaps" algorithm. For example, $\theta^{(l)} = (\theta_1^{(l)}, \ldots, \theta_n^{(l)})$ denote the posterior sample of the θ vector, $k^{(l)}$ denote the number of distinct values among $\theta^{(l)}$, $\phi^{(l)} = (\phi_1^{(l)}, \ldots, \phi_{k^{(l)}}^{(l)})$ denote the set of distinct observations among $\theta^{(l)}$, and $n_j^{(l)}$ represents the number of elements in cluster j at the lth step of the "no gaps" algorithm.

3.2 "No Gaps" Algorithm

In the DPM model, the observations are $X = \{X_{-m+1}, X_{m+2}, \ldots, X_n\}$ and the model is presented equivalently in the two equations (1.2.5) and (1.2.7). It is more convenient to work with equation (1.2.7) if we want to estimate the interested quantity numerically through MCMC method. We rewrite the model as follows.

$$\begin{aligned} & P \sim D_{\alpha G_0} \\ & \theta_1, \ldots, \theta_n | P \overset{\text{i.i.d.}}{\sim} P \\ & X_i | Z_i, \tau_i, \sigma_i \sim N\left(H(Z_i, \tau_i), \sigma_i^2\right) \text{ for } i \geq 1 \end{aligned} \tag{3.2.1}$$

where $\theta_i = (\tau_i, \sigma_i)$, $\tau_i = (\beta_{0i}, \beta_{1i}, \gamma_i)$, $\gamma_i = (\gamma_{0i}, \ldots, \gamma_{mi})$, $Z_i = (1, X_{i-m}, \ldots, X_{i-1})$ and

$$H(Z_i, \tau_i) = \beta_{0i} + \beta_{1i} \frac{1}{1 + \exp(-\gamma_i^T Z_i)} \tag{3.2.2}$$

Furthermore we assume that $Z_1 = (X_{-m+1}, \ldots, X_0)$ is fixed or has a known distribution and that the distribution G_0 is nonatomic and has density g_0 with respect to Lebesgue measure. We denote the conditional transition density of X_i given Z_i and $\theta_i = (\tau_i, \sigma_i)$ by

$$f(X_i | Z_i, \theta_i) = \phi_{\sigma_i}\left(X_i - H(Z_i, \tau_i)\right). \tag{3.2.3}$$

It is equivalent to viewing that the data are sampled from $f(X_i|Z_i, \theta_i)$ while the time-varying parameter θ_i has a prior joint distribution given in (c) of Theorem 3.4.1 or suggested by the generalized Polya urn sampling scheme as in equation (2.1.2). Thus the random probability measure P is integrated out so that we can design the MCMC algorithm based on its equivalent parametric form when the sample size is finite. As noted in the last section, a key feature of the Dirichlet process is that with positive probability, some of the θ_i's are identical due to the discreteness of the random measure P, and that the marginal prior distribution of θ_i is G_0. Since we assume that G_0 is nonatomic, the ties among θ_i's is not caused by chance when

drawing from G_0 and the distinct elements among θ_i's are i.i.d. from the prior G_0. The joint prior distribution of θ_i's is complex. In developing MCMC algorithms, it is convenient to work with the vector ϕ of distinct elements and the configuration vector s. Appendix 2 gave a short review of several popular MCMC sampling schemes.

Theorem 3.2.1. *Suppose that the Dirichlet Mixture model is given in equation* (3.2.1). *The posterior distribution of θ_i given θ_{-i} and X is a mixture distribution*

$$\theta_i \mid (\theta_{-i}, X) \sim q_{0,i}\, G_{-i}(\theta_i) + \sum_{j \neq i} q_{j,i}\, \delta_{\theta_j}(\theta_i),$$

where the weights satisfy

$$q_{j,i} \propto f(X_i|Z_i, \theta_j),$$
$$q_{0,i} \propto \alpha \int f(X_i|Z_i, \theta) dG_0(\theta),$$
$$\sum_{j \neq i} q_{j,i} + q_{0,i} = 1,$$

and $G_{-i}(\theta_i)$ is the posterior distribution of θ_i given X_i, Z_i based on the prior $G_0(\theta_i)$ and the likelihood $f(X_i|Z_i, \theta_i)$, that is,

$$G_{-i}(\theta_i) \sim \frac{f(X_i|Z_i, \theta_i) g_0(\theta_i)}{\int f(X_i|Z_i, \theta_i) dG_0(\theta_i)}$$

The posterior distribution of θ_i given θ_{-i} and X could be simplified as

$$\theta_i \mid (\theta_{-i}, X) \sim q_0 G_{-i} + \sum_{\phi_k \in \phi_{-i}} n_{-i,k} q_{k,i} \delta_{\phi_k}(\theta_i)$$

where $q_{k,i} \propto f(X_i|Z_i, \phi_k)$.

Proof. By Theorem 3.4.1, the conditional distribution of $\theta_i|\theta_{-i}$ is given by

$$\theta_i|\theta_{-i} \sim \frac{\alpha}{\alpha + n} G_0(\theta_i) + \sum_{j \neq i} \frac{1}{\alpha + n} \delta_{\theta_j}(\theta_i).$$

An application of the Bayes theorem yields that

$$\begin{aligned} dP\left(\theta_i|\theta_{-i}, X\right) &= \frac{f(X|\theta_i, \theta_{-i})dP(\theta_i|\theta_{-i})}{\int_{\theta_i} f(X|\theta_i, \theta_{-i})dP(\theta_i|\theta_{-i})} \\ &\propto f(X_i|Z_i, \theta_i)dP(\theta_i|\theta_{-i}) \\ &\propto \alpha f(X_i|Z_i, \theta_i)dG_0(\theta_i) + \sum_{j\neq i} f(X_i|Z_i, \theta_i)\delta_{\theta_j}(d\theta_i) \\ &\propto q_{0,i}G_{-i}(\theta_i) + \sum_{j\neq i} q_{j,i}\delta_{\theta_j}(d\theta_i) \end{aligned}$$

□

Sampling θ_i from the posterior distribution of θ_i given θ_{-i} and X is equivalent to sampling s_i first given s_{-i}, ϕ_{-i} and X, and then sampling θ_i given s_i, ϕ_{-i}. If $s_i = j \leq k_{-i}$, let the new $\theta_i = \phi_j$ if $s_i = k_{-i} + 1$, then let the new value θ_i be sampled from G_{-i}. This argument leads to the following result.

Corollary 3.2.1. *Under the assumption and notation of Theorem* 3.2.1*, the distribution of s_i given s_{-i} and X is*

$$\begin{aligned} &\Pr\left(s_i = j|s_{-i}, \phi_{-i}, X\right) \propto n_{-i,j}f(X_i|Z_{i-1}, \phi_j), \text{ for } j = 1, \ldots, k_{-i}, \\ &\Pr\left(s_i = k_{-i} + 1|s_{-i}, \phi_{-i}, X\right) \propto \alpha \int f(X_i|Z_{i-1}, \theta)dG_0(\theta). \end{aligned} \tag{3.2.4}$$

Theorem 3.2.2. *Suppose that the Dirichlet Mixture model is given by equation* (3.2.1)*. Let the distinct elements among $\theta = (\theta_1, \ldots, \theta_n)$ be denoted by $\phi = (\phi_1, \ldots, \phi_k)$. Let I_j be the corresponding cluster related to ϕ_j such that $\theta_i = \phi_j$ for $i \in I_j$ and $1 \leq j \leq k$. Then in the posterior distribution, $\phi_1, \ldots, \phi_k$ are conditionally independent given X and s and is given as follows:*

$$Pr(\phi_j|s, X) \propto g_0(\phi_j) \prod_{i\in I_j} f(X_i|Z_i, \phi_j) \text{ for } j = 1, \ldots, k. \tag{3.2.5}$$

Proof. By Theorem 3.1.2, $\phi_1, \ldots, \phi_k$ are i.i.d. according to the law G_0 (with density g_0) given k (or s). By Bayes theorem, the conditional distribution of ϕ given (s, X) is,

$$\begin{aligned} dP(\phi|s, X) &\propto dP(\phi|s) f(X|\phi, s, \sigma) \\ &\propto \left[\prod_{j=1}^{k} dG_0(\phi_j)\right] \left[\prod_{i=1}^{n} f(X_i|Z_i, \phi_{s_i})\right] \\ &\propto \prod_{j=1}^{k} \left[g_0(\phi_j) \prod_{i \in I_j} f(X_i|Z_i, \phi_j)\right] \end{aligned}$$

Hence given s and X, $\phi_1, \ldots, \phi_k$ are conditionally independent with the posterior densities given in equation (3.2.5). □

Given the conditional distributions, the above theorems yield a very simple MCMC scheme for the Dirichlet mixture model.

> Repeat the following steps until the MCMC algorithm converges:
> For $i = 1, \ldots, n$, draw a new value θ_i from the posterior $(\theta_i|\theta_{-i}, X)$ given in Theorem 3.2.1 or equivalently s_i and then θ_i as the method described by Corollary 3.2.1.

This MCMC sampling scheme does not work well for our DPM model. Two possible reasons are cited below. When α is not large enough, the sum $\sum_{j \neq i} q_{i,j}$ would be very large relative to q_{0i}. So it is very unlikely to generate a "new" value of θ_i different from all values in θ_{-i}. Hence the θ vector will get stuck in the algorithm. The first case could be prevented by "remixing" the distinct values ϕ_j, that is, resampling ϕ_j according to Theorem 3.2.2 besides updating the θ vector at each MCMC step. The MCMC scheme with an additional remixing step has been studied by Bush and MacEachern (1996), West, Muller and Escobar (1994) and others. The remixing step is shown to improve the convergence of the MCMC algorithm. Another difficulty is

that it is not easy to calculate $q_{0i} = \alpha \int f(X_i|Z_i,\theta)dG_0(\theta)$ if $g_0(\theta)$ is not conjugate with respect to $f(X_i|Z_i,\theta)$. In the absence of conjugacy, Muller and MacEachern (1998) developed the "no gaps" algorithm in which the parameter space is cleverly augmented. Neal (2000) developed algorithms based on the Metropolis-Hasting sampling scheme. In our numerical demonstration, we shall use Muller and MacEachern's (1998) "no gaps" sampling scheme.

In the "no gaps" algorithm, the ϕ vector is augmented to

$$\{\underbrace{\phi_1,\ldots,\phi_k}_{\phi_F},\underbrace{\phi_{k+1},\ldots,\phi_n}_{\phi_E}\}$$

Muller and MacEachern refer the $\phi_F = (\phi_1,\ldots,\phi_k)$ as "full" clusters, which are composed of the k distinct values among θ, and $\phi_E = (\phi_{k+1},\ldots,\phi_n)$ as "empty" clusters or "potential" clusters. The augmentation relies upon the constraint that there be "no gaps" in the values of the s_i, that is, $n_j > 0$ for $j = 1,\ldots,k$ and $n_j = 0$ for $j = k+1,\ldots,n$. In the "no gaps", it is allowed to permute the index of $\phi_F = (\phi_1,\ldots,\phi_k)$ arbitrarily. There are total $k!$ possible pairs of (ϕ_F, s). Equal probabilities are assigned to each of the $k!$ permutations.

We should note that the conditional distributions of s_i given (s_{-i}, X, ϕ) are different from the one given (s_{-i}, X, ϕ_{-i}) as in Corollary 3.2.1. The reason is that once a new candidate $\phi_{k_{-i}+1}$ for θ_i is generated, no permutation is performed; however, in the "no gaps" algorithm the probability is calculated under the assumption that the new candidate from ϕ_E is permuted with those in ϕ_F.

Theorem 3.2.3. (a) *In the "no gaps" algorithm, the conditional distribution of* s_i *given* s_{-i} *and* $\phi = \phi_F \cup \phi_E$ [†] *is*

$$\Pr(s_i = j|s_{-i},\phi) \propto n_{-ij} \text{ for } j = 1,\ldots,k_{-i}$$

[†]The mass is calculated after ϕ_F is permuted.

$$\Pr(s_i = n_{-ij} + 1 | s_{-i}, \phi) \propto \frac{\alpha}{k_{-i} + 1}.$$

if all $n_{-ij} \geq 1$ *for* $j = 1, \ldots, k_{-i}$. *The conditional distribution is degenerated if* $n_{-ij} = 0$ *for some* j *under the "no gaps" constraint, that is,*

$$\Pr(s_i = j | s_{-i}, \phi) = 1.$$

(b) *The conditional distribution of* s_i *given* s_{-i}, ϕ *and* X *is*

$$\Pr\left(s_i = j | s_{-i}, \phi, X\right) \propto \Pr(s_i = j | s_i, \phi) f(X_i | X_{i-1}, \phi_j)$$

for $j = 1, \cdots, k_{-i} + 1$, *where* $\Pr(s_i = j | s_{-i}, \phi)$ *is defined in (a).*

Proof. (a) An application of the Bayes theorem yields that

$$\Pr(s_i = j | s_{-i}, \phi) = \frac{\Pr(s_i = j, s_{-i} | \phi)}{\sum_j \Pr(s_i = j, s_{-i} | \phi)} \propto \Pr(s_i = j, s_{-i} | \phi).$$

By Theorem 3.1.2,

$$\Pr(s = (s_1, \ldots, s_n) | \phi) = \frac{\alpha^k \prod_{j=1}^{k} (n_j - 1)!}{\alpha^{[n]} k!}$$

In the case $n_{-ij} \geq 1$ for $j = 1, \ldots, k_{-i}$, when $j = 1, \ldots, k_{-i}$, there are only k_{-i} distinct values, so

$$\Pr(s_i = j, s_{-i} | \phi) = \frac{\alpha^{k_{-i}} \prod_{j=1}^{k_{-i}} (n_{-ij} - 1)!}{\alpha^{[n]} k_{-i}!}.$$

When $j = k_{-i} + 1$, there are $k_{-i} + 1$ distinct values. The new s_i form a cluster with size 1, so

$$\Pr(s_i = k_{-i} + 1, s_{-i} | \phi) = \frac{\alpha^{k_{-i}+1} \prod_{j=1}^{k_{-i}} (n_j - 1)!}{\alpha^{[n]} (k_{-i} + 1)!}.$$

Thus

$$\Pr(s_i = j | s_{-i}, \phi) \propto \Pr(s_i = j, s_{-i} | \phi)$$

$$\propto \begin{cases} n_{-ij} & \text{if } j \leq k_{-i}, \\ \frac{\alpha}{k_{-i}+1} & \text{if } j = k_{-i} + 1. \end{cases}$$

In the case $n_{-ij} = 0$ for some j, the "no gaps" constraint makes the distribution of $(s_i|s_{-i}, \phi)$ degenerate.

(b) By the Bayes theorem,

$$\begin{aligned}\Pr(s_i = j|s_{-i}, \phi, X) &\propto f(X|s_i = j, s_{-i}, \phi) \Pr(s_i = j|s_{-i}, \phi) \\ &\propto \Pr(s_i = j|s_{-i}, \phi) f(X_i|Z_i, \phi_j).\end{aligned}$$

□

"No gaps" Algorithm One (The Version Telling the Idea):

Initialize the s and (ϕ_F, ϕ_E) and update them according to following mechanism until the algorithm converges.

i) Repeating (ia) and (ib) for $i = 1, \ldots, n$.
   ia) Given ϕ_F or $\theta = (\theta_1, \ldots, \theta_n)$, randomly permute the index of ϕ_F , with each permutation having probability $1/k!$.
   ib) Sample $(s_i|s_{-i}, \phi, X)$ according to Theorem 3.2.3.

ii) Sample $(\phi_i|s, X)$ as in theorem (3.2.2) for $i = 1, \ldots, n$ since ϕ_i's are conditional independent given s, X. $\phi_{k+1}, \ldots, \phi_n$ could be sampled directly from G_0 since $n_j = 0$ for $j = k + 1, \ldots, n$. From here, we should note that, in (ib) of each MCMC step, once a new θ value is generated, it is from the law G_0, not G_i defined in Theorem 3.2.1.

Note that in the step (ia), the permutation of the index ϕ_F and s does not change the values in the θ vector. The implementation of this algorithm may be simplified and speeded by discarding unnecessary draws that do not alter the chain itself. So in step (i) the permutation (ia) will be performed only when $n_{s_i} = 1$. In this case, $s_i = k$ with probability $1/k$ and $s_i < k$ with probability $1-1/k$. The former case leads to a nondegenerate posterior conditional distribution $\Pr(s_i|s_{-i}, \phi, X)$, as in Theorem 3.2.3. The latter case leads to a degenerate posterior for s_i. Also, we note that in

a typical cycle of the algorithm, most of the $\phi_j \in \phi_E$ will not be used. So they are generated only when needed. Thus the "no gaps" algorithm may be simplified as follows.

"No gaps" Algorithm (Simplified Version):

Initialize the s and ϕ_F and update them according to the following mechanism until the algorithm converges.

i) Set ϕ_E empty. For $i = 1, \ldots, n$, repeating (ia) and (ib)

(ia) If $n_{s_i} > 1$, then $k_{-i} = k$. If ϕ_E is empty, draw a new value ϕ_{k+1} from G_0 and add it to ϕ_E. Sample s_i according to non-degenerated posterior distribution $\Pr(s_i|s_{-i}, \phi, X)$ in Theorem 3.2.3. If the new $s_i \leq k$, ϕ_F are unchanged. If the new $s_i = k + 1$, move the first element from ϕ_E to ϕ_F, so $\phi_F = (\phi_1, \ldots, \phi_{k+1})$.

(ib) If $n_{s_i} = 1$, $k_{-i} = k - 1$. With probability $1 - 1/k$, leave s_i unchanged. Nothing is done. Otherwise relabel clusters such that $s_i = k$ and then resample s_i according to the non-degenerate posterior distribution $\Pr(s_i|s_{-i}, \phi, X)$ in Theorem 3.2.3. So if the new s_i happened to be equal $k_{-i}+1 = k$, then the preceding relabeling kept the previous values of θ_i as ϕ_k and nothing is changed except possible relabing of ϕ_F and hence s. If the new $s_i \leq k_{-i}$, the last element after relabeling in ϕ_F is moved to ϕ_E.

ii) Only draw $\phi_i|(s, X)$ as in Theorem 3.2.2 for $i = 1, \ldots, k$.

In the implementation of the MCMC sampling scheme, it is critical for the algorithm to converge to the posterior distribution. Otherwise the strong law of large numbers do not hold for the Markov process designed in the MCMC algorithm, which leads to a wrong Bayes estimate based on the sample from the MCMC. Muller and MacEachern (1997, 1998) showed that their "no gaps" algorithm converges to the

posterior distribution under some mild sufficient conditions.

In the "no gaps" algorithm, the state space of the designed Markov process is $s = (s_1, \ldots, s_n)$ and $\phi = (\phi_F, \phi_E)$. Given n, the number of all possible configuration vectors s is finite. The prior distribution of s is given in Theorem 3.1.2. The distinct elements ϕ_i $(i = 1, \ldots, n)$ are i.i.d. from the prior G_0 with density g_0. Given s and ϕ, the joint distribution of X is $\prod_{i=1}^{n} f(X_i|Z_i, \phi_{s_i})$. Thus the posterior distribution of s and ϕ is

$$\pi(s, \phi|X) \propto \Pr(s) \left[\prod_{i=1}^{n} g_0(\phi_i)\right] \left[\prod_{i=1}^{n} f(X_i|Z_i, \phi_{s_i})\right]. \tag{3.2.6}$$

The "no gaps" is designed specially in the case when $G_0(\theta_i)$ is not conjugated with the kernel $f(X_i|Z_i, \theta_i)$. Let the support of G_0 be denoted by Θ which is a subset of the Euclidean space $\mathbb{R}^{m+2} \times \mathbb{R}^+ \times \mathbb{R}^+$. Suppose that G_0 has density g_0 which is positive for $\theta \in \Theta$. In the "no gaps", we need to update the ϕ_i sequentially according to their posterior, given in Theorem 3.2.2. Typically, we need to turn to the Metropolis-Hasting sampler (the Gibbs sampler is a special case) as introduced in Appendix 2. Suppose further that each ϕ_i vector could be decomposed to $\eta_{i,1}, \ldots, \eta_{i,l}$ components (η_{ij} may be a vector or scalar) and we update each $\eta_{i,j}$ for $j = 1, \ldots, l$ and $i = 1, \ldots, k$ sequentially in a fixed order with the Metropolis-Hasting sampler. For each $\eta_{i,j}$, its posterior distribution given X and s and other $\eta_{i,j}$'s is

$$f(\eta_{i,j}|X, s, \eta_{i,1}, \ldots, \eta_{i,j-1}, \eta_{i,j+1}, \ldots, \eta_{i,l}) \propto f(\phi_i|X, s),$$

where $f(\phi_i|X, s)$ is given in Theorem 3.2.2 and is positive everywhere for $\phi_i \in \Theta$. When updating $\eta_{i,j}$, we sample a candidate $\widetilde{\eta}_{i,j}$ from a proposed distribution

$$q_{ij}(\eta_{i,j}, \widetilde{\eta}_{i,j}) = q(\widetilde{\eta}_{i,j}|s, \eta_{i,1}, \ldots, \eta_{i,j-1}, \eta_{i,j}, \eta_{i,j+1}, \eta_{i,l}) \tag{3.2.7}$$

such that q_{ij} is positive everywhere. Then with probability

$$\alpha_{ij}(\eta_{i,j}, \widetilde{\eta}_{i,j}) = \min\left\{1, \frac{f(\widetilde{\eta}_{i,j}|X, s, \eta_{i,1}, \ldots, \eta_{i,j-1}, \eta_{i,j+1}, \ldots, \eta_{i,l}) q_{ij}(\eta_{i,j}, \widetilde{\eta}_{i,j})}{f(\eta_{i,j}|X, s, \eta_{i,1}, \ldots, \eta_{i,j-1}, \eta_{i,j+1}, \ldots, \eta_{i,l}) q_{ij}(\widetilde{\eta}_{i,j}, \eta_{i,j})}\right\}. \tag{3.2.8}$$

we update $\eta_{i,j}$ as $\widetilde{\eta}_{i,j}$ and otherwise we leave $\eta_{i,j}$ unchanged. The special case of the Metropolis-Hasting sampler is the Gibbs sampler in which the proposed distribution $q_{ij} = f(\eta_{i,j}|X, s, \eta_{i,1}, \ldots, \eta_{i,j-1}, \eta_{i,j+1}, \ldots, \eta_{i,l})$ and with acceptance probability $\alpha_{ij} = 1$ the $\eta_{i,j}$ is replaced by $\widetilde{\eta}_{i,j}$. In application, the assumption given in this paragraph is very mild. However, these assumptions do ensure that the "no gaps" converges almost surely to the target (posterior) distribution.

Theorem 3.2.4. *Suppose that the center measure G_0 has a density g_0 which is positive everywhere in its support Θ. Assume that the ϕ_i vector could be decomposed into $\eta_{i,1}, \ldots, \eta_{i,l}$ components which are updated sequentially in a fixed order by the Metropolis-Hasting sampler (Gibbs sampler is a special case) with proposed distributions q_{ij} given in equation* (3.2.7) *which are everywhere positive. Further assume that s_i is sampled directly from the conditional distribution as in Theorem* 3.2.3.

(a) *Then the Markov kernel Q defined after a cycle of updating all parameters $\omega = (s, \phi)$ of the algorithm is irreducible, aperiodic and positive with unique invariant distribution π as in equation* (3.2.6).

(b) *For π-almost all initial starting point $\omega^{(0)} = \left(s^{(0)}, \phi^{(0)}\right)$,*

$$\left\| Q^n\left(\omega^{(0)}, \cdot\right) - \pi \right\| \to 0,$$

where Q^n is the n-step transition kernel of Q. That is, the algorithm converges to the posterior distribution almost surely in total variation distance.

(c) *For π-almost all initial starting point, and any π-integrable function f,*

$$\frac{1}{n}\sum_{l=1}^{n} f\left(\omega^{(l)}\right) \to \int f d\pi \ \ a.s.\, \pi,$$

where $\omega^{(l)} = \left(s^{(l)}, \phi^{(l)}\right)$ is the state of chain at end of l-th cycle.

Proof. The theorem follows from Theorem A.1.3 if all three conditions of Theorem A.1.3 could be verified. Since the "no gaps" algorithm consists of Gibbs and Metropolis steps, and each step defines an aperiodic and irreducible sub-Markov kernel, the

three conditions will be satisfied according to Tierney (1994). The details are given as follows.

(i) To verify that π is one of the invariant distribution of the cycle kernel Q defined by the "no gaps" algorithm. This condition is automatically true since each step of (ia) (ib) and (ii) of the "no gaps" is typically a Gibbs sampler or Metropolis-Hasting sampler. Appendix 2 gives the brief argument.

(ii) To verify that Q is π-irreducible. We shall show for any A such that $\pi(A) > 0$, $Q(\omega^{(0)}, A) > 0$ for each initial $\omega^{(0)} = \left(s^{(0)}, \phi^{(0)}\right)$. Thus Q will be π-irreducible. Any A may be partitioned as $A = \cup_s A_s$, where the elements of the partition are indexed by the configuration vector. Also, the invariant distribution has a unique representation as $\pi = \sum_s \pi_s$, where

$$\pi_s(d\phi_1, \ldots, d\phi_n) \propto p(s) \left[\prod_{i=1}^{n} g_0(\phi_i)\right] \left[\prod_{i=1}^{n} f(X_i|Z_{i-1}, \phi_{s_i})\right] [d\phi_1 \ldots d\phi_n] .$$

so that $\pi(A) = \sum_s \pi_s(A_s)$ and there exists some A_{s^*} for which $\pi_{s^*}(A_{s^*}) > 0$.

Note that n is fixed. Given n, there are only a finite number of the possible configuration vectors s which receive positive prior probability. The first stage (i) of the "no gaps" algorithm involves the generation of a new configuration s through a sequence of smaller generations. Given any ϕ and X, $f(X_i|Z_i, \phi_j)$ will always be positive. So after n steps of repeating (ia) and (ib), there is a positive probability of a transition to each vector s which receives positive prior probability. So given any initial $\omega^{(0)} = \left(s^{(0)}, \phi^{(0)}\right)$, there is a positive probability for the chain to move to s^*.

At the second stage (ii) of the "no gaps" algorithm, we focus on the generation of new ϕ_i $(i = 1, \ldots, k)$ by Metropolis-Hastings sampler (Gibbs sampler is a special case). Given the old ϕ_i and the new s^*, the updating is implemented sequentially in a fixed order as $\eta_{i,j}$ for $j = 1, \ldots, l$ and $i = 1, \ldots, k$. When we update a parameter, say η_{ij}, we sample a candidate $\widetilde{\eta}_{ij}$ from the proposal transition kernel with density q_{ij} which may depend on other η_{ij}'s and accept it with probability $\alpha(\eta_{ij}, \widetilde{\eta}_{ij})$ as in equation

(3.2.8). The component transition kernel is

$$\begin{aligned}
&P_{ij}(\eta_{ij}, d\widetilde{\eta}_{ij}|X, S^*, \eta_{i,1}, \ldots, \eta_{i,j-1}, \eta_{i,j+1}, \ldots, \eta_{i,l}) \\
=&q_{ij}(\eta_{ij}, d\widetilde{\eta}_{ij})\alpha_{ij}(\eta_{ij}, \widetilde{\eta}_{ij})d\widetilde{\eta}_{ij} + \left[1 - \int q_{ij}(\eta_{ij}, d\widetilde{\eta}_{ij})\alpha_{ij}(\eta_{ij}, \widetilde{\eta}_{ij})d\widetilde{\eta}_{ij}\right]\delta_{\eta_{ij}}(d\widetilde{\eta}_{ij}) \\
\geq&q_{ij}(\eta_{ij}, d\widetilde{\eta}_{ij})\alpha_{ij}(\eta_{ij}, \widetilde{\eta}_{ij})d\widetilde{\eta}_{ij}
\end{aligned}$$

the same way as that in equation (B.3). Thus given (s^*, X), the overall transition after the updating of $\eta_{i,j}$ for $j = 1, \ldots, l$ and $i = 1, \ldots, k$ in step (ii) of the no gaps satisfies

$$\begin{aligned}
&P(\phi, d\widetilde{\phi}|X, s^*) \\
=&\prod_{i=1}^{k}\prod_{j=1}^{l}\left[P_{ij}(\eta_{i1}, d\widetilde{\eta}_{ij}|X, s^*, d\widetilde{\eta}_{i,1}, \ldots, d\widetilde{\eta}_{i,j-1}, \eta_{i,j+1}, \ldots, \eta_{i,l})\right] \\
\geq&\prod_{i=1}^{k}\prod_{j=1}^{l}\left[q_{ij}(\eta_{ij}, d\widetilde{\eta}_{ij})\alpha_{ij}(\eta_{ij}, \widetilde{\eta}_{ij})d\widetilde{\eta}_{ij}\right].
\end{aligned}$$

So its absolutely continuous part has density everywhere positive as suggested in above equation and is hence dominated by the invariant measure π_{s^*} since π_{s^*} is absolutely continuous. So $\pi_{s^*}(A_{s^*}) > 0$ implies that $\Pr(\phi \in A_{s^*}|X, s^*) > 0$. Hence

$$Q(\omega^{(0)}, A) \geq Q(\omega^{(0)}, A_{s^*}) = \Pr(s^*|\phi^{(0)}, s^{(0)})\Pr(\phi \in A_{s^*}|s^*) > 0.$$

(iii) To verify that the chain is aperiodic.

Suppose that the chain is not aperiodic. Then by definition, the overall parameter space could be divided into a sequence $\{E_0, E_1, \cdots, E_{d-1}\}$ $(d \geq 2)$ of d nonempty disjointed sets such that for all $\omega = (\phi, s) \in E_i$,

$$Q(\omega, E_j) = 1 \text{ for } j \equiv i + 1 \pmod{d}.$$

Hence for any i and any $\omega = (\phi, s) \in E_i$, $Q(\omega, E_i) = 0$. We note as in the verification of the irreducibility condition in step (ii), that if $\pi(A) > 0$, then for any $\omega = (\phi, s)$,

$Q(\omega, A) > 0$. So the equation $Q(\omega, E_i) = 0$ leads to $\pi(E_i) = 0$ for any i. We conclude that $\pi(\cup_{i=0}^{d-1} E_i) = 0$. This contradicts the fact that π is an invariant probability measure. So the chain is aperiodic. □

3.3 "No Gaps" Algorithm with a Specified Prior

In the last section, we have discussed the "no gaps" algorithm and its convergence property. To complete our model, we need to specify the prior, that is, the parameter α and G_0 of the Dirichlet process. When we implement the Dirichlet mixture model, we specify a center measure G_0 in which all the parameters $\theta = (\beta_0, \beta_1, \gamma_0, \ldots, \gamma_{m+1}, \sigma)$ are independent with distribution as follows.

$$\begin{aligned} &\beta_0 \sim N(u_0, V_\beta), \\ &\beta_1 \sim N(0, V_\beta), \\ &\gamma_j \sim N(0, V_{\gamma_1}) \text{ for } j = 1, \ldots, m, \\ &\log(\gamma_{m+1}) \sim N(0, V_{\gamma_2}), \\ &\frac{1}{\sigma^2} \sim \mathcal{G}(a, b). \end{aligned} \tag{3.3.1}$$

where $N(a, b)$ is the normal distribution with mean a and variance b and $\mathcal{G}(a, b)$ is the gamma distribution with shape parameter a and inverse scale parameter b. When implementing the "no gaps" algorithm, one needs to specify the constants α, u_0, V_β, V_{γ_1}, V_{γ_2}, a and b.

Lemma 3.3.1. *Given h, β and X, the random vector Y has multivariate normal distribution $N(X\beta, h^{-1}I)$. If the prior distribution of β is $N(u, V)$, then the posterior distribution of β given X, Y and h is the multivariate normal with mean vector $b_h = B_h(hX^TY + V^{-1}u)$ and variance matrix B_h where $B_h = (V^{-1} + hX^TX)^{-1}$.*

Proof.

$$\begin{aligned} f(\beta|h,X,Y) &\propto f(\beta)f(Y|\beta,X,h) \\ &\propto \exp\left[-\frac{(\beta-u)^T V^{-1}(\beta-u) - h(y-X\beta)^T(y-X\beta)}{2}\right] \\ &\propto \exp\left[-\frac{(\beta-b_h)^T B_h^{-1}(\beta-b_h)}{2}\right] \end{aligned}$$

So the conditional posterior distribution of β is $N(b_h, B_h)$. □

The Complete "No Gaps" Algorithm:

Initialize the number of distinct elements k and the configuration vector $s = (s_1, \ldots, s_n)$ and ϕ_F. Repeating the following step until the algorithm converges.

i) Set $\phi = \phi_F \cup \phi_E$. ϕ_E comes after ϕ_F. Empty ϕ_E first. Elements will be added to ϕ_E only when needed. For $i = 1, \ldots, n$, repeat (ia) and (ib).

(ia) If $n_{s_i} > 1$, $k_{-i} = k$. If ϕ_E is empty, draw a new value from G_0 and add it to the first position of ϕ_E. Obviously ϕ_{k+1} is the first value in ϕ_E at this time. Resample s_i according to the following multinomial distribution.

$$s_i = \begin{cases} j \text{ for } j = 1, \ldots, k_{-i} & \text{with probability} \quad n_{-ij} f(X_i|Z_i,\theta_i), \\ k_{-i}+1 & \text{with probability} \quad \frac{\alpha}{\alpha+k_{-i}+1} f(X_i|Z_i,\theta_i). \end{cases} \tag{3.3.2}$$

If the new $s_i \leq k$, ϕ_F and k are kept unchanged. But one new element may be added to ϕ_E depending on whether it was previously empty or not. If the new $s_i = k+1$, move the first element from ϕ_E to ϕ_F and change k to $k+1$ and the configuration vector correspondingly. The vector ϕ_E loses one element if it was not empty previously and remains empty if it was empty previously.

(ib) If $n_{s_i} = 1$, $k_{-i} = k-1$. With probability $1 - 1/k$ leave s_i unchanged.

Nothing is done. Otherwise relabel the indices of ϕ_F and change the configuration vector s accordingly such that $s_i = k$ and $\phi_k = \theta_i$, and then resample s_i according to the multinomial distribution given in equation (3.3.2). If the new s_i happens to be equal $k_{-i} + 1 = k$, then the preceding relabeling keeps the previous values of θ_i as ϕ_k. The values in $\phi_F = (\phi_1, \ldots, \phi_k)$ are never changed in this step except for the possible relabeling of the indices. If the new $s_i \leq k_{-i}$, ϕ_F becomes $\phi_F = (\phi_1, \ldots, \phi_{k-1})$, and add ϕ_k to the first position in the set ϕ_E since it may be used in step (ia) later.

ii) Updating ϕ_j $(j = 1, \ldots, k)$ sequentially according to the following scheme: Let I_j, $j = 1, \ldots, k$, denote the clusters in which $i \in I_j$ if and only if $\theta_i = \pi_j$.

(iia) Updating σ_j:

Note that the prior for σ_j^2 is the inverse gamma distribution. The posterior of $h_j = 1/\sigma_j^2$ is

$$\begin{aligned} f(h_j|X, s, \beta_j, \gamma_j) \propto & g_0(h_j) \prod_{i \in I_j} f(X_i|Z_i, \beta_j, \gamma_j) \\ \propto & h_j^{a-1} e^{-bh_j} \prod_{i \in I_j} \left[h_j^{\frac{1}{2}} e^{-\frac{h_j (X_i - H(Z_i, \beta_j, \gamma_j))^2}{2}} \right] \\ \propto & h_j^{a+n_j-1} e^{-h_j[b + SSE_j/2]} \end{aligned} \tag{3.3.3}$$

where n_j is the size of the jth cluster, and $SSE_j = \sum_{i \in I_j} (X_i - H(Z_i, \beta_j, \gamma_j))^2$ is the sum of squares of the error of the jth cluster. So the posterior of $h_j = 1/\sigma_j^2$ is a gamma distribution $\mathcal{G}(a + n_j, b + SSE_j/2)$. It is straightforward to update σ_j with a Gibbs sampler.

(iib) updating $\beta_j = (\beta_{j0}, \beta_{j1})^T$:

Since the prior for β_j is a bivariate normal distribution with mean vector $u_\beta = (u_0, 0)^T$ and covariate matrix $V_\beta I_2$ where I_j is the $j \times j$ identity matrix, and $(X_m, m \in I_j)$ are conditional normal given $(\gamma_j, \sigma_j, Z_m)$, by

Lemma 3.3.1, the posterior of β_j is a bivariate normal distribution with covariance matrix

$$B_j = \left(\frac{1}{V_\beta} I_2 + \frac{1}{\sigma_j^2} S_j^T S_j \right)^{-1},$$

and mean vector $U_j = B_j(\frac{1}{\sigma_j^2} S_j^T Y_j + \frac{1}{V_\beta} u_\beta)$, where

$$S_j^T = \begin{pmatrix} 1 & \cdots & 1 \\ \left[1 + \exp(-\gamma_j^T Z_{j_1})\right]^{-1} & \cdots & \left[1 + \exp(-\gamma_j^T Z_{j_{n_j}})\right]^{-1} \end{pmatrix},$$

and $j_1, j_2, \ldots, j_{n_j}$ are the n_j elements in I_j. The β_j is updated with a Gibbs sampler.

(iic) Updating $\gamma_j = (\gamma_{j0}, \ldots, \gamma_{j,m-1}, \gamma_{jm})^T$:

Updating γ_j is equivalent to update $\gamma_j^* = [\gamma_{j0}, \ldots, \gamma_{j,m-1}, \gamma_{jm}^*]^T$ where $\gamma_{jm}^* = \log(\gamma_{jm})$. Recall that the prior for γ_j^* is a multivariate normal distribution with mean vector 0 and covariance matrix $V_\gamma = \text{diag}(V_{\gamma_1}, \ldots, V_{\gamma_1}, V_{\gamma_2})$. The posterior distribution of γ_j^* given β_j, σ_j and X_m's is

$$f(\gamma_j^* | \beta_j, \sigma_j, X) \propto g_0(\gamma_j^*) \prod_{i \in I_j} f(X_i | Z_i, \beta_j, \sigma_j, \gamma_j^*), \tag{3.3.4}$$

where $g_0(\gamma_j^*)$ is the density of the multivariate normal distribution $N(0, V_\gamma)$. The complex conditional distribution function does not allow efficient random variate generation to implement a Gibbs sampling step. Instead we realize a random walk chain step. Generate a candidate $\tilde{\gamma}_j^* = (\tilde{\gamma}_{j0}, \ldots, \tilde{\gamma}_{jm-1}, \tilde{\gamma}_{jm}^*)$ from $q(\gamma_j^*, \tilde{\gamma}_j^*)$, the multivariate normal distribution with mean vector γ_j^* and covariance matrix $V_\gamma^* = cV_\gamma = \text{diag}(cV_{\gamma_1}, \cdots, cV_{\gamma_1}, cV_{\gamma_2})$. The acceptance probability is

$$\begin{aligned} \alpha(\gamma_j^*, \tilde{\gamma}_j^*) &= \min\left\{1, \frac{f(\tilde{\gamma}_j^* | \beta_j, \sigma_j, X) q(\tilde{\gamma}_j^*, \gamma_j^*)}{f(\gamma_j^* | \beta_j, \sigma_j, X) q(\gamma_j^*, \tilde{\gamma}_j^*)}\right\} \\ &= \min\left\{1, \frac{f(\tilde{\gamma}_j^* | \beta_j, \sigma_j, X)}{f(\gamma_j^* | \beta_j, \sigma_j, X)}\right\}. \end{aligned}$$

where $f(\gamma_j^*|\beta_j, \sigma_j, X)$ is defined in equation (3.3.4). With probability $\alpha(\gamma_j^*, \tilde{\gamma}_j^*)$ replace γ_j^* by $\tilde{\gamma}_j^*$; otherwise keep γ_j^*. In practice, the value of c is determined automatically by the program such that the average acceptance probability is between 0.20 and 0.65. The choice of c makes a compromise between the jump distance in the parameter space and the acceptance frequency; both of them ensure the efficiency of the MCMC algorithm.

3.4 Bayes Estimate

Theorem 3.4.1. *Suppose that the Dirichlet Mixture model is given by equation* (3.2.1) *and* $g(\theta) = g(\tau, \sigma)$ *is a measurable function.*
The Bayes estimate of $\int g(\theta)\, dP(\theta)$ *given* $X = \{X_{-m+1}, \ldots, X_n\}$ *under squared error loss is*

$$\mathrm{E}\left(\int g(\theta)\, dP(\theta) \,\middle|\, X\right) = \frac{\alpha \int g(\theta)\, dG_0(\theta) + \mathrm{E}\left[\sum_{i=1}^n g(\theta_i) \,\middle|\, X\right]}{\alpha + n}.$$

Proof. Since the joint distribution of X, given $(\theta_1, \ldots, \theta_n)$ and P, is free of P, the distribution of P, given $(\theta_1, \ldots, \theta_n)$ and X depends on $(\theta_1, \ldots, \theta_n)$ only. By (a) of Theorem 3.4.1, given $(\theta_1, \ldots, \theta_n)$ the distribution of P is Dirichlet process $D_{G_{0n}}$ where $G_{0n} = \alpha G_0 + \sum_{i=1}^n \delta_{\theta_i}$. A further use of Theorem 3.4.1 leads to

$$\begin{aligned}\mathrm{E}\left(\int g(\theta)\, dP(\theta) \,\middle|\, \theta_1, \ldots, \theta_n, X\right) =& \mathrm{E}\left(\int g(\theta)\, dP(\theta) \,\middle|\, \theta_1, \ldots, \theta_n\right) \\ =& \frac{\alpha \int g dG_0 + \sum_{i=1}^n g(\theta_i)}{\alpha + n}.\end{aligned}$$

Therefore

$$\mathrm{E}\left(\int g(\theta)\,dP(\theta)\;\middle|\;X\right) = \mathrm{E}\left[\mathrm{E}\left(\int g(\theta)\,dP(\theta)\;\middle|\;\theta_1,\ldots,\theta_n,\,X\right)\;\middle|\;X\right]$$
$$= \frac{\alpha\int g(\theta)\,dG_0(\theta) + \mathrm{E}\left[\sum_{i=1}^{n} g(\theta_i)\;\middle|\;X\right]}{\alpha+n}.$$

□

Corollary 3.4.1. *Suppose that the Dirichlet Mixture model is given by equation* (3.2.1) *and the conditions of Theorem* 3.2.4 *hold. If* $\mathrm{E}\left(\int g(\theta)dP(\theta)|X\right)$, *the Bayes estimate of* $\int g(\theta)dP(\theta)$, *exists, then for almost all starting values, given any fixed sample* X,

$$\frac{\alpha\int g(\theta)dG_0(\theta)}{\alpha+n} + \frac{\sum_{i=1}^{M}\left[\sum_{j=1}^{k^{(i)}} n_j^{(i)} g(\phi_j^{(i)})\right]}{M(\alpha+n)} \to E\left(\int g(\theta)dP(\theta)|X\right) \;\; a.s.$$

as $M\to\infty$.

Proof. As shown in Theorem 3.4.1, the Bayes estiamte of $\int g(\theta)dP(\theta)$ is

$$\mathrm{E}\left(\int g(\theta)dP(\theta)\;\middle|\;X\right) = \frac{\alpha\int g(\theta)dG_0(\theta) + \mathrm{E}\left(\sum_{i=1}^{n} g(\theta_i)|X\right)}{\alpha+n}.$$

A use of part (c) of Theorem 3.2.4 leads to the conclusion. □

We are interested in the predictive distribution of X_{n+1} and prediction of the future value X_{n+1} given all the past. Under squared error loss, the Bayes estimate is just the posterior mean. Corollary 3.4.1 is one of the key results for us to estimate the quantity of interest. In Corollary 3.4.1, the posterior sample at every step is used in the approximation. In practice, we may use the posterior sample after the "burn-in" period, at which the algorithm has reached convergence approximately. This will improve the approximation. Also, we may use the posterior sample every r steps to

reduce the variance of the Bayes estimate given that the same M is used in (3.4.2) since the covariance between rth lagged value tends to be quite small in a Markov process.

The Bayes estimate of the transition kernel is equal to the posterior expected transition density, that is,

$$\hat{f}(x|z) = \mathrm{E}\left(\int f(x|z,\theta) dP(\theta) \;\middle|\; X_1, X_2, \ldots, X_n \right) \tag{3.4.1}$$

where $f(x|z,\theta)$ is given in equation (3.2.3). By Corollary 3.4.1, the Bayes estimate of the transition kernel is approximated by

$$\hat{f}(x|z) = \frac{\alpha \int f(x|z,\theta) dG_0(\theta)}{\alpha + n} + \frac{\sum_{i=1}^{M} \left[\sum_{j=1}^{k^{(N+ri)}} n_j^{(N+ri)} f\left(x|z, \phi_j^{(N+ri)} \right) \right]}{M(\alpha + n)} \tag{3.4.2}$$

where N is the number of steps in the "burn in" period, and M is a large integer. Suppose that the prior is given in equation (3.3.1). Given Z and $\theta = (\beta_0, \beta_1, \gamma, \sigma)$, X has a normal distribution with mean $\beta_0 + \beta_1/[1+\exp(-\gamma^T Z_{n+1})]$ and variance σ^2, and (β_0, β_1) has a bivariate normal distribution with mean $(u_0, 0)^T$ and covariance matrix $V_\beta I_2$. So given Z, γ, σ, the distribution of X is normal with mean u_0 and variance $V_g = \sigma^2 + V_\beta(1 + 1/[1 + \exp(-\gamma^T Z_{n+1})]^2)$. The integration in the first part of equation (3.4.2) can be simplified as

$$\int f(x|z,\theta) dG_0(\theta) = \int \frac{1}{\sqrt{2\pi} V_g} \exp\left[-\frac{(x-u_0)^2}{2V_g} \right] dG_0(\gamma, \sigma). \tag{3.4.3}$$

One can simulate N_2 pairs of (γ_i, σ_i) from the posterior and get the corresponding $V_{gi} = \sigma_i^2 + V_\beta(1 + 1/[1 + \exp(-\gamma_i^T Z_{n+1})]^2)$. By the strong law of large numbers, the integration could be approximated by

$$\int f(x|z,\theta) dG_0(\theta) \cong N_2^{-1} \sum_{i=1}^{N_2} \left(\frac{1}{\sqrt{2\pi} V_{gi}} \exp\left[-\frac{(x-u_0)^2}{2V_{gi}} \right] \right). \tag{3.4.4}$$

The density of the predictive distribution of X_{n+1} given $X_0, \ldots, X_n$ is just $\hat{f}(X_{n+1}|Z_{n+1})$. Under the squared error loss, the Bayes predictor of X_{n+1}, is just the conditional mean from $\hat{f}(X_{n+1}|Z_{n+1})$ as in following equation,

$$\begin{aligned}\mu_{n+1} =& \mathrm{E}\left(\int_\theta \int_x x f(x|Z_{n+1},\theta) dx dP(\theta)|X_1, X_2, \ldots, X_n\right) \\ =& \mathrm{E}\left[\int_\theta \left(\beta_0 + \frac{\beta_1}{1+\exp(-\gamma^T Z_{n+1})}\right) dP(\theta)|X_1, X_2, \ldots, X_n\right].\end{aligned} \tag{3.4.5}$$

By Corollary 3.4.1, the Bayes predictor could be approximated by

$$\hat{\mu}_{n+1} \cong \frac{\alpha \int g(\theta) dG_0(\theta)}{\alpha + n} + \frac{\sum_{i=1}^{M}\left[\sum_{j=1}^{k^{(N+ri)}} n_j^{(N+ri)} g\left(\phi_j^{(N+ri)}\right)\right]}{M(\alpha+n)}, \tag{3.4.6}$$

where $g(\theta) = \beta_0 + \frac{\beta_1}{1+\exp(-\gamma^T Z_{n+1})}$. Given the prior as in equation (3.3.1), the Bayes predictor can be simplified as

$$\hat{\mu}_{n+1} \cong \frac{\alpha u_0}{\alpha + n} + \frac{\sum_{i=1}^{M}\left[\sum_{j=1}^{k^{(N+ri)}} n_j^{(N+ri)} g\left(\phi_j^{(N+ri)}\right)\right]}{M(\alpha+n)}. \tag{3.4.7}$$

The variance of the Bayes predictor is given by the following equation:

$$\begin{aligned}\mathrm{var}(\hat{\mu}_{n+1}) =& \mathrm{E}\left(\int_\theta \int_x x^2 f(x|Z_{n+1},\theta) dx dP(\theta)|X_1, X_2, \ldots, X_n\right) - \mu_{n+1}^2 \\ =& \mathrm{E}\left[\int_\theta h(\theta) dP(\theta)|X_1, X_2, \ldots, X_n\right] - \mu_{n+1}^2,\end{aligned} \tag{3.4.8}$$

where

$$h(\theta) = \left(\beta_0 + \frac{\beta_1}{1+\exp(-\gamma^T Z_{n+1})}\right)^2 + \sigma^2.$$

By Corollary 3.4.1, the variance of Bayes predictor could be approximated by

$$\hat{\mathrm{var}}(\hat{\mu}_{n+1}) \cong \frac{\alpha \int h(\theta) dG_0(\theta)}{\alpha + n} + \frac{\sum_{i=1}^{M}\left[\sum_{j=1}^{k^{(N+ri)}} n_j^{(N+ri)} h\left(\phi_j^{(N+ri)}\right)\right]}{M(\alpha+n)} - \hat{\mu}_{n+1}^2 \tag{3.4.9}$$

Again, the integration in first term of equation (3.4.10) could be found by direct integration or by numerical integration as in equation (3.4.4). Suppose that the prior is given in equation (3.3.1). Then the variance of the Bayes predictor is approximated by

$$\begin{aligned}\hat{\text{var}}(\hat{\mu}_{n+1}) \cong & \frac{\alpha \int \left(\beta_0 + \frac{\beta_1}{1+\exp(-\gamma^T Z_{n+1})}\right)^2 dG_0(\theta)}{\alpha+n} + \frac{\alpha}{\alpha+n}\frac{b}{a-1} \\ & + \frac{\sum_{i=1}^{M}\left[\sum_{j=1}^{k^{(N+ri)}} n_j^{(N+ri)} h\left(\phi_j^{(N+ri)}\right)\right]}{M(\alpha+n)} - \hat{\mu}_{n+1}^2.\end{aligned} \tag{3.4.10}$$

We note that the variance of the Bayes predictor does not exist when $a \leq 1$.

Chapter 4

Posterior Consistency On Transition Densities

This chapter establishes sufficient conditions under which the posterior will be consistent in our DPM models when the state space is the real line $\mathbb{R}$. Posterior consistency is important in validating the Bayes procedure in that the procedure should be able to find the true mechanism closely, as more and more data come in. In consistent Bayesian procedures, the data will eventually swamp the prior, that is, different priors will ultimately lead to very close predictive distributions (refer to Section 2.2 for more discussion). The definition of posterior consistency depends on the topology on the relevant space. Section 4.1 defines different topologies on the space of transition density functions. As pointed out in Section 2.2, Schwartz's (1965) theorem and its various extensions have been the main tools for studying posterior consistency in nonparametric problem with i.i.d. observations. Section 4.3 extends Schwartz's theorem to Markov models. Schwartz's theorem consists of two critical conditions. One is that the prior puts positive mass in any Kullback-Leibler neighborhood of the true parameter. The other is the existence of an exponentially consistent sequence of tests. For a Markov process, Section 4.4 gives a sufficient condition so that a transition density lies in the Kullback-Leibler support of the prior. Sufficient conditions

for the existence of exponentially consistent tests in weak star topology are given in Section 4.5 and in sup-L_1 metric are given in Section 4.6. Our main results are concluded in Section 4.2.

4.1 Topologies on the Space of Transition Densities

Let $\mathbb{R}$ be the real line with its Borel σ-algebra $\mathcal{B}$. Let $\mathcal{M}(\mathbb{R})$ be the space of all probablity measures on $(\mathbb{R}, \mathcal{B})$. Let $\pounds(\mathbb{R})$ be a proper subset of the space of Markov transition density functions on $(\mathbb{R}, \mathcal{B})$. In our approach, we assume that Doeblin's condition holds for any Markov model in $\pounds(\mathbb{R})$. Thus any Markov model $f(\cdot|x) \in \pounds(R)$ is uniformly ergodic with a unique stationary probability density function π_f. Let $\pounds_\pi(\mathbb{R})$ denote the space of all invariant density functions corresponding to $f(\cdot|x) \in \pounds(R)$. There are many natural topologies on the space $\mathcal{M}(\mathbb{R})$, as reviewed in Subsection 4.1.1. Subsection 4.1.2 discusses how to extend the topologies on $\mathcal{M}(\mathbb{R})$ to the space $\pounds(\mathbb{R})$.

4.1.1 Topologies on the Space of Probability Measures

We first review topologies on $\mathcal{M}(\mathbb{R})$. Among these topologies, the "topology of weak convergence" (the "weak-star topology") is the most frequently used. The weak-star topology on $\mathcal{M}(\mathbb{R})$ is the one for which the basic neighborhoods of P are the sets of the form

$$\left\{ Q : \left| \int \psi_i dQ - \int \psi_i dP \right| \leq \epsilon, i = 1, \ldots, k \right\}, \tag{4.1.1}$$

where $\epsilon > 0$ and $\psi_1, \ldots, \psi_k$ are bounded continuous functions on $\mathbb{R}$. Thus $\{P_n\}$ converges weakly to P if and only if $\{P_n\}$ converges to P in the weak-star topology.

Following Portmanteau theorem C.0.1, three other bases for the weak convergence topology could be defined as follows.

(a) the sets $\{ Q : Q(F_i) < P(F_i) + \epsilon,\ i = 1, \ldots, k \}$ with F_i closed.

(b) the sets $\{ Q : Q(G_i) > P(G_i) - \epsilon,\ i = 1, \ldots, k \}$ with G_i open.

(c) the sets $\{ Q : |Q(A_i) - P(A_i)| < \epsilon,\ i = 1, \ldots, k \}$ with $P(\partial A_i) = 0$, where $\partial A = \overline{A} - A^o$ is the boundary of A.

Theorem 4.1.1 (Billingley (1968), Page 236). *The bases (a), (b), (c) and the basis given in equation* (4.1.1) *are equivalent and all generate the same weak star topology. The equivalence means that given any neighborhood O_1 in one basis, we could find the neighborhood O_2 in any three other bases satisfying $O_2 \subset O_1$.*

Since $\mathbb{R}$ is separable, $\mathcal{M}(\mathbb{R})$ is complete and seperable under the topology of weak convergence (Billingley 1999). The σ-algebra generated by the weak star topology is just the smallest one that makes all the function $\{P \mapsto P(B) : B \in \mathcal{B}\}$ measurable. The weak star topology has the fewest open sets and weak neighborhoods are large. The other two frequently used metrics are the "total variation metric" and the "Hellinger metric". They are extremely useful if we are interested in the subspace L_μ, all of probability measures dominated by the σ-finite measure μ. Let P and Q be two probability measures on the space $(\mathbb{R}, \mathcal{B})$. The "total variation distance" between P and Q is defined as

$$d_{TV}(P, Q) = \|P - Q\| = 2 \sup_{A \in \mathcal{B}(\chi)} |P(A) - Q(A)|.$$

The "Hellinger distance" between P, Q is

$$H(P, Q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2 d\mu}$$

where p, q are the densities of P and Q respectively with respect to μ. Associated with the Hellinger metric, there is a useful quantity $A(P, Q)$ called *Affinity* defined

by,

$$A(P,Q) = \int \sqrt{p}\sqrt{q}d\mu = 1 - \frac{H^2(P,Q)}{2}.$$

The "Kullback Leibler divergence" between two probability densities p, q (with respect to μ) is defined as

$$K(P,Q) = K(p,q) = \mathrm{E}_P \log \frac{p}{q} = \int p \log \frac{p}{q} d\mu.$$

Though not a metric, it has played a central role in the classical theory of estimation and testing and the consistency of Bayesian estimation.

Theorem 4.1.2.

$$\|P - Q\| = \sup_{|f|\le 1} \left| \int f\, dP - \int f\, dQ \right| = 2 \sup_{0\le f\le 1} \left| \int f\, dP - \int f\, dQ \right|.$$

Let p, q be the densities of P, Q respectively with respect to some measure μ. For convenience, we may denote the total variation distance of P and Q as $\|p-q\|$. Then

$$\begin{aligned}\|p - q\| &= \int |p - q|\, d\mu = 2 \int_{p(x)\ge q(x)} (p(x) - q(x)) d\mu(x) = 2 \int_{q(x)\ge p(x)} (q(x) - p(x)) d\mu(x) \\ &= \sup_{|f|\le 1} \left| \int fp\, d\mu - \int fq\, d\mu \right| = 2 \sup_{0\le f\le 1} \left| \int fp\, d\mu - \int fq\, d\mu \right|.\end{aligned}$$

Theorem 4.1.3. *(i)* $\|P - Q\|^2 \le 2H^2(P,Q)(1 + A(P,Q)) \le 4H^2(P,Q)$.
(ii) $H^2(P,Q) \le \|P - Q\|$.
(iii) $A(P,Q) = \int \sqrt{pq} d\mu = \mathrm{E}_P \sqrt{\frac{q}{p}} = \mathrm{E}_Q \sqrt{\frac{p}{q}} \le \sqrt{1 - \frac{\|P-Q\|^2}{4}}$.
(iv) $K(P,Q) \ge \frac{\|P-Q\|^2}{4}$.

From Theorem 4.1.3, $\|P_n - Q_n\| \to 0$ if and only if $H(P_n, Q_n) \to 0$. So the total variation metric is equivalent to the Hellinger metric. If $\{P_n\}$ converges to P under either of the two metrics, it means $P_n(B)$ converges to $P(B)$ uniformly for any $B \in \mathcal{B}$. While if $\{P_n\}$ converges to P weakly, it only means that $P_n(B)$ converges to $P(B)$ if $P(\partial B) = 0$. Therefore convergence under total variation metric or the Hellinger metric is much stronger than the weak convergence.

4.1.2 Topologies on the Space of Transition Densities

There are many topologies on $\mathcal{M}(\mathbb{R})$. However, there is no natural one on the space $\pounds(\mathbb{R})$ due to the dependent structure of Markov processes. The metrics defined on the space $\mathcal{M}(\mathbb{R})$ cannot be used directly on $\pounds(\mathbb{R})$. For each $x \in \mathbb{R}$, a basic open set for the weak star topology of $f_0(y|x)$ is given by

$$\left\{ f : \int \psi_1(y) f(y|x) dy < \int \psi_1(y) f_0(y|x) dy + \varepsilon, \ldots, \right.$$
$$\left. \int \psi_k(y) f(y|x) dy < \int \psi_k(y) f_0(y|x) dy + \varepsilon \right\}.$$

where $\psi_1, \ldots, \psi_k$ are bounded continuous functions on $\mathbb{R}$. This depends on x which must be integrated out. It is natural to integrate out x with respect to the corresponding invariant distributions, so that we consider weak neighborhoods of the type

$$\left\{ f : \iint \psi_1(y) f(y|x) dy \, \pi_f(x) dx < \iint \psi_1(y) f_0(y|x) dy \, \pi_0(x) dx + \varepsilon, \right.$$
$$\left. \ldots, \iint \psi_k(y) f(y|x) dy \, \pi_f(x) dx < \iint \psi_k(y) f_0(y|x) dy \, \pi_0(x) dx + \varepsilon \right\}.$$

Since $\int f(y|x)\pi_f(x)dx = \pi_f(y)$, the basic open set for the weak topology reduces to a weak neighborhood of the invariant measure:

$$\left\{ f : \int \psi_1(y) \, \pi(y) dy < \int \psi_1(y) \, \pi_0(y) dy + \varepsilon, \ldots, \right.$$
$$\left. \int \psi_k(y) \, \pi(y) dy < \int \psi_k(y) \, \pi_0(y) dy + \varepsilon \right\}. \quad (4.1.2)$$

This is just the weak neighborhood of π_0 in the space $\pounds_\pi(R)$.

Definition 4.1.1. The *weak neighborhood* of a transition density f_0 is defined in equation (4.1.2), that is, just the weak neighborhood of its invariant probability measures of π_0 in the space $\pounds_\pi(R)$.

The weak-star neighborhood defined above is too large. Another issue with respect to the weak-star neighborhood is the possible non-identification of the transition densities in a neighborhood since different transition densities may have the same invariant probability measures. The sup-L_1 metric defined below is a very strong neighborhood and avoids the non-identity issue.

Definition 4.1.2. The sup-L_1 *neighborhood* of a transition density f_0 is defined as

$$\left\{ f : \sup_x \|f(y|x) - f_0(y|x)\| < \varepsilon \right\}. \tag{4.1.3}$$

In the literature, there are some other topologies on the space of transition densities. One example would be the Birge's metric (1978). Let ν be a probability measure on the space of $\mathcal{M}(\mathbb{R})$. Birge's distance between two transition densities is

$$d_\nu(f_0, f) = \frac{1}{2} \iint \left(\sqrt{f_0(y|x)} - \sqrt{f(y|x)} \right)^2 dy\, \nu(dx). \tag{4.1.4}$$

Since Birge's metric is much weaker than the sup-L_1 metric, posterior consistency in sup-L_1 metric implies the consistency in Birge's metric. In this thesis, we shall not discuss too much about the Birge's metric.

Similarly, we could define the Kullback-Leibler divergence for the Markov processes as follows.

Definition 4.1.3. In $\pounds(R)$, the *Kullback-Leibler divergence* between two transition densities f, f_0 is

$$K(f_0, f) = \mathrm{E}_{\pi_0 \otimes f_0} \left[\log \frac{f_0(y|x)}{f(y|x)} \right] = \int \pi_0(x) f_0(y|x) \log \frac{f_0(y|x)}{f(y|x)} dy dx. \tag{4.1.5}$$

The true transition density f_0 is said to be in the Kullback-Leibler (K-L) support of the prior Π, if for any $\epsilon > 0$, $\Pi(K_\epsilon(f_0)) > 0$, where $K_\epsilon(f_0)$ stands for

$$K_\epsilon(f_0) = \{ g : K(f_0, g) < \epsilon \}. \tag{4.1.6}$$

Suppose that P_f^μ denote the overall law of the Markov process with transition density f and initial distribution μ of X_0. In nonparametric Bayesian procedure, Π is a prior on the space $\pounds(\mathbb{R})$ and Π_n denote the posterior distribution when the sample size is n.

Definition 4.1.4. The sequence of posterior distributions $\{\Pi_n,\, n \geq 1\}$ is said to be *consistent at* f_0 under some topology if for every neighborhood U of f_0, $\Pi_n(U|X_1, \cdots, X_n) \to 1$ almost surely the law defined by f_0, i.e., $P_{f_0}^\mu$.

Following the Portmanteau theorem C.0.1, the consistency of the posterior at f_0 is equivalent to requiring that the posterior $\{\Pi_n\}$ converges weakly to δ_{f_0} almost surely in the corresponding topology.

Definition 4.1.5. The sequence of posterior distributions $\{\Pi_n,\, n \geq 1\}$ is said to be *weakly consistent at* f_0 if for every weak neighborhood U of f_0 given in equation (4.1.2), $\Pi_n(U|X_1, \cdots, X_n) \to 1$ almost surely $P_{f_0}^\mu$. The sequence of posterior distributions $\{\Pi_n,\, n \geq 1\}$ is said to be *strongly consistent at* f_0 *under* sup-L_1 *metric* if for every neighborhood U of f_0 given in equation (4.1.3), $\Pi_n(U|X_1, \cdots, X_n) \to 1$ almost surely $P_{f_0}^\mu$.

4.2 Main results

In this chapter, we establish the posterior consistency under weak-star topology for models one and three, and the posterior consistency under the sup-L_1 metric for models two and four. Models one to four are defined as follows:

Model One:

$$X_i = \theta_{1i} + F(X_{i-1}, \theta_{2i}) + \varepsilon_i, \tag{4.2.1}$$

where F could be any continuous function bounded by a value, say a, θ_{2i} may be a vector, ε_i's are i.i.d. from the normal distribution with mean 0 and variance σ_i^2, and $\theta_i = (\theta_{1i}, \theta_{2i}^T, \sigma_i)^T$ are i.i.d. according to the law P. We assume that P has a Dirichlet process prior $D_{\alpha G_0}$, where the center measure G_0 satisfies

$$G_0(\sigma \geq \underline{\sigma}) = 1.$$

for a fixed positive $\underline{\sigma}$.

Model Two:

Given the parameter P, X_i's follows a Markov process with transition density

$$f_P(y|x) = \int \phi_\sigma \left(y - v - \varphi \frac{1}{1 + \exp[-\gamma(x-u)]} \right) dP(v, \varphi, \gamma, u, \sigma).$$

We assume that P has a Dirichlet process prior $D_{\alpha G_0}$, where G_0 are compactly supported, and

$$G_0(\gamma \in [\underline{\gamma}, \overline{\gamma}]) = 1,$$

where $\underline{\gamma}$ and $\overline{\gamma} > 0$ have the same signs.

Model Three:

$$X_i = \theta_{1i} + F(X_{i-1}, \theta_{2i}) + \varepsilon_i, \tag{4.2.2}$$

where F could be any continuous function bounded by a value, say a, θ_{2i} may be a vector, ε_i are i.i.d. from the normal distribution with mean 0 and variance σ^2, and $\theta_i = (\theta_{1i}, \theta_{2i}^T)^T$ are i.i.d. according to the law P. We assume that P has a Dirichlet process prior $D_{\alpha G_0}$, and σ has a prior μ satisfying

$$\mu(\sigma > \underline{\sigma}) = 1, \text{ where } \underline{\sigma} > 0 \text{ is fixed.}$$

Model Four:

Given the parameter P, X_i's follows a Markov process with transition density

$$f_P(y|x) = \int \phi_\sigma \left(y - v - \varphi \frac{1}{1 + \exp[-\gamma(x-u)]} \right) dP(v, \varphi, \gamma, u, \sigma).$$

We assume that P has a Dirichlet process prior $D_{\alpha G_0}$ and σ has a prior μ, where both G_0 and μ are compactly supported, and

$$G_0(\sigma \geq \underline{\sigma}) = 1.$$

for a fixed positive $\underline{\sigma}$.

Models three and four are less flexible than Model one and two, as mentioned at the end of Section 1.2, since in Models three and four, σ is a static parameter with a prior μ. In fact, Model three and four are special cases of Model one and two respectively. So the posterior consistency of Model three and four hold automatically if we can show the posterior consistency of Models one and two. All arguments for establishing posterior consistency are based on the Schwartz-type Theorem in Section 4.3. Details are given in Section 4.4, 4.5 and 4.6.

4.3 General Theories on Posterior Consistency

Schwartz's (1965) theorem and its various extensions have been the main tools for studying posterior consistency in nonparametric problem with i.i.d. observations. In this section, we extend Schwartz's theorem to Markov models. Schwartz's theorem consists of two critical conditions. One is that the true transition lies in the K-L (Kullback-Leibler) support of the prior, the other is the existence of an exponentially consistent sequence of tests. Kullback-Leibler support is defined in Section 4.1 while the definition of exponentially consistent tests is given as follows. Our two fundamental Schwartz-type results are given in Theorem 4.3.1 and Corollary 4.3.1. The next three sections establish sufficient conditions so that the two assumptions of Schwartz's theorem hold.

Definition 4.3.1. A sequence of test functions $\{\phi(X_0, X_1, \ldots, X_n) : n \geq 1\}$ is uniformly exponentially consistent for testing $H_0 : f = f_0$ *vs* $H_1 : f \in V_n^c$, if there exists C, β, n_0 such that for all $n > n_0$, the type I and type II are bounded by $Ce^{-n\beta}$, that is,

$$\begin{aligned} \mathrm{E}_{f_0}(\phi(X_0, X_1, \ldots, X_n)) &\leq Ce^{-n\beta} \\ \sup_{f \in V_n^c} \mathrm{E}_f(1 - \phi(X_0, X_1, \ldots, X_n)) &\leq Ce^{-n\beta} \end{aligned} \tag{4.3.1}$$

Theorem 4.3.1 (Extension Schwartz's theorem* for ergodic Markov processes). *Let Π be a prior and V_n and f_0 satisfy* †

(i) f_0 is in the K-L support of Π,

(ii) there exists a sequence of test functions which is uniformly exponentially consistent for testing $H_0 : f = f_0$ vs $H_1 : f \in V_n^c$.

Then $\Pi(V_n | X_0, X_1, \ldots, X_n) \to 1$ a.s.

Proof. (i) Since f_0 is in the K-L support of Π, by Lemma 4.3.1 we have for any $\beta > 0$,

$$\liminf_{n\to\infty} e^{n\beta} \int \prod_{i=1}^{n} \frac{f(X_i|X_{i-1})}{f_0(X_i|X_{i-1})} \Pi(df) = \infty \text{ a.s. } P_{f_0}^{\infty}. \tag{4.3.2}$$

(ii) There exists a sequence of test functions which is uniformly exponentially consistent for testing $H_0 : f = f_0$ *vs* $H_1 : f \in V_n^c$. By Lemma 4.3.2, for some $\beta_0 > 0$,

$$\liminf_{n\to\infty} e^{n\beta_0} \int_{V_n^c} \prod_{i=1}^{n} \frac{f(X_i|X_{i-1})}{f_0(X_i|X_{i-1})} \Pi(df) = 0 \text{ a.s. } P_{f_0}^{\infty}. \tag{4.3.3}$$

By taking $\beta = \beta_0$ in equation (4.3.2), it easily follows that the ratio

$$\begin{aligned} \Pi(V_n^c | X_0, X_1, \ldots, X_n) &= \frac{\int_{V_n^c} \prod_{i=1}^{n} f(X_i|X_{i-1}) \Pi(df)}{\int \prod_{i=1}^{n} f(X_i|X_{i-1}) \Pi(df)} \\ &= \frac{e^{n\beta_0} \int_{V_n^c} \prod_{i=1}^{n} \frac{f(X_i|X_{i-1})}{f_0(X_i|X_{i-1})} \Pi(df)}{e^{n\beta_0} \int \prod_{i=1}^{n} \frac{f(X_i|X_{i-1})}{f_0(X_i|X_{i-1})} \Pi(df)} \to 0 \text{ a.s. } P_{f_0}^{\infty}. \end{aligned}$$

*Refer to Section 2.2 for original Schwartz's theorem

†In practice, we may take $V_n = V$ for all n.

by equation (4.3.2) and (4.3.3). These arguments leads to the proof of our fundamental consistency theorem. □

Corollary 4.3.1. *Let Π be a prior and V and f_0 satisfy*

(i) f_0 is in the K-L support of Π,

(ii) For each $i = 1, \ldots, k$, there exists a sequence of test functions which is uniformly exponentially consistent for testing $H_0 : f = f_0$ vs $H_1 : f \in U_i$, where $V^c = \bigcup_{i=1}^{k} U_i$ where k is a finite integer and U_i's may overlap with each other. .

Then $\Pi(V|X_0, X_1, \ldots, X_n) \to 1$ a.s. $P_{f_0}^{\infty}$.

Proof. By Theorem 4.3.1, for any $i = 1, \ldots, k$,

$$\Pi(V_i|X_0, X_1, \ldots, X_n) \to 0 \text{ a.s. } P_{f_0}^{\infty}.$$

We have $0 \leq \Pi(V^c|X_0, X_1, \ldots, X_n) \leq \sum_{i=1}^{k} \Pi(V_i|X_0, X_1, \ldots, X_n) \to 0$ a.s. $P_{f_0}^{\infty}$.

So $\Pi(V|X_0, X_1, \ldots, X_n) \to 1$ a.s. $P_{f_0}^{\infty}$. □

Lemma 4.3.1. *If $f_0(y|x)$ is in the Kullback-Leibler support of Π, then for any $\beta > 0$,*

$$\liminf_{n\to\infty} e^{n\beta} \int \prod_{i=1}^{n} \frac{f(X_i|X_{i-1})}{f_0(X_i|X_{i-1})} \Pi(df) = \infty \text{ a.s. } P_{f_0}^{\infty}.$$

Proof. Let $\epsilon = \beta/2$ and note that

$$\int_{\pounds_\mu} \prod_{i=1}^{n} \frac{f(X_i|X_{i-1})}{f_0(X_i|X_{i-1})} \Pi(df) \geq \int_{K_\epsilon(f_0)} e^{-\sum_{i=1}^{n} \log \frac{f_0(X_i|X_{i-1})}{f(X_i|X_{i-1})}} \Pi(df)$$

Since f_0 defines an Ergodic Markov chain which is Harris positive and aperiodic, by Theorem A.1.2,

$$-\frac{1}{n} \sum_{i=1}^{n} \log \frac{f_0(X_i|X_{i-1})}{f(X_i|X_{i-1})} \to -K(f_0, f) > -\epsilon \text{ a.s. } P_{f_0}^{\infty}$$

for each $f \in K_\epsilon(f_0)$. Equivalently for each $f \in K_\epsilon(f_0)$,

$$\exp\left\{ n \left[2\epsilon - \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_0(x_i|x_{i-1})}{f(x_i|x_{i-1})} \right] \right\} \to \infty \text{ a.s. } P_{f_0}^{\infty}. \tag{4.3.4}$$

Hence by Fubini's theorem there is $\Omega_0 \subset \Omega$ of $P_{f_0}^{\infty}$ measure 1 such that, for each $\omega \in \Omega_0$, for all $f \in K_\epsilon(f_0)$, outside a set of Π measure 0, equation (4.3.4) holds. By Fatou's lemma,

$$\begin{aligned}
&\liminf_{n\to\infty} e^{2n\epsilon} \int_{L_\mu} \prod_{i=1}^{n} \frac{f(X_i|X_{i-1})}{f_0(X_i|X_{i-1})} \Pi(df) \\
&\geq \liminf_{n\to\infty} e^{2n\epsilon} \int_{K_\epsilon(f_0)} e^{-\sum_{i=1}^{n} \log \frac{f_0(X_i|X_{i-1})}{f(X_i|X_{i-1})}} \Pi(df) \\
&\geq \int_{K_\epsilon(f_0)} \liminf_{n\to\infty} e^{n\left[2\epsilon - \frac{1}{n}\sum_{i=1}^{n} \log \frac{f_0(X_i|X_{i-1})}{f(X_i|X_{i-1})}\right]} \Pi(df) \to \infty.
\end{aligned}$$

a.s. $P_{f_0}^{\infty}$. □

Proposition 4.3.1. *Let v be any probability measure on U_n. Suppose that there is a nonnegative test function $\phi_n(X_0, X_1, \cdots, X_n)$ bounded by 1 such that*

$$\mathrm{E}_{f_0}(\phi_n(X_0, X_1, \ldots, X_n)) \leq Ce^{-n\beta},$$

$$\sup_{f\in U_n} \mathrm{E}_f(1 - \phi_n(X_0, X_1, \ldots, X_n)) \leq Ce^{-n\beta}.$$

Let $p_n(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_0(x_i|x_{i-1})$ and $q_n(x_1, x_2, \ldots, x_n) = \int_{U_n} \prod_{i=1}^{n} f(x_i|x_{i-1}) v(df)$, then

$$||p_n - q_n|| \geq 2(1 - 2Ce^{-n\beta}).$$

Proof. By Fubini's theorem,

$$\begin{aligned}
&\int_{x_1} \cdots \int_{x_n} \phi_n(x_0, \cdots, x_n) q_n(x_1, x_2, \ldots, x_n) dx_1 \cdots dx_n \\
&= \int_{x_1} \cdots \int_{x_n} \int_{U_n} \phi_n(x_0, \cdots, x_n) \prod_{i=1}^{n} f(x_i|x_{i-1}) \Pi(df) dx_1 \cdots dx_n \\
&= \int_{U_n} E_f\left[\phi_n(X_0, X_1, \cdots, X_n)\right] \Pi(df) \\
&\geq 1 - Ce^{-n\beta}
\end{aligned} \tag{4.3.5}$$

and by the assumption,

$$\int_{x_1} \cdots \int_{x_n} \phi_n(x_0, \cdots, x_n) p_n(x_1, x_2, \ldots, x_n) dx_1 \cdots dx_n = \mathrm{E}_{f_0}(\phi_n) \leq Ce^{-n\beta}. \tag{4.3.6}$$

Hence, by Theorem 4.1.2

$$\begin{aligned}
&||p_n - q_n|| \\
&= 2 \sup_{0 \le h(x_1,\ldots,x_n) \le 1} \Big| \int_{x_1} \cdots \int_{x_n} h(x_1, \ldots, x_n) \\
&\qquad\qquad [p_n(x_1, \ldots, x_n) - q_n(x_1, \ldots, x_n)] dx_1 \cdots dx_n | \\
&\ge 2 \left(\int_{x_1} \cdots \int_{x_n} \phi_n(x_1, \ldots, x_n) p_n(x_1, \ldots, x_n) dx_1 \cdots dx_n - \right. \\
&\qquad\qquad \left. \int_{x_1} \cdots \int_{x_n} \phi_n(x_1, \ldots, x_n) q_n(x_1, \ldots, x_n) dx_1 \cdots dx_n \right) \\
&\ge 2(1 - 2Ce^{-n\beta})
\end{aligned}$$

by equation (4.3.5) and (4.3.6) □

Lemma 4.3.2. *If there exists a sequence of test functions which is uniformly exponentially consistent for testing* $H_0 : f = f_0$ *vs* $H_1 : f \in U_n^c$, and $\Pi(U_n^c) \ge a > 0$ for all $n > n_0$, then for some $\beta_0 > 0$,

$$\lim_{n\to\infty} e^{n\beta_0} \int_{U_n^c} \prod_{i=1}^{n} \frac{f(X_i|X_{i-1})}{f_0(X_i|X_{i-1})} \Pi(df) = 0 \text{ a.s. } P_{f_0}^{\infty}.$$

Proof. Obviously $v(df) = \Pi(df)/\Pi(U_n^c)$ defines a probability measure on the set U_n^c. Set $q_n(x_0, \ldots, x_n) = \frac{1}{\Pi(U_n^c)} \int_{U_n^c} \prod_{i=1}^{n} f(x_i|x_{i-1}) \Pi(df)$ and $p_n(x_0, \ldots, x_n) = \prod_{i=1}^{n} f_0(x_i|x_{i-1})$. Since there exists a sequence of test functions which is uniformly exponentially consistent for testing $H_0 : f = f_0$ *vs* $H_1 : f \in U_n^c$, by Proposition 4.3.1,

$$||q_n - p_n|| \ge 2(1 - 2Ce^{-n\beta}) \tag{4.3.7}$$

Let

$$A(p_n, q_n) = \int \cdots \int \sqrt{\prod_{i=1}^{n} f_0(x_i|x_{i-1})} \sqrt{q_n(x_0, x_1, \cdots, x_n)} dx_1 \cdots dx_n.$$

Then by Theorem 4.1.3 and equation (4.3.7),

$$A(p_n, q_n) \le \sqrt{1 - \frac{||q_n - p_n||^2}{4}} \le \sqrt{1 - \frac{[2(1 - 2Ce^{-n\beta})]^2}{4}} \le \sqrt{4C} e^{-\frac{n\beta}{2}}$$

provided that $Ce^{-n\beta} \leq 1$. Thus by Markov's inequality,

$$P_{f_0}\left\{\sqrt{\frac{q_n(X_0, X_1, \ldots, X_n)}{\prod_{i=1}^n f_0(X_i|X_{i-1})}} \geq e^{\frac{-n\beta}{4}}\right\} \leq e^{\frac{n\beta}{4}}\sqrt{4C}e^{-\frac{n\beta}{2}} = \sqrt{4C}e^{-\frac{n\beta}{4}}$$

An application of Borel-Cantelli lemma yields that

$$\sqrt{\frac{q_n(X_0, X_1, \ldots, X_n)}{\prod_{i=1}^n f_0(X_i|X_{i-1})}} \leq e^{-\frac{n\beta}{4}} \quad \text{a.s.} \quad P_{f_0}^{\infty}.$$

Hence when $\beta_0 < \beta/4$,

$$e^{\frac{n\beta}{4}}\int_{U_n^c}\prod_{i=1}^n \frac{f(x_i|x_{i-1})}{f_0(x_i|x_{i-1})}\Pi(df) \leq e^{-n(\frac{\beta}{4}-\beta_0)} \to 0 \text{ a.s. } P_{f_0}^{\infty}.$$

□

4.4 Kullback-Leibler Support of f_0

In Section 4.1, we introduced the concept of the K-L support of Π, which is one of the key sufficient conditions for posterior consistency. For any transition density $f_0(y|x)$, we denote by $K_\epsilon(f_0)$ the Kullback- Leibler neighborhood

$$\left\{f : \iint \pi_0(x)f_0(y|x)\log\{\frac{f_0(y|x)}{f(y|x)}\}dydx < \epsilon\right\}.$$

Say that f_0 is in the *Kullback-Leibler support* of Π if for any $\epsilon > 0$,

$$\Pi(K_\epsilon(f_0)) > 0.$$

Our main result on Kullback-Leibler support for Markov processes is given in Theorem 4.4.1.

Theorem 4.4.1. *Given model one* [‡] *introduced in Section 4.2, we assume that the true model has the form*

$$f_0(y|x) = f_{P_0}(y|x) = \int \phi_\sigma(y - \theta_1 - F(x, \theta_2))dP_0(\theta_1, \theta_2, \sigma).$$

[‡]Note that Model two is a special case of Model one

Suppose that the support of P_0 belongs to the support of G_0 (that is, P_0 is in the weak-star support of $D_{\alpha G_0}$). If $\mathrm{E}_{P_0}(\theta_1^2) < \infty$ and $\mathrm{E}_{P_0}(\sigma^2) < \infty$, then f_0 is in the K-L support of the prior.

The proof of Theorem 4.4.1 will be given after Proposition 4.4.2..

Lemma 4.4.1. *Suppose that the assumptions of Model one hold. If $f = f_P$ is any Markov transition density where the support of P belongs to the support of G_0 (that is, P is in the weak-star support of $D_{\alpha G_0}$), then it is uniformly ergodic and*

$$\sup_{x\in\mathbb{R}} \|f_P^n(\cdot|x) - \pi\| \leq \left[\int_{|y|\leq a/\underline{\sigma}} \phi(y)dy\right]^n,$$

where π is the invariant probability measure of f, and $\phi(y)$ is the density of standard normal distribution.

Proof. Observe that

$$\begin{aligned} f_P(y|x) &= \int \phi_\sigma\left(y - \theta_1 - F(x,\theta_2)\right) dP(\theta_1,\theta_2,\sigma) \\ &\geq g(y) = \int \phi_\sigma(|y-\theta_1| + a) dP(\theta_1,\sigma) \end{aligned}$$

where a bounds the function F, and that

$$\begin{aligned} 1 > c &= \int g(y)dy \\ &= \iint \phi_\sigma(|y-\theta_1|+a)dP(\theta_1,\sigma)\,dy \\ &= \iint \phi_\sigma(|y-\theta_1|+a)dy\,dP(\theta_1,\sigma) \\ &= \int\int_{y>\theta_1} \phi_\sigma(y-\theta_1+a)dy\,dP(\theta_1,\sigma) + \int\int_{y<\theta_1} \phi_\sigma(y-\theta_1-a)dy\,dP(\theta_1,\sigma) \\ &= \int\int_{y>a} \phi_\sigma(y)dy\,dP(\theta_1,\sigma) + \int\int_{y<-a} \phi_\sigma(y)dy\,dP(\theta_1,\sigma) \\ &= \int\int_{|y|>a/\sigma} \phi(y)dy\,dP(\theta_1,\sigma) \\ &\geq \int_{|y|>a/\underline{\sigma}} \phi(y)dy \end{aligned}$$

Thus Doeblin's condition holds since the transition density satisfies $f(y|x) \geq cg(y)/c$, where $g(y)/c$ is a density on the real line. The lemma follows from the Theorem A.1.7. □

Proposition 4.4.1. *Suppose that the true transition densition has the form*

$$f_0(y|x) = f_{P_0}(y|x) = \int \phi_\sigma(y - \theta_1 - F(x, \theta_2)) dP_0(\theta_1, \theta_2, \sigma).$$

Then

(i) the invariant probability measure, say π_0, uniquely exists,

(ii) if $\mathrm{E}_{P_0}(\theta_1^2) < \infty$ and $\mathrm{E}_{P_0}(\sigma^2) < \infty$, then $\mathrm{E}_{\pi_0}(x^2) < \infty$.

Proof. (i) by Lemma 4.4.1, the chain is uniformly ergodic. The invariant distribution π_0 uniquely exists.

(ii) Note that

$$y^2 = (\theta_1 + F(x, \theta_2) + \varepsilon)^2 \leq 3(\theta_1^2 + a^2 + \varepsilon^2),$$

where $\varepsilon \sim N(0, \sigma^2)$ and $(\theta_1, \theta_2, \sigma) \sim P_0$. So

$$\mathrm{E}_{f_0}(y^2|x) \leq U = 3\mathrm{E}_{P_0}\left[\theta_1^2 + a^2 + \mathrm{E}(\varepsilon^2)\right] = 3\left[\mathrm{E}_{P_0}(\theta_1^2) + a^2 + \mathrm{E}_{P_0}(\sigma^2)\right].$$

for any x. So

$$\mathrm{E}_{\pi_0}(x^2) = \mathrm{E}_{\pi_0}\{\mathrm{E}_{f_0}(y^2|x)\} \leq U < \infty.$$

□

Proposition 4.4.2. *If $P_2(|\theta_1| \leq k, \sigma \in [\sigma_1, \sigma_2]) \geq b$, where $b > 0$ and $\sigma_1 > 0$, then*

$$\left|\log \frac{f_{P_1}(y|x)}{f_{P_2}(y|x)}\right| \leq \log \frac{\sigma_2}{b\underline{\sigma}} + \frac{3(y^2 + k^2 + a^2)}{2\sigma_1^2}$$

Proof. If $P_2(|\theta_1| \le k, \sigma \in [\sigma_1, \sigma_2]) \ge b$, then we have

$$\begin{aligned}
f_{P_2}(y|x) &= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\theta_1-F(x,\theta_2))^2}{2\sigma^2}\right] dP_2(\theta_1,\theta_2,\sigma) \\
&\ge \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{3(y^2+\theta_1^2+F(x,\theta_2)^2)}{2\sigma^2}\right] dP_2(\theta_1,\theta_2,\sigma) \\
&\ge \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{3(y^2+\theta_1^2+a^2)}{2\sigma^2}\right] dP_2(\theta_1,\theta_2,\sigma) \\
&\ge \int_{|\theta_1|\le k, \sigma\in[\sigma_1,\sigma_2]} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{3(y^2+\theta_1^2+a^2)}{2\sigma^2}\right] dP_2(\theta_1,\theta_2,\sigma) \\
&\ge \int_{|\theta_1|\le k, \sigma\in[\sigma_1,\sigma_2]} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{3(y^2+k^2+a^2)}{2\sigma_1^2}\right] dP_2(\theta_1,\theta_2,\sigma) \\
&\ge b\frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{3(y^2+k^2+a^2)}{2\sigma_1^2}\right]
\end{aligned}$$

and that

$$f_{P_1}(y|x) \le \frac{1}{\sqrt{2\pi}\underline{\sigma}}.$$

We get the result by dividing the above two inequalities. □

Proof of Theorem 4.4.1. The Theorem is equivalent to that for any $\epsilon > 0$, we could find a weak-star neighborhood N_p of P_0 such that $\mathit{\Pi}(N_p) > 0$, and for any $P \in N_{p\sigma}$,

$$\mathrm{E}_{\pi_0\otimes f_0}\left[\log \frac{f_{\sigma,P_0}(y|x)}{f_{\sigma,P}(y|x)}\right] \le \epsilon. \tag{4.4.1}$$

Note that P_0 is a given probability measure on the space of $(\theta_1, \theta_2, \sigma)$. There exists constants k_1, σ_1, σ_2, and a compact set A_{θ_2} in the space of θ_2 such that

$$P_0(\theta_1 \in [-k_1, k_1], \sigma \in [\sigma_1, \sigma_2]) \ge P_0(\theta_1 \in [-k_1, k_1], \sigma \in [\sigma_1, \sigma_2], \theta_2 \in A_{\theta_2}) \ge \frac{1}{2}.$$

Since $\mathrm{E}_{P_0}(\theta_1^2) < \infty$ and $\mathrm{E}_{P_0}(\sigma^2) < \infty$, by Proposition 4.4.1, $\log \frac{2\sigma_2}{\underline{\sigma}} + \frac{3(Y^2+k_1^2+a^2)}{2\sigma_1^2}$ is integrable under the product probability measure $\pi_0 \otimes f_0$. So there exists K_1 and K_2 such that

$$\mathrm{E}_{\pi_0\otimes f_0}\left[\left(\log\frac{2\sigma_2}{\underline{\sigma}} + \frac{3(Y^2+k_1^2+a^2)}{2\sigma_1^2}\right) I(|Y| > K_2)\right] < \frac{3\epsilon}{8},$$

$$\mathrm{E}_{\pi_0\otimes f_0}\left[\left(\log\frac{2\sigma_2}{\underline{\sigma}}+\frac{3(Y^2+k_1^2+a^2)}{2\sigma^2}\right)I(|X|>K_1)\right]<\frac{3\epsilon}{8}.$$

If we restrict N_p be the subset of $P(|\theta_1|\le k_1,\sigma\in[\sigma_1,\sigma_2])\ge 1/2$, then by Proposition 4.4.2, for any $P\in N_p$,

$$\left|\log\frac{f_{P_0}(y|x)}{f_P(y|x)}\right|\le\log\frac{2\sigma_2}{\underline{\sigma}}+\frac{3(y^2+k_1^2+a^2)}{2\sigma_1^2}.$$

So we have

$$\begin{aligned}E_{\pi_0\otimes f_0}\left[\left|\log\frac{f_{P_0}(Y|X)}{f_P(Y|X)}\right|I(|Y|>K_2)\right]&<\frac{3\epsilon}{8},\\ E_{\pi_0\otimes f_0}\left[\left|\log\frac{f_{P_0}(Y|X)}{f_P(Y|X)}\right|I(|X|>K_1)\right]&<\frac{3\epsilon}{8}.\end{aligned}\tag{4.4.2}$$

Now we try to construct N_p such that for any $P\in N_p$,

$$\mathrm{E}_{\pi_0\otimes f_0}\left\{\log\left(\frac{f_{P_0}(Y|X)}{f_P(Y|X)}\right)I(|X|\le K_1,|Y|\le K_2)\right\}<\frac{\epsilon}{4}.$$

Then the Theorem follows equation (4.4.6)

Let

$$d=\inf_{|x|<K_1,|y|<K_2}\left[\frac{\varepsilon}{8}\int_{\theta_1\in[-k_1,k_1],\,\theta_2\in A_{\theta_2},\,\sigma\in[\sigma_1,\sigma_2]}\phi_\sigma[y-\theta_1-F(x,\theta_2)]dP_0(\theta_1,\theta_2)\right].$$

Then $d>0$ since for any $(x,\,y,\theta_1,\theta_2,\sigma)$ in a compact set, $\phi_\sigma[y-\theta_1-F(x,\theta_2)]$ is bounded below.

We could find the compact set B and B_L on the space $(\theta_1,\theta_2,\sigma)$ and a continuous function $t(\theta_1,\theta_2,\sigma)$ such that

(a) $B^*\subset B\subset B_L$ where

$$B^*=\{(\theta_1,\theta_2,\sigma):\theta_1\in[-k_1,k_1],\theta_2\in A_{\theta_2},\sigma\in[\sigma_1,\sigma_2]\},$$

(b) $\frac{1}{\underline{\sigma}\sqrt{2\pi}}P_0(B^c)\le d$,

(c) $I_B(\theta_1,\theta_2,\sigma)\le t(\theta_1,\theta_2,\sigma)\le I_{B_L}(\theta_1,\theta_2,\sigma)$.

Hence for any $|x| < K_1$ and $|y| < K_2$,

$$\begin{aligned}
&\int_{B^c} \phi_\sigma(y-\theta_1-F(x,\theta_2))dP_0(\theta_1,\theta_2,\sigma)\\
\leq &\int_{B^c} \frac{1}{\underline{\sigma}\sqrt{2\pi}} dP_0(\theta_1,\theta_2,\sigma)\\
\leq &\frac{1}{\underline{\sigma}\sqrt{2\pi}} P_0(B^c)\\
\leq &d = \frac{\varepsilon}{8}\int_{\theta_1\in[-k_1,k_1],\theta_2\in A_{\theta_2},\sigma\in[\sigma_1,\sigma_2]} \phi_\sigma(y-\theta_1-F(x,\theta_2))dP_0(\theta_1,\theta_2,\sigma)\\
\leq &\frac{\varepsilon}{8}\int_B \phi_\sigma(y-\theta_1-F(x,\theta_2))dP_0(\theta_1,\theta_2,\sigma)\\
\leq &\frac{\varepsilon}{8}\int \phi_\sigma(y-\theta_1-F(x,\theta_2))t(\theta_1,\theta_2)dP_0(\theta_1,\theta_2,\sigma).
\end{aligned}$$

So for any $|x| < K_1$ and $|y| < K_2$,

$$\begin{aligned}
&\int \phi_\sigma(y-\theta_1-F(x,\theta_2))dP_0(\theta_1,\theta_2,\sigma)\\
< &\left(1+\frac{\varepsilon}{8}\right)\int \phi_\sigma(y-\theta_1-F(x,\theta_2))t(\theta_1,\theta_2)dP_0(\theta_1,\theta_2,\sigma)
\end{aligned}$$

and hence for any $P \in N_p$, $|x| < K_1$ and $|y| < K_2$,

$$\begin{aligned}
&\log\frac{f_{P_0}(y|x)}{f_P(y|x)}\\
\leq &\log\frac{\int \phi_\sigma(y-\theta_1-F(x,\theta_2))dP_0(\theta_1,\theta_2,\sigma)}{\int \phi_\sigma(y-\theta_1-F(x,\theta_2))t(\theta_1,\theta_2,\sigma)dP(\theta_1,\theta_2,\sigma)}\\
\leq &\log\left(1+\frac{\varepsilon}{8}\right)+\log\frac{\int \phi_\sigma(y-\theta_1-F(x,\theta_2))t(\theta_1,\theta_2,\sigma)dP_0(\theta_1,\theta_2,\sigma)}{\int \phi_\sigma(y-\theta_1-F(x,\theta_2))t(\theta_1,\theta_2,\sigma)dP(\theta_1,\theta_2,\sigma)}\\
\leq &\frac{\varepsilon}{8}+\left|\frac{\int \phi_\sigma(y-\theta_1-F(x,\theta_2))t(\theta_1,\theta_2,\sigma)dP_0(\theta_1,\theta_2,\sigma)}{\int \phi_\sigma(y-\theta_1-F(x,\theta_2))t(\theta_1,\theta_2,\sigma)dP(\theta_1,\theta_2,\sigma)}-1\right|
\end{aligned} \tag{4.4.3}$$

since $\log(x) < \log(1+|x-1|) < |x-1|$ for any $x > 0$.

Clearly,

$$c = \inf_{|x|\leq K_1,|y|\leq K_2}\ \inf_{(\theta_1,\theta_2,\sigma)\in B_L} \phi_\sigma\left(y-\theta_1-F(x,\theta_2)\right) > 0$$

The family of functions $\{\phi_\sigma\left(y-\theta_1-F(x,\theta_2)\right) : |x| \leq K_1, |y| \leq K_2\}$, viewed as a set of functions of $(\theta_1,\theta_2,\sigma) \in B_L$, is uniformly continuous. By the Arzela-Ascoli

theorem, given any $\delta > 0$, there exist finitely many points $(x_1, y_1), \ldots, (x_m, y_m)$ such that for any $|x| \leq K_1, |y| \leq K_2$, there exists an i with

$$\sup_{(\theta_1,\theta_2,\sigma)\in B_L} |\phi_\sigma(y - \theta_1 - F(x, \theta_2) - \phi_\sigma(y_i - \theta_1 - F(x_i, \theta_2)| < c\delta,$$

so

$$\sup_{(\theta_1,\theta_2,\sigma)} |\phi_\sigma(y-\theta_1-F(x,\theta_2))t(\theta_1,\theta_2,\sigma)-\phi_\sigma(y_i-\theta_1-F(x_i,\theta_2))t(\theta_1,\theta_2,\sigma)| < c\delta. \quad (4.4.4)$$

Let

$$N_p = \left\{ P : P(B^*) > \frac{1}{2}, \left| \int \phi_\sigma(y_i - \theta_1 - F(x_i, \theta_2))t(\theta_1, \theta_2, \sigma)dP(\theta_1, \theta_2, \sigma) \right.\right.$$

$$\left.\left. - \int \phi_\sigma(y_i - \theta_1 - F(x_i, \theta_2))t(\theta_1, \theta_2, \sigma)dP_0(\theta_1, \theta_2, \sigma) \right| < c\delta; i = 1, \cdots, m \right\}$$

Since N_p is a weak neighborhood of P_0 (refer to Theorem 4.1.1), $\Pi(N_p) > 0$. Let $P \in N_p$. For any $|x| \leq K_1, |y| \leq K_2$, choosing the appropriate (x_i, y_i) from equation (4.4.4), using a simple triangulation argument we get

$$\left| \int \phi_\sigma(y - \theta_1 - F(x, \theta_2))t(\theta_1, \theta_2, \sigma)dP(\theta_1, \theta_2, \sigma) \right.$$

$$\left. - \int \phi_\sigma(y - \theta_1 - F(x, \theta_2))t(\theta_1, \theta_2, \sigma)dP_0(\theta_1, \theta_2, \sigma) \right| < 3c\delta.$$

Also, $\int \phi_\sigma(y - \theta_1 - F(x, \theta_2))t(\theta_1, \theta_2, \sigma)dP_0(\theta_1, \theta_2, \sigma) \geq cP_0(B) \geq c/2$. We get

$$\left| \frac{\int \phi_\sigma(y - \theta_1 - F(x, \theta_2))t(\theta_1, \theta_2, \sigma)dP(\theta_1, \theta_2, \sigma)}{\int \phi_\sigma(y - \theta_1 - F(x, \theta_2))t(\theta_1, \theta_2, \sigma)dP_0(\theta_1, \theta_2, \sigma)} - 1 \right| < 6\delta.$$

So when $\delta = \epsilon/(6\epsilon + 48)$, for any $|x| \leq K_1, |y| \leq K_2, P \in N_p$,

$$\left| \frac{\int \phi_\sigma(y - \theta_1 - F(x, \theta_2))t(\theta_1, \theta_2, \sigma)dP_0(\theta_1, \theta_2, \sigma)}{\int \phi_\sigma(y - \theta_1 - F(x, \theta_2))t(\theta_1, \theta_2, \sigma)dP(\theta_1, \theta_2, \sigma)} - 1 \right| < \frac{6\delta}{1 - 6\delta} \leq \frac{\epsilon}{8}.$$

So for any $P \in N_p$, we have that by equations (4.4.3)

$$\log\left(\frac{f_{P_0}(Y|X)}{f_P(Y|X)}\right) I(|X| \leq K_1, |Y| \leq K_2) < \frac{\epsilon}{4}$$

$$\mathrm{E}_{\pi_0\otimes f_0}\left\{\log\left(\frac{f_{P_0}(Y|X)}{f_P(Y|X)}\right)I(|X|\le K_1,|Y|\le K_2)\right\}<\frac{\epsilon}{4} \tag{4.4.5}$$

Following equations (4.4.5) and (4.4.2), we have that for any $P\in N_p$,

$$\begin{aligned}
&\mathrm{E}_{\pi_0\otimes f_0}\left\{\log\frac{f_{P_0}(Y|X)}{f_P(Y|X)})\right\}\\
\le&\mathrm{E}_{\pi_0\otimes f_0}\left\{(\log\left(\frac{f_{P_0}(Y|X)}{f_P(Y|X)}\right)I(|X|\le K_1,|Y|\le K_2)\right\}\\
&+\mathrm{E}_{\pi_0\otimes f_0}\left\{\left|\log\left(\frac{f_{P_0}(Y|X)}{f_P(Y|X)}\right)\right|I(|Y|>K_2)\right\}\\
&+\mathrm{E}_{\pi_0\otimes f_0}\left\{\left|\log\left(\frac{f_{P_0}(Y|X)}{f_P(Y|X)}\right)\right|I(|X|>K_1)\right\}\\
<&\epsilon
\end{aligned} \tag{4.4.6}$$

Hence given any $\epsilon>0$, we could find a neighborhood N_p of P_0 such that for any $P\in N_p$, equation (4.4.1) holds. □

4.5 Weak Consistency on Densities

In Section 4.2, we extended Schwartz's theorem to Markov processes. Section 4.3 provided sufficient conditions for a transition density to lie in the Kullback-Leibler support of the prior. In this section, we establish sufficient condition for the existence of an exponentially consistent sequence of tests in weak-star topology. If the data are i.i.d., an exponentially consistent sequence of tests will always exist by Hoeffding's inequality, so that the posterior consistency will automatically follow. This is not true in Markov processes since Hoeffding's inequality may not hold again, because of the dependence among data. However, if we focus on the subset of $\pounds(R)$, in which all Markov processes are uniformly ergodic, for any bounded function g, we could decompose the sum series $\{\sum_{i=1}^n[g(X_i)-\pi_f(g)],\, n\ge 1\}$ into a sum of bounded martingale difference sequence plus a bounded term and use Azuma's inequality to

find the uniformly exponentially consistent sequence of tests. Our result on Hoeffding-type inequality in uniformly ergodic Markov processes and the construction of an exponentially consistent sequence of tests is given in Theorem 4.5.2. Theorem 4.5.1 presents a result on posterior consistency in the weak-star topology.

Theorem 4.5.1. *Suppose that the assumption of Model one introduced in Section 4.2 holds, the true model has the form*

$$f_0(y|x) = f_{P_0}(y|x) = \int \phi_\sigma(y - \theta_1 - F(x, \theta_2))dP(\theta_1, \theta_2, \sigma),$$

and f_0 is in the Kullback-Leibler support of the prior. Furthermore, the initial point x_0 is either fixed or has a known distribution. Then the posterior is weakly consistent at any f_0.

Proof. Let

$$V_w = \left\{ f_P : \left| \int g_i(x)\pi(x)dx - \int g_i(x)\pi_0(x)dx \right| < \epsilon, \|g_i\| \leq M_i, i = 1, \ldots, k \right\},$$

where g_i's are continuous functions, be a weak neighborhood of f_0. Let $g^*_{i1} = \frac{1}{2} + \frac{g_i}{2M_i}$, $g^*_{i2} = 1 - g_{i1}$ and $\delta_i = \epsilon/(2M_i)$, then g_{i1} and g_{i2} is continuous functions with values lying in $[0, 1]$, and $V_w = \cap_{i=1}^k \cap_{j=1}^2 V_{ij}$, where

$$V_{ij} = \{f_P : \int g^*_{ij}(x)\pi(x)dx - \int g^*_{ij}(x)\pi_0(x)dx < \delta_i\}.$$

By Lemma 4.5.1, there exist an exponentially consistent sequence of tests for

$$H_0 : f = f_0 \quad vs \quad H_{ij} : f \in V^c_{ij}.$$

So for $j = 1, 2$, $i = 1, \ldots, k$,

$$\Pi(V^c_{ij}|X_0, \ldots, X_n) \to 0 \text{ a.s. } P^\infty_{f_0}$$

by Theorem 4.3.1 as f_0 is in the K-L support of the prior. So

$$\Pi(V^c_w|X_0, \ldots, X_n) \leq \sum_{i=1}^k \sum_{j=1}^2 \Pi(V^c_{ij}|X_0, \ldots, X_n) \to 0 \text{ a.s. } P^\infty_{f_0}.$$

Equivalently,

$$\Pi(V_w|X_0,\ldots,X_n) \to 1 \text{ a.s. } P_{f_0}^{\infty}.$$

□

Theorem 4.5.2. *Consider a Markov process with the state space $(\chi, \mathcal{B}(\chi))$ and transition probability measure P which is uniformly ergodic, that is, there exists $\varepsilon > 0$ and a probability v such that for all $A \in \mathcal{B}(\chi)$,*

$$\inf_{x\in\chi} P^m(x,A) \geq \varepsilon v(A).$$

Let π denote the invariant probability measure for the chain, and $g : \chi \to [l,u]$ be a measurable function. Then
(i) $G(x) = \sum_{i=0}^{\infty}[P^i(x,g) - \pi(g)]$ solves the Poisson equation

$$G(x) - PG(x) = g(x) - \pi(g),$$

and $|G(x)|$ is bounded by

$$R = \frac{u-l}{2[1-(1-\varepsilon)^{1/m}]}$$

(ii) If a satisfies $na > 2R$, then for any initial distribution of X_0,

$$\Pr\left(\sum_{i=1}^{n}[g(X_i) - \pi(g)] \geq na\right) \leq \exp\left[\frac{-n(a-2R/n)^2}{2R^2}\right] \tag{4.5.1}$$

$$\Pr\left(\sum_{i=1}^{n}[g(X_i) - \pi(g)] \leq -na\right) \leq \exp\left[\frac{-n(a-2R/n)^2}{2R^2}\right] \tag{4.5.2}$$

Proof. (i) Let $\tilde{g}(x) = g(x) - (u-l)/2$. Then $\tilde{g}$ is bounded by $(u-l)/2$. If G solve the Poisson equation $G(x) - PG(x) = g(x) - \pi(g)$, it also solves the Poisson equation $G(x) - PG(x) = \tilde{g}(x) - \pi(\tilde{g})$. By Theorem A.2.8, for any $x \in \chi$,

$$|G(x)| \leq R = \frac{u-l}{2[1-(1-\varepsilon)^{1/m}]}$$

(ii) As denoted in equation (A.2), the sum series $S_n(\overline{g})$ could be written as

$$S_n(\overline{g}) = \sum_{i=1}^{n}[G(X_i) - PG(X_i)] = \sum_{i=1}^{n}[G(X_i) - PG(X_{i-1})] + PG(x_0) - PG(x_n),$$

where $[G(X_i) - PG(X_{i-1})]$, $i = 1, 2, \ldots$, is a martingale difference sequence and the remaining term $PG(X_0) - PG(X_n)$ is bounded by $2R$. So we have

$$\begin{aligned}
&\Pr\left(\sum_{i=1}^{n}[g(X_i) - \pi(g)] \geq na\right) \\
=&\Pr\left(\sum_{i=1}^{n}[G(X_i) - PG(X_{i-1})] + PG(X_0) - PG(X_n) \geq na\right) \\
\leq&\Pr\left(\sum_{i=1}^{n}[G(X_i) - PG(X_{i-1})] \geq n(a - 2R/n)\right)
\end{aligned}$$

and similarly

$$\Pr\left(\sum_{i=1}^{n}[g(X_i) - \pi(g)] \leq -na\right) \leq \Pr\left(\sum_{i=1}^{n}[G(X_i) - PG(X_{i-1})] \leq -n(a - 2R/n)\right)$$

Note that $-R - PG(X_{i-1}) \leq [G(X_i) - PG(X_{i-1})] \leq R - PG(X_{i-1})$. An application of Azuma's inequality (Theorem C.0.2) yields the inequalities (4.5.1) and (4.5.2). □

Corollary 4.5.1. *Let $f = f_P$ be a Markov transition density function, where P is in the weak-star support of $D_{\alpha G_0}$. Let g be any nonnegative function bounded by* 1. *Then for any δ, when $n \geq 4M/\delta$ where $M = 1/\left[2\int_{|y|>a/\underline{\sigma}}\phi(y)dy\right]$ and $\phi(y)$ is the density of the standard normal distribution,*

$$\begin{aligned}
\Pr\left(\sum_{i=1}^{n}[g(X_i) - \pi(g)] \geq n\delta\right) &\leq \exp\left[\frac{-n\delta^2}{8M^2}\right] \\
\Pr\left(\sum_{i=1}^{n}[g(X_i) - \pi(g)] \leq -n\delta\right) &\leq \exp\left[\frac{-n\delta^2}{8M^2}\right]
\end{aligned} \tag{4.5.3}$$

Proof. By Lemma 4.4.1, the chain is uniformly ergodic. This corollary follows from Theorem 4.5.2 with $R = M$. So when $n \geq 2R/\delta$,

$$\Pr\left(\sum_{i=1}^{n}[g(X_i) - \pi(g)] \geq n\delta\right) \leq \exp\left[\frac{-n(\delta - 2R/n)^2}{2R^2}\right],$$

$$\Pr\left(\sum_{i=1}^{n}[g(X_i)-\pi(g)]\le -n\delta\right)\le \exp\left[\frac{-n(\delta-2R/n)^2}{2R^2}\right].$$

Note $R=M$. When $n\ge 4M/\delta$,

$$\exp\left[\frac{-n(\delta-2R/n)^2}{2R^2}\right]\le \exp\left[\frac{-n\delta^2}{8M^2}\right].$$

So equation (4.5.3) holds. □

Lemma 4.5.1. *Under the assumption of Theorem* 4.5.1*, there exists a sequence of tests which are exponentially consistent for testing*

$$H_0: f=f_0 \quad vs\ H_1: f\in\left\{V:\int g(x)\pi(dx)>\int g(x)\pi_0(dx)+\varepsilon\right\}$$

for any nonnegative measurable function g bounded by 1*, i.e. there exists C, β and n_0 and a nonnegative sequence test ϕ_n bounded by* 1 *such that for all $n>n_0$.*

$$\mathrm{E}_{f_0}(\phi_n)\le C\exp\{-n\beta\}\ and\ \sup_{f\in V}\mathrm{E}_f(1-\phi_n)\ge C\exp(-n\beta) \tag{4.5.4}$$

Proof. Let $\alpha=\pi_0(g)$ and $\gamma=\inf_{f\in V}\pi(g)>\alpha$. Then

$$\phi_n=I\left(\sum_{i=1}^{n}g(X_i)>n(\alpha+\gamma)/2\right),n\ge 1,$$

is such a sequence of test functions. By Corollary 4.5.1, When $n\ge n_0=8M/(\gamma-\alpha)$ where $M=1/\left[2\int_{|y|>a/\underline{\sigma}}\phi(y)dy\right]$, we have that

$$\mathrm{E}_{f_0}(\phi_n)=\Pr_{f_0}\left(\sum(g(X_i)-\pi_0(g))>n(\gamma-\alpha)/2\right)\le\exp\left[\frac{-n(\gamma-\alpha)^2}{32M^2}\right]$$

and when $n\ge n_0$, for any $f\in V$, we have that

$$\mathrm{E}_f(1-\phi_n)\le\Pr_f\left(\sum(g(X_i)-\pi(g))<-n(\gamma-\alpha)/2\right)\le\exp\left[\frac{-n(\gamma-\alpha)^2}{32M^2}\right]$$

□

4.6 Strong Consistency Under sup-L_1 metric

In the last section, we presented a result on posterior consistency in the weak-star topology. As discussed in Section 4.1, the weak-star neighborhoods are large, and the transition densities in a weak-star neighborhood may not be identical since different transition densities may be corresponding to the same invariant probability measure. The sup-L_1 metric defined below is a very strong neighborhood and avoids the non-identity issue. A result on posterior consistence in the sup-L_1 topology is given in Theorem 4.6.1.

Theorem 4.6.1. *Denote* $\theta = (v, \varphi, \gamma, u, \sigma)$,

$$\psi_\theta(y|x) = \psi_{v,\varphi,\gamma,u,\sigma}(y|x) = \phi_\sigma\left(y - v - \varphi\frac{1}{1+\exp[-\gamma(x-u)]}\right).$$

and $f_P(y|x) = \int \psi_\theta(y|x)dP(\theta)$. *Under the assumption of Model two defined in Section 4.2, assume that the true transition density be of the form* $f_0(y|x) = f_{P_0}(y|x)$. *Suppose that*

$$G_0([\underline{v}, \overline{v}], [\underline{\gamma}, \overline{\gamma}], [\underline{\varepsilon}, \overline{\varepsilon}], [\underline{u}, \overline{u}], [\underline{\sigma}, \overline{\sigma}]) = 1,$$

and the interval $[\underline{\gamma}, \overline{\gamma}]$ *does not contain 0,*
(i) If P_0 *in the weak-star support of Dirichlet prior* $D_{\alpha G_0}$, *then* f_0 *is in the K-L support of the prior,*
(ii) If f_0 *is in the K-L support of the prior, the posterior distribution is strongly consistent at* f_0 *under the* $\sup - L_1$ *metric.*

The key observation is that the prior Π is supported on a compact subspace of $\pounds(R)$ under sup-L_1 topology. Here, the support of γ cannot contain 0. Otherwise the space under sup-L_1 metric may not be compact. Let see the counterexample below.

Counterexample 4.6.1. Consider the class of the degenerate measures $\varpi = \{P : P = \delta_{u=0,v=0,\varphi=1,\gamma,\sigma=1} : 0 \leq \gamma \leq 1\}$. Let $P_1 = \delta_{(0,0,1,\gamma_1,1)}$ and $P_2 = \delta_{(0,0,1,\gamma_2,1)}$ and

suppose that $a = \gamma_2/\gamma_1 > 1$. Then

$$\begin{aligned}
&\sup_x ||f_{P_1}(\cdot|x) - f_{P_2}(\cdot|x)|| \\
&= \sup_x \int \frac{1}{\sqrt{2\pi}} \left| \exp\left[-\frac{1}{2}(y - \frac{1}{1+\exp(-x\theta_1)})^2\right] - \exp\left[-\frac{1}{2}(y - \frac{1}{1+\exp(-x\theta_2)})^2\right] \right| dy \\
&= \sup_x \frac{4}{\sqrt{2\pi}} \int_0^{\frac{1}{2}\left|[1+\exp(-x\theta_1)]^{-1} - [1+\exp(-x\theta_2)]\right|} e^{-y^2/2} dy \\
&= \sup_x \frac{4}{\sqrt{2\pi}} \int_0^{\frac{1}{2}\left|[1+\exp(-x)]^{-1} - [1+\exp(-xa)]^{-1}\right|} e^{-y^2/2} dy
\end{aligned}$$

From the above equation, we see that $\sup_x ||f_{P_1}(\cdot|x) - f_{P_2}(\cdot|x)||$ depends on a only and increases with a. If we want to find the cardinality of the minimal η-net of ϖ. let a be satisfy $\eta = \sup_x ||f_{P_1}(\cdot|x) - f_{P_2}(\cdot|x)||$. Then we can find infinitely many *eta*-separated points $f_{\delta_{(0,0,1,\gamma_i,1)}}$ where $\gamma_i = 1/(1+a)^i$ for $i = 1, 2, \ldots$.

proof of Theorem 4.6.1. (i) follows from Theorem 4.4.1.

(ii) By Lemma 4.6.4, we could find finite nets $V_1, \ldots, V_k$ covering V^C. For each V_i, $i = 1, \ldots, k$ we could find a sequence of exponentially consistent tests for testing f_0 against $f \in V_i$. Since f_0 is in the K-L support of the prior, the theorem follows from Corollary 4.3.1. □

Proposition 4.6.1. [§] *Let a pair of random variables (u, β) be independent of ϵ, and ϵ has density $f(\epsilon)$ with respect to the Lebesgue measure. Suppose that $f_1(y)$ and $f_2(y)$ are the density function of $y = u + \beta + \epsilon$ and $y = u + \beta I(|\beta| \le a) + \epsilon$ respectively, The total variation distance between f_1 and f_2 is bounded by*

$$||f_1 - f_2|| \le 2\Pr(|\beta| \ge a).$$

Proof. Let the distribution of (u, β) be P. Then the density of $y = u + \beta + \epsilon$ is given by

$$f_1(y) = \int_\beta \int_u f(y - u - \beta) dP(u, \beta),$$

[§]A similar result is given in [37] in which $u = 0$.

and the density of $y = u + \beta I(|\beta| \leq a) + \epsilon$ is given by

$$f_2(y) = \frac{\int_{|\beta|\leq a} \int_u f(y-u-\beta) dP(u,\beta)}{\Pr(|\beta| \leq a)}.$$

$$\begin{aligned}
&||f_1 - f_2|| \\
&= \int_y |f_1(y) - f_2(y)| dy \\
&= \int_y \left| \int_\beta \int_u f(y-u-\beta) dP(u,\beta) - \frac{\int_{|\beta|\leq a} \int_u f(y-u-\beta) dP(u,\beta)}{\Pr(|\beta| \leq a)} \right| dy \\
&= \int_y \left| \int_{|\beta|\leq a} \int_u f(y-u-\beta) dP(u,\beta) \left(\frac{1}{\Pr(|\beta| \leq a)} - 1 \right) - \right. \\
&\qquad \left. \int_{|\beta|\geq a} \int_u f(y-u-\beta) dP(u,\beta) \right| dy \\
&= \int_y \int_{|\beta|\leq a} \int_u f(y-u-\beta) dP(u,\beta) \left(\frac{1}{\Pr(|\beta| \leq a)} - 1 \right) dy + \\
&\qquad \int_y \int_{|\beta|\geq a} \int_u f(y-u-\beta) dP(u,\beta) dy \\
&= \Pr(|\beta| \leq a) \left(\frac{1}{\Pr(|\beta| \leq a)} - 1 \right) + \Pr(|\beta| \geq a) \\
&= 2 \Pr(|\beta| \geq a)
\end{aligned}$$

□

Proposition 4.6.2. ¶ *Let $N(u, \sigma^2)$ denote the normal distribution with mean u and variance σ^2. Then the total variation distance between $N(\theta_1, \sigma^2)$ and $N(\theta_2, \sigma^2)$ is*

$$\frac{4}{\sqrt{2\pi}} \int_0^{\frac{|\theta_2-\theta_1|}{2\sigma}} \exp\left[-\frac{x^2}{2}\right] dx \leq \sqrt{\frac{2}{\pi}} \frac{|\theta_2 - \theta_1|}{\sigma}.$$

¶This proposition is given in [36].

Proof. Without loss of generality, we assume that $\theta_1 \leq \theta_2$. Now

$$\begin{aligned}
&\int \left| \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\theta_1)^2}{2\sigma^2}\right] - \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\theta_2)^2}{2\sigma^2}\right] \right| dx \\
&= \int \left| \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x+(\theta_2-\theta_1)/2)^2}{2\sigma^2}\right] - \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-(\theta_2-\theta_1)/2)^2}{2\sigma^2}\right] \right| dx \\
&= \frac{4}{\sqrt{2\pi}} \int_0^{\frac{|\theta_2-\theta_1|}{2\sigma}} \exp\left[-\frac{x^2}{2}\right] dx \\
&\leq \sqrt{\frac{2}{\pi}} \frac{|\theta_2-\theta_1|}{\sigma}.
\end{aligned}$$

□

Proposition 4.6.3. [‖] *Let* $\wp_N = \{(P_1, P_2, \cdots, P_N) : P_i \geq 0, \sum_{i=1}^N P_i = 1\}$ *be the* N*-dimensional probability simplex and let* $\wp_N^*$ *be a* δ*-net in* $\wp_N$*, i.e., given* $P \in \wp_N$*, there is* $P^* = (P_1^*, \cdots, P_N^*) \in \wp_N^*$ *such that* $\sum_{i=1}^N |P_i - P_i^*| < \delta$*. The cardinality of the minimal* δ*-net of* P_N *is smaller than* $(N/\delta)^N(1+\delta)^N/N!$*.*

Proof. Since $|P_i - P_i^*| < \delta/N$ for all i implies that $\sum_{i=1}^N |P_i - P_i^*| < \delta$. An upper bound for the cardinality of the minimal δ-net of P_N is given by

$$\#\text{cubes of length } \delta/N \text{ covering } [0,1]^N \times \text{ volume of}$$
$$\left\{(P_1, P_2, \cdots, P_N) : P_i \geq 0, \sum_{i=1}^N P_i \leq 1+\delta\right\} = (N/\delta)^N(1+\delta)^N/N!$$

□

Lemma 4.6.1. *Let*

$$F_{h,\underline{\tau},\overline{\tau}} = \left\{ \int \phi_h\left(y|x,\tau\right) dP(\tau) : P([\underline{\gamma},\overline{\gamma}],[\underline{v},\overline{v}],[\underline{u},\overline{u}],[\underline{\varphi},\overline{\varphi}]) = 1 \right\},$$

where $\tau = (\gamma, u, v, \varphi)$*, and* $\phi_h(y|x,\tau) = \phi_h\left(y - v - \varphi \frac{1}{1+\exp[-\gamma(x-u)]}\right)$*. If* $[\underline{\gamma},\overline{\gamma}]$ *does not contain* 0 *and* $h \in [\underline{\sigma},\overline{\sigma}]$ *is a fixed point, the space* $F_{h,\underline{\tau},\overline{\tau}}$ *is compact with respect to the* sup-L_1 *distance.*

[‖] This proposition is given in [36].

Proof. Note that $\underline{\gamma} > 0$.

(i) When x is small enough, (say $x \leq -M_1 < 0$),

$$\left| \frac{\varphi}{1+\exp[-\gamma(x-u)]} \right| \leq \frac{|\overline{\varphi}|}{1+\exp[\underline{\gamma}(\underline{u}+M_1)]} \leq \frac{\delta}{4}\sqrt{\frac{\pi}{2}}\frac{\underline{\sigma}}{\max|\varphi|} \tag{4.6.1}$$

(ii) When x is big enough, (say $x \geq M_2 > 0$),

$$\frac{1}{1+\exp[-\gamma(x-u)]} \geq \frac{1}{1+\exp[-\underline{\gamma}(M_2-\overline{u})]} \geq 1 - \frac{\delta}{8}\sqrt{\frac{\pi}{2}}\frac{\underline{\sigma}}{\max|\varphi|} \tag{4.6.2}$$

(iii) When $|x| \leq M$ (where $M = \max\{M_1, M_2\}$), by Proposition 4.6.2,

$$\begin{aligned}
&||\phi_{\tau_1,h}(\cdot|x) - \phi_{\tau_2,h}(\cdot|x)|| \\
&\leq \sqrt{\frac{2}{\pi}} \frac{\left| v_1 + \frac{\varphi_1}{1+\exp[-\gamma_1(x-u_1)]} - v_2 - \frac{\varphi_2}{1+\exp[-\gamma_2(x-u_2)]} \right|}{h} \\
&\leq \sqrt{\frac{2}{\pi}} \frac{|v_1 - v_2| + |\varphi_1 - \varphi_2| + \max|\varphi| \left| \frac{1}{1+\exp[-\gamma_1(x-u_1)]} - \frac{1}{1+\exp[-\gamma_2(x-u_2)]} \right|}{h}
\end{aligned}$$

By the Arzela-Ascoli theorem, given $\delta > 0$, there exists $c > 0$ such that, whenever $|u_1 - u_2| < c$, $|v_1 - v_2| < c$, $|\theta_1 - \theta_2| < c$ and $|\varphi_1 - \varphi_2| < c$,

$$\sup_{h\in[\underline{\sigma},\overline{\sigma}],|x|\leq M} |\phi_{u_1,v_1,\gamma_1,\varphi_1,h}(\cdot|x) - \phi_{u_2,v_2,\gamma_2,\varphi_2,h}(\cdot|x)| \leq \delta \tag{4.6.3}$$

Let $d = \min\{c, \frac{\underline{\sigma}\delta}{4\max|\varphi|}\sqrt{\frac{\pi}{2}}\}$. Let N_1 be the smallest integer greater than $(\overline{u} - \underline{u})/d$, N_2 be the smallest integer greater than $(\overline{v} - \underline{v})/d$, N_3 be the smallest integer greater than $(\overline{\gamma} - \underline{\gamma})/d$, N_4 be the smallest integer greater than $(\overline{\varphi} - \underline{\varphi})/d$. Divide the support of $\tau = (u, v, \varphi, \gamma)$ into $N = N_1N_2N_3N_4$ boxes $E_1, E_2, \ldots, E_N$. Let τ_i, $i = 1, \cdots, N$, be the corresponding center points in each box E_i. Let $\wp_N = \{(P_1, P_2, \ldots, P_N) : P_i \geq 0, \sum_{i=1}^N P_i = 1\}$ be the N-dimensional probability simplex. Let $\wp_N^*$ be a δ-net of $\wp_N$, that is, given $p \in \wp_N$, there exists $P^* = (P_1^*, \ldots, P_N^*) \in \wp_N^*$ such that $\sum_{i=1}^N |P_i - P_i^*| < \delta$. By Proposition 4.6.3, the cardinality of $\wp_N^*$ is finite given N.

Let $F^* = \{\sum_{i=1}^N P_i^* \phi_{\tau_i,h} : P^* = (P_1^*, \cdots, P_N^*) \in \wp_N^*\}$. Then we can prove that F^* is a 2δ-net of F_h. If $f_{h,P} = \phi_h * P \in F_h$, set $P_i = P(E_i)$ and let $P^* \in \wp_N^*$ be such that $\sum_{i=1}^N |P_i - P_i^*| < \delta$, then

$$\begin{aligned}
&\left\| \int \phi_{\tau,h}(\cdot|x) dP(\tau) - \sum_{i=1}^N \int P_i^* \phi_{\tau_i,h}(\cdot|x) \right\| \\
\leq &\left\| \int \phi_{\tau,h}(\cdot|x) dP(\tau) - \sum_{i=1}^N \int I_{E_i}(\tau) \phi_{\tau_i,h}(\cdot|x) dP(\tau) \right\| \\
&+ \left\| \sum_{i=1}^N P_i^* \phi_{\tau_i,h}(\cdot|x) - \sum_{i=1}^N P_i \phi_{\tau_i,h}(\cdot|x) \right\| \qquad (4.6.4) \\
&\leq \int \sum_{i=1}^N I_{E_i}(\tau) \|\phi_{\tau,h}(\cdot|x) - \phi_{\tau_i,h}(\cdot|x)\| dP(\tau) + \sum_{i=1}^N |P_i - P_i^*| \\
&\leq \int \sum_{i=1}^N I_{E_i}(\tau) \|\phi_{\tau,h}(\cdot|x) - \phi_{\tau_i,h}(\cdot|x)\| dP(\tau) + \delta
\end{aligned}$$

Now we try to bound the term $\|\phi_{\tau,h}(\cdot|x) - \phi_{\tau_i,h}(\cdot|x)\|$ where τ lie in the box E_i and τ_i is the center of E_i for any i. When $|x| \leq M$, by equation (4.6.3)

$$\sup_{|x| \leq M} \|\phi_{\tau,h}(.|x) - \phi_{\tau_i,h}(.|x)\| \leq \delta.$$

When $x > M$, by equation (4.6.1) and Proposition 4.6.2,

$$\begin{aligned}
&\|\phi_{\tau,h}(.|x) - \phi_{\tau_i,h(.|x)}\| \\
\leq &\sqrt{\frac{2}{\pi}} \frac{\left| v + \frac{\varphi}{1+\exp[-\gamma(x-u)]} - v_i + \frac{\varphi_i}{1+\exp[-\gamma_i(x-u_i)]} \right|}{h} \\
\leq &\sqrt{\frac{2}{\pi}} \frac{|v - v_i| + |\varphi - \varphi_i| + \max|\varphi| \left| \frac{1}{1+\exp[-\gamma(x-u)]} - \frac{1}{1+\exp[-\gamma_i(x-u_i)]} \right|}{h} \\
\leq &\sqrt{\frac{2}{\pi}} \frac{\frac{\delta}{2}\sqrt{\frac{\pi}{2}}\underline{\sigma} + \frac{\delta}{4}\sqrt{\frac{\pi}{2}}\underline{\sigma} + \frac{\delta}{4}\sqrt{\frac{\pi}{2}}\underline{\sigma}}{h} \\
= &\delta.
\end{aligned}$$

When $x < -M$, by equation (4.6.2) and Proposition 4.6.2,

$$\begin{aligned}
&\|\phi_{\tau,h}(\cdot|x) - \phi_{\tau_i,h}(\cdot|x)\| \\
\leq&\sqrt{\frac{2}{\pi}}\frac{\left|v + \frac{\varphi}{1+\exp[-\gamma(x-u)]} - v_i + \frac{\varphi_i}{1+\exp[-\gamma_i(x-u_i)]}\right|}{h} \\
\leq&\sqrt{\frac{2}{\pi}}\frac{|v - v_i| + \frac{|\varphi|}{1+\exp[-\gamma(x-u)]} + \frac{|\varphi_i|}{1+\exp[-\gamma_i(x-u_i)]}}{h} \\
\leq&\sqrt{\frac{2}{\pi}}\frac{|v - v_i| + 2\frac{\delta}{4}\sqrt{\pi}2\underline{\sigma}}{h} \\
\leq&\delta
\end{aligned}$$

So for any x, $||\phi_{\tau,h}(\cdot|x) - \phi_{\tau_i,h}(\cdot|x)|| \leq \delta$ provided that τ, τ_i lie in the same box E_i. So we can see that by equation (4.6.4), for any x,

$$\begin{aligned}
&\| \int \phi_{\tau,h}(\cdot|x)dP(\tau) - \sum_{i=1}^{N} \int P_i^* \phi_{\tau_i,h}(\cdot|x)\| \\
\leq& \int \sum_{i=1}^{N} I_{E_i}(\tau)\|\phi_{\tau,h}(\cdot|x) - \phi_{\tau_i,h}(\cdot|x)\|dP(\tau) + \delta \\
\leq&2\delta
\end{aligned}$$

Since the cardinality of $\wp_N^*$ is finite, the space of F_h is compact. □

Lemma 4.6.2. *Let* $F_{\underline{\theta},\overline{\theta}} = \{P * \phi(y|x,\theta) : P([\underline{\gamma},\overline{\gamma}],[\underline{v},\overline{v}],[\underline{u},\overline{u}],[\underline{\varphi},\overline{\varphi}],[\underline{\sigma},\overline{\sigma}]) = 1\}$, *where* $\phi(y|x,\theta) = \phi_\sigma\left(y - v - \varphi\frac{1}{1+\exp[-\gamma(x-u)]}\right)$ *and* $[\underline{\gamma},\overline{\gamma}]$ *does not contain* 0. *Then* $F_{\underline{\theta},\overline{\theta}}$ *is compact with respect to the* sup-L_1 *distance.*

Proof. Note that for any Markov process in $F_{\underline{\theta},\overline{\theta}}$, its transition density f_P is given by

$$X_{n+1} = v + \frac{\varphi}{1+\exp[-\gamma(X_n - u)]} + \varepsilon_{n+1},$$

or equivalently by

$$X_{n+1} = v + \varepsilon_{n+1,2} + \frac{\varphi}{1+\exp[-\gamma(X_n - u)]} + \varepsilon_{n+1,1}, \tag{4.6.5}$$

where

$$(u, v, \varphi, \gamma, \sigma) \sim P,$$

$$\varepsilon_{n+1} \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad \text{given } P,$$

$$\varepsilon_{n+1,1} \overset{i.i.d.}{\sim} N(0, \underline{\sigma}^2),$$

$$\varepsilon_{n+1,2} \overset{i.i.d.}{\sim} N(0, \sigma^2 - \underline{\sigma}^2), \quad \text{given } P,$$

and the series of random variables $\{\varepsilon_{n+1,1} : n \geq 0\}$ are independent of $\{\varepsilon_{n+1,2} : n \geq 0\}$.

Let F^* denote the set of all such models:

$$X_{n+1} = v' + \frac{\varphi}{1 + \exp[-\theta(X_n - u)]} + \varepsilon_{\underline{\sigma},n+1},$$

where $v' = v + \varepsilon_{n+1,2} I\{|\varepsilon_{n+1,2}| \leq b\}$, with corresponding transition density f_P^*. By Proposition 4.6.1,

$$\begin{aligned} \sup_x ||f_P(\cdot|x) - f_P^*(\cdot|x)|| &\leq 2\Pr(|\varepsilon_{n+1,2}| \geq b) \\ &\leq 2\mathrm{E}(\varepsilon_{n+1,2}^2)/b^2 \\ &= 2\mathrm{E}_P\left[\mathrm{E}(\varepsilon_{n+1,2}^2|P)\right]/b^2 \\ &= 2\mathrm{E}_P[\sigma^2 - \underline{\sigma}^2|P]/b^2 \\ &\leq 2(\overline{\sigma}^2 - \underline{\sigma}^2)/b^2 \end{aligned} \tag{4.6.6}$$

So we can find a sufficiently large b such that

$$(\overline{\sigma}^2 - \underline{\sigma}^2)/b^2 \leq 2\delta.$$

So $J(3\delta, F_{\underline{\sigma},\overline{\sigma},\underline{\tau},\overline{\tau}}) \leq J(\delta, F^*)$, where $J(\delta, \mathcal{F})$ is the logarithm of the minimal number of balls of radius δ in total variation metric needed to cover the $\mathcal{F}$. By Lemma 4.6.1, F^* and hence $F_{\underline{\theta},\overline{\theta}}$ is compact with respect to the sup-L_1 distance.

□

Proposition 4.6.4. *Under the assumption of Theorem* 4.6.1 *and* $|\overline{v}| + |\overline{\varphi}| \leq a$, *the following inequalities hold for any* $f_{P_1}, f_{P_2} \in F_{\underline{\theta},\overline{\theta}}$, *any* x_n *and any Borel set* A.

$$\frac{1}{\sqrt{2\pi}\overline{\sigma}} \exp\left[-\frac{(|x_{n+1}| + |a|)^2}{2\underline{\sigma}^2}\right] \leq f_{P_2}(x_{n+1}|x_n)$$
$$\leq \frac{1}{\sqrt{2\pi}\underline{\sigma}} \exp\left[-\frac{(|x_{n+1}| - |a|)^2}{2\overline{\sigma}^2}\right] \exp\left[\frac{a^2}{2\underline{\sigma}^2}\right],$$

$$\Pr_{f_2}(X_{n+1} \in A|X_n) \geq \frac{1}{\sqrt{2\pi}\overline{\sigma}} \int_A \exp\left[-\frac{(|y| + a)^2}{2\underline{\sigma}^2}\right] dy$$

and

$$\frac{f_{P_1}(x_{n+1}|x_n)}{f_{P_2}(x_{n+1}|x_n)} \leq \frac{\overline{\sigma}}{\underline{\sigma}} \exp\left[\frac{(|x_{n+1}| + a)^2}{2\underline{\sigma}^2}\right]$$

Proof. The first equation is easy to prove if we note that for any $|b| \leq a$,

$$(|x_{n+1}| - |b|)^2 \leq (x_{n+1} - b)^2 \leq (|x_{n+1}| + |b|)^2 \leq (|x_{n+1}| + a)^2.$$

Then for any $\sigma \in [\underline{\sigma}, \overline{\sigma}]$,

$$\frac{1}{\sqrt{2\pi}\overline{\sigma}} \exp\left[-\frac{(|x_{n+1}| + a)^2}{2\underline{\sigma}^2}\right] \leq \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(|x_{n+1}| + a)^2}{2\sigma^2}\right]$$
$$\leq \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left(x_{n+1} - v - \varphi\frac{1}{1+\exp[-\gamma(x-u)]}\right)^2}{2\sigma^2}\right]$$
$$\leq \frac{1}{\sqrt{2\pi}\underline{\sigma}} \exp\left[-\frac{\left(|x_{n+1}| - |v + \varphi\frac{1}{1+\exp[-\gamma(x-u)]}|\right)^2}{2\sigma^2}\right]$$
$$\leq \frac{1}{\sqrt{2\pi}\underline{\sigma}} \exp\left[-\frac{(|x_{n+1}| - |a|)^2}{2\sigma^2}\right] \exp\left[\frac{a^2}{2\sigma^2}\right]$$
$$\leq \frac{1}{\sqrt{2\pi}\underline{\sigma}} \exp\left[-\frac{(|x_{n+1}| - |a|)^2}{2\overline{\sigma}^2}\right] \exp\left[\frac{a^2}{2\underline{\sigma}^2}\right]$$

The first equation is easy to get by integrating the above inequalities with respect to P_2. The second equation is easy to get from the first equation. The third equation is straightforward if we note the following two equations,

$$f_{P_1}(x_{n+1}|x_n) \leq \frac{1}{\sqrt{2\pi}\underline{\sigma}},$$

and

$$f_{P_2}(x_{n+1}|x_n) \geq \frac{1}{\sqrt{2\pi}\overline{\sigma}} \exp\left[-\frac{(|x_{n+1}| + |a|)^2}{2\underline{\sigma}^2}\right]$$

□

Proposition 4.6.5. *Under the assumption of Theorem* 4.6.1 *and* $|\overline{v}| + |\overline{\varphi}| \leq a$*, given a positive number* b_1*, there exists* η *such that for all* x_n*, any* $f_{P_1}, f_{P_2}, f_{P_3} \in F_{\underline{\theta},\overline{\theta}}$*, and* $\lambda = \underline{\sigma}^2/(2\overline{\sigma}^2)$*, the inequality*

$$\int \left[\frac{f_{P_1}(x_{n+1}|x_n)}{f_{P_2}(x_{n+1}|x_n)}\right]^{\lambda} [f_{P_3}(x_{n+1}|x_n) - f_{P_2}(x_{n+1}|x_n)]dx_{n+1} \leq b_1,$$

holds whenever

$$\sup_{x_n} \|f_{P_3}(\cdot|x_n) - f_{P_2}(\cdot|x_n)\| \leq \eta.$$

Proof. If we could find b such that

$$\int_{|x_{n+1}| \geq 3a+b} \left[\frac{f_{P_1}(x_{n+1}|x_n)}{f_{P_2}(x_{n+1}|x_n)}\right]^{\lambda} f_{P_3}(x_{n+1}|x_n)dx_{n+1} \leq b_1/2 \tag{4.6.7}$$

for all x_n and any f_{P_1}, f_{P_2}, $f_{P_3} \in F_{\underline{\theta},\overline{\theta}}$,
$\eta = b_1/(2T)$ solves this problem where $T = \sup_{|x_{n+1}| \leq 3a+b} \left\{ \left[\frac{\overline{\sigma}}{\underline{\sigma}}\right]^{\lambda} \exp\left[\frac{\lambda(|x_{n+1}|+a)^2}{2\underline{\sigma}^2}\right] \right\}$.

To see this, note that for any x_n,

$$\begin{aligned}
&\int \left[\frac{f_{P_1}(x_{n+1}|x_n)}{f_{P_2}(x_{n+1}|x_n)}\right]^{\lambda} [f_{P_3}(x_{n+1}|x_n) - f_{P_2}(x_{n+1}|x_n)]dx_{n+1} \\
\leq &\int_{|x_{n+1}|\geq 3a+b} \left[\frac{f_{P_1}(x_{n+1}|x_n)}{f_{P_2}(x_{n+1}|x_n)}\right]^{\lambda} f_{P_3}(x_{n+1}|x_n)dx_{n+1} \\
&\quad + \int_{|x_{n+1}|\leq 3a+b} \left[\frac{f_{P_1}(x_{n+1}|x_n)}{f_{P_2}(x_{n+1}|x_n)}\right]^{\lambda} [f_{P_3}(x_{n+1}|x_n) - f_{P_2}(x_{n+1}|x_n)]dx_{n+1} \\
\leq &\int_{|x_{n+1}|\geq 3a+b} \left[\frac{f_{P_1}(x_{n+1}|x_n)}{f_{P_2}(x_{n+1}|x_n)}\right]^{\lambda} f_{P_3}(x_{n+1}|x_n)dx_{n+1} \\
&\quad + \sup_{|x_{n+1}|\leq 3a+b} \left\{ \left[\frac{f_{P_1}(x_{n+1}|x_n)}{f_{P_2}(x_{n+1}|x_n)}\right]^{\lambda} \right\} ||f_{P_3}(\cdot|x_n) - f_{P_2}(\cdot|x_n)|| \\
\leq &\frac{b_1}{2} + \sup_{|x_{n+1}|\leq 3a+b} \left\{ \left[\frac{\overline{\sigma}}{\underline{\sigma}}\right]^{\lambda} \exp\left[\frac{\lambda(|x_{n+1}|+a)^2}{2\underline{\sigma}^2}\right] \right\} \eta \qquad \text{(by Proposition 4.6.4)} \\
\leq &\frac{b_1}{2} + T\eta = b_1 \text{ for all } x_n
\end{aligned}$$

Since $\int \exp\left[-\frac{y^2}{4\overline{\sigma}^2}\right] dy < \infty$, we can find b satisfying

$$\left[\frac{\overline{\sigma}}{\underline{\sigma}}\right]^{\lambda} \frac{1}{\underline{\sigma}\sqrt{2\pi}} \exp\left[\frac{a^2}{2\underline{\sigma}^2}\right] \exp\left[\frac{2a^2}{\overline{\sigma}^2}\right] \int_{|y|\geq b} \exp\left[-\frac{(|y|)^2}{4\overline{\sigma}^2}\right] dy \leq \frac{b_1}{2} \tag{4.6.8}$$

Now to show b satisfies equation (4.6.7), we observe that

$$\begin{aligned}
&\int_{|x_{n+1}|\geq 3a+b} \left[\frac{f_{P_1}(x_{n+1}|x_n)}{f_{P_2}(x_{n+1}|x_n)}\right]^{\lambda} f_{P_3}(x_{n+1}|x_n)dx_{n+1} \\
\leq &\int_{|x_{n+1}|\geq 3a+b} \left[\frac{\overline{\sigma}}{\underline{\sigma}}\right]^{\lambda} \exp\left[\frac{\lambda(|x_{n+1}|+a)^2}{2\underline{\sigma}^2}\right] \frac{1}{\sqrt{2\pi}\underline{\sigma}} \exp\left[-\frac{(|x_{n+1}|-a)^2}{2\overline{\sigma}^2}\right] \exp\left[\frac{a^2}{2\underline{\sigma}^2}\right] dx_{n+1} \\
&\qquad \text{(by Proposition 4.6.4)} \\
\leq &\int_{|x_{n+1}|\geq 3a+b} \left[\frac{\overline{\sigma}}{\underline{\sigma}}\right]^{\lambda} \exp\left[\frac{(|x_{n+1}|+a)^2}{4\overline{\sigma}^2}\right] \frac{1}{\sqrt{2\pi}\underline{\sigma}} \exp\left[-\frac{(|x_{n+1}|-a)^2}{2\overline{\sigma}^2}\right] \exp\left[\frac{a^2}{2\underline{\sigma}^2}\right] dx_{n+1} \\
\leq &\left[\frac{\overline{\sigma}}{\underline{\sigma}}\right]^{\lambda} \frac{1}{\underline{\sigma}\sqrt{2\pi}} \exp\left[\frac{a^2}{2\underline{\sigma}^2}\right] \exp\left[\frac{2a^2}{\overline{\sigma}^2}\right] \int_{|x_{n+1}|\geq 3a+b} \exp\left[-\frac{(|x_{n+1}|-3a)^2}{4\overline{\sigma}^2}\right] dx_{n+1} \\
\leq &\left[\frac{\overline{\sigma}}{\underline{\sigma}}\right]^{\lambda} \frac{1}{\underline{\sigma}\sqrt{2\pi}} \exp\left[\frac{a^2}{2\underline{\sigma}^2}\right] \exp\left[\frac{2a^2}{\overline{\sigma}^2}\right] \int_{|y|\geq b} \exp\left[-\frac{(|y|)^2}{4\overline{\sigma}^2}\right] dy \\
\leq &\frac{b_1}{2} \qquad \text{(by equation (4.6.8))}
\end{aligned}$$

□

Lemma 4.6.3. *Let P and Q be two probability measures and p, q be their densities with respect to some σ-finite measure μ. Then for any $0 < \alpha < 1$,*

$$\int p^{\alpha} q^{1-\alpha} d\mu \leq \left[1 - \frac{1}{4}\|p-q\|^2\right]^{\beta},$$

where $\|p-q\|$ is the total variation distance between P and Q, and $\beta = \min\{\alpha, 1-\alpha\}$.

Proof. By Lyapounov's Inequality (Lemma C.0.1) and Theorem 4.1.3, when $0 < \alpha \leq \frac{1}{2}$,

$$\int p^{\alpha} q^{1-\alpha} d\mu = \mathrm{E}_Q \left[\frac{p}{q}\right]^{\alpha} \leq \left[\mathrm{E}_Q \sqrt{\frac{p}{q}}\right]^{2\alpha} \leq \left[1 - \frac{1}{4}\|p-q\|^2\right]^{\alpha}.$$

and when $\frac{1}{2} \leq \alpha < 1$,

$$\int p^{\alpha} q^{1-\alpha} d\mu = \mathrm{E}_P \left[\frac{q}{p}\right]^{\beta} \leq \left[\mathrm{E}_P \sqrt{\frac{q}{p}}\right]^{2\beta} \leq \left[1 - \frac{1}{4}\|p-q\|^2\right]^{\beta}.$$

□

Lemma 4.6.4. *Let V_i be a subset of $F_{\underline{\theta},\overline{\theta}}$, $f_0 \in F_{\underline{\theta},\overline{\theta}}$, A_i be a Borel set, ζ and δ_2 are positive numbers such that*

$$2\delta_2 \leq \inf_{x_n \in A_i} ||f_0(\cdot|x_n) - f(\cdot|x_n)||, \text{ and } \zeta \leq \inf_{x_n, f \in V_i} \Pr_f(X_{n+1} \in A_i | x_n).$$

Let $\lambda = \underline{\sigma}^2/(2\overline{\sigma}^2)$, $b_1 = \left(1 - [1-\delta_2^2]^{2\lambda}\right)\zeta/2$ and η be the value given in Proposition 4.6.5 *corresponding to b_1. If for any $f_1, f_2 \in V_i$,*

$$\sup_{x_n} ||f_1(\cdot|x_n) - f_2(\cdot|x_n)|| \leq \eta,$$

then there exists a sequence of exponentially consistent tests for

$$H_0 : f_0 \quad vs\ H_1 : f \in V_i$$

Proof. Let f_1 be any Markov model in V_i. Define the function

$$h(k) = \sum_{l=1}^{k} \log \frac{f_1(x_{2l}|x_{2l-1})}{f_0(x_{2l}|x_{2l-1})}.$$

Consider the following sequence of testing functions:

$$\phi_n = I\{h(k) > 0\} \text{ if } n = 2k \text{ or } n = 2k+1 \tag{4.6.9}$$

We will show that ϕ_n is a sequence of uniformly exponentially consistent tests.

By Lemma 4.6.3,

$$\begin{aligned}
&\int \left[\frac{f_1(x_{n+1}|x_n)}{f_0(x_{n+1}|x_n)}\right]^{\lambda} f_0(x_{n+1}|x_n)dx_{n+1} \\
=&\int [f_1(x_{n+1}|x_n)]^{\lambda}[f_0(x_{n+1}|x_n)]^{1-\lambda}dx_{n+1} \\
\leq&\left[1 - \frac{\|f_1(x_{n+1}|x_n) - f_0(x_{n+1}|x_n)\|^2}{4}\right]^{\lambda} \\
\leq&\begin{cases} [1-\delta_2^2]^{\lambda}, & \text{if } x_n \in A_i, \\ 1, & \text{if } X_n \in A_i^c \end{cases} \\
=&1 + \left([1-\delta_2^2]^{\lambda} - 1\right) I(x_n \in A_i).
\end{aligned} \tag{4.6.10}$$

similarly,

$$\begin{aligned}&\int \left[\frac{f_0(x_{n+1}|x_n)}{f_1(x_{n+1}|x_n)}\right]^{\lambda} f_1(x_{n+1}|x_n)dx_{n+1}\\ \leq& 1+\left([1-\delta_2^2]^{\lambda}-1\right) I(x_n\in A_i).\end{aligned} \tag{4.6.11}$$

Let $\beta = b_1+1+\left([1-\delta_2^2]^{2\lambda}-1\right)\zeta = 1+\left([1-\delta_2^2]^{2\lambda}-1\right)\zeta/2 < 1$. We get by equation (4.6.10) that

$$\begin{aligned}&\iint \left[\frac{f_1(x_{n+1}|x_n)}{f_0(x_{n+1}|x_n)}\right]^{\lambda} f_0(x_{n+1}|x_n)f_0(x_n|x_{n-1})dx_{n+1}dx_n\\ \leq& 1+\left([1-\delta_2^2]^{\lambda}-1\right)\int I(x_n\in A_i)f_0(x_n|x_{n-1})dx_n\\ \leq& 1+\left([1-\delta_2^2]^{\lambda}-1\right)\zeta\\ \leq&\beta\end{aligned} \tag{4.6.12}$$

For any $f\in V_i$, by Proposition 4.6.5 and equation (4.6.11),

$$\begin{aligned}&\int \left[\frac{f_0(x_{n+1}|x_n)}{f_1(x_{n+1}|x_n)}\right]^{\lambda} f(x_{n+1}|x_n)dx_{n+1}\\ =&\int \left[\frac{f_0(x_{n+1}|x_n)}{f_1(x_{n+1}|x_n)}\right]^{\lambda} [f(x_{n+1}|x_n)-f_1(x_{n+1}|x_n)]dx_{n+1}+\int \left[\frac{f_0(x_{n+1}|x_n)}{f_1(x_{n+1}|x_n)}\right]^{\lambda} f_1(x_{n+1}|x_n)dx_{n+1}\\ \leq& b_1+1+\left([1-\delta_2^2]^{\lambda}-1\right) I(x_n\in A_i).\end{aligned}$$

So

$$\begin{aligned}&\iint \left[\frac{f_0(x_{n+1}|x_n)}{f_1(x_{n+1}|x_n)}\right]^{\lambda} f(x_{n+1}|x_n)f(x_n|x_{n-1})dx_{n+1}dx_n\\ \leq& b_1+1+\left([1-\delta_2^2]^{\lambda}-1\right)\int I(x_n\in A_i)f(x_n|x_{n-1})dx_n\\ \leq& b_1+1+\left([1-\delta_2^2]^{\lambda}-1\right)\zeta\\ \leq&\beta.\end{aligned} \tag{4.6.13}$$

To compute the type I error of the sequence of tests, let $n = 2k$ or $2k+1$

$$
\begin{aligned}
&\mathrm{E}_{f_0}(\phi_n) \\
&= \mathrm{E}_{f_0}(\phi_{2k}) \\
&= \Pr_{f_0}(h(k) > 0) \\
&\le \mathrm{E}_{f_0}\{\exp[\lambda h(k)]\} \qquad \text{(by Markov inequality)} \\
&= \mathrm{E}_{f_0}\left\{\exp[\lambda h(k-1)] \iint \left[\frac{f_1(x_{2k}|x_{2k-1})}{f_0(x_{2k}|x_{2k-1})}\right]^{\lambda} f_0(x_{2k}|x_{2k-1}) f_0(x_{2k-1}|x_{2k-2}) dx_{2k} dx_{2k-1}\right\} \\
&\le \beta \mathrm{E}_{f_0}\{\exp[\lambda h(k-1)]\} \qquad \text{(by equation (4.6.12))} \\
&\le \beta^k \qquad \text{(recursively using equation (4.6.12))} \\
&\le \beta^{n/2} \\
&= \exp\left[-n \log(1/\beta)/2\right].
\end{aligned}
$$

The type II error of the sequence of tests is bounded by

$$
\begin{aligned}
&\sup_{f \in V_i} \mathrm{E}_f(1 - \phi_n) \\
&= \sup_{f \in V_i} \mathrm{E}_f(1 - \phi_{2k}) \\
&= \sup_{f \in V_i} \Pr_f(-h(k) > 0) \\
&\le \sup_{f \in V_i} \mathrm{E}_f\{\exp[-\lambda h(k)]\} \qquad \text{by Markov Inequality} \\
&= \sup_{f \in V_i} \mathrm{E}_f\left\{\exp[-\lambda h(k-1)] \iint \left[\frac{f_0(x_{2k}|x_{2k-1})}{f_1(x_{2k}|x_{2k-1})}\right]^{\lambda} f(x_{2k}|x_{2k-1}) f(x_{2k-1}|x_{2k-2}) dx_{2k} dx_{2k-1}\right\} \\
&\le \beta \sup_{f \in V_i} \mathrm{E}_f\{\exp[-\lambda h(k-1)]\} \qquad \text{(by equation (4.6.13))} \\
&\le \beta^k \qquad \text{(recursively using equation (4.6.13))} \\
&\le \beta^{n/2} \\
&= \exp[-n \log(1/\beta)/2].
\end{aligned}
$$

Since both of the type I and II error are bounded by $\exp[-n\log(1/\beta)/2]$, ϕ_n are a sequence of exponentially consistent tests. □

Lemma 4.6.5. *Under the condition of Theorem* 4.6.1*, there exist finite nets* $V_1, \ldots, V_k$ *covering the set* $V = \{f_P : \sup_x ||f_P - f_0|| > \delta\}$ *and for each* $i = 1, \ldots, k$*, there exists a sequence of tests which is uniformly exponentially consistent for testing*

$$H_0 : f = f_{P_0} \qquad vs \qquad H_1 : f_P \in V_i.$$

Proof. We partition V twice. The first partition $V_1, \ldots, V_m$ is made in order to find intervals $A_1, \ldots, A_m$ satisfying

$$\min_{x_n \in A_i, f_P \in V_i} ||f_0(\cdot|x_n) - f_P(\cdot|x_n)|| \geq \frac{\delta}{2} \text{ for } i = 1, \ldots, m.$$

Then refine the partition of each V_i in order to find uniformly exponentially consistent tests.

By Lemma 4.6.2, the space $F_{\underline{\theta},\overline{\theta}}$ is compact. the cardinality of $\delta/8$-net is finite, so we can find a finite disjoint partition $\{V_1, \ldots, V_m\}$ of V such that for any two transition density f_1,f_2 in the same V_i $(i = 1, \ldots, m)$,

$$\sup_x ||f_1(y|x) - f_2(y|x)|| \leq \frac{\delta}{4}. \tag{4.6.14}$$

We arbitrarily pick up a model $f_i = f_{P_i}$ from each V_i. Since $\sup_x ||f_0(\cdot|x) - f_i(\cdot|x)|| > \delta$ and $||f_0(\cdot|x) - f_i(\cdot|x)||$ is a continuous function of x, so we can find intervals A_i satisfying

$$\min_{x \in A_i} ||f_0(\cdot|x) - f_i(\cdot|x)|| \geq \frac{3\delta}{4} \text{ for } i = 1, \ldots, m$$

By equation (4.6.14),

$$||f_0(\cdot|x) - f_{\sigma,P}(\cdot|x)|| \geq \frac{3\delta}{4} - \frac{\delta}{4} = \frac{\delta}{2}$$

for all $f_{\sigma,P} \in V_i$ and $x \in A_i$ and $i = 1, \ldots, m$.

Suppose that $|v + \varphi/[1 + \exp(-\theta(x - u))]| \leq |\overline{v}| + |\overline{\varphi}| \leq a$. By proposition 4.6.4,

$$\inf_{\forall x_n,\, P} f_P(x_{n+1} \in A_i | x_n) \geq d_i$$

where $d_i = \frac{1}{\sqrt{2\pi}\overline{\sigma}} \int_{A_i} \exp\left[-\frac{(|y|+a)^2}{2\underline{\sigma}^2}\right] dy$ for $i = 1, \ldots, m$.
Let $\zeta = \min\{d_1, \cdots, d_m\}$, (Obviously $\zeta > 0$[**]), $\delta_2 = \delta/4$, $\lambda = \underline{\sigma}^2/(2\overline{\sigma}^2)$, and $b_1 = \left(1 - [1 - \delta_2^2]^\lambda\right)\zeta/2$. Now we come to discuss the finer partition of $V_1, \ldots, V_m$. Given b_1, let η be the value got as in Proposition 4.6.5 and $\eta_1 = \min\{\eta, \delta/8\}/2$. Again, due to the compactness of the space $F_{\underline{\theta},\overline{\theta}}$, the cardinality of η_1-net is finite, so we can find a finer finite disjoint partitions $V_{ij}, j = 1, \ldots, i_{m_i}, i = 1, \ldots, m$ that satisfy

$$V_i = V_{i1} + \cdots + V_{i,k_i} \; \forall i = 1, \ldots, m,$$
$$\sup_x ||f_1(\cdot|x) - f_2(\cdot|x)|| \leq 2\eta_1 \leq \eta \quad \forall f_1, f_2 \in V_{ij}, \quad \forall i, j. \tag{4.6.15}$$

By Lemma 4.6.4, we can find sequences of uniformly exponentially consistent test for $H_0 : f = f_0 \quad vs \quad \mathrm{H_{ij}} : \mathrm{f} \in \mathrm{V_{ij}}$ for any $j = 1, \ldots, i_{m_i}, i = 1, \ldots, m$. □

[**]From here, we see compactness is the key factor.

Chapter 5

Numerical Examples

This chapter evaluates the performance of the DPM model through analyzing simulation data and real data. In the DPM model, the data are modeled by a mixture of kernel in which the kernel is known and the mixing distribution is unknown and could have any shape. The fundamental idea is to use the mixture of kernel to approximate the transition density. The advantage of the DPM models is that the data will automatically determine the shape of the mixing distribution through the Bayesian framework, and thus the complex model identification procedures could be avoided. How well could the true transition density be recovered by the DPM model? In Section 1, we simulate data from some known Markov processes and compare the Bayes estimates of the transition densities with the true ones. Simulation studies show that the DPM model recovers the transition densities well for models lying in the support of the prior as well as for many other popular models. In addition, the Bayes estimates from the Dirichlet mixture models are shown to be insensitive to the specification of the prior parameters over a wide range. In the Chapter 1, we argued that statistical models under i.i.d. assumption will fail when the data show evidence of dependence on the past. In Section 2, we will compare the prediction powers of the DPM models under the Markov and the i.i.d. assumptions under the criterion

of the predictive Mean square error (PMSE). In the final section, we analyze the famous Canadian lynx data using our general DPM models. For real data analysis, our interest is focused on prediction. The results from the general DPM models are very encouraging, compared to the literature.

5.1 Estimating the Transition Density

In this section, we illustrate the performance of the DPM model with simulated data by comparing the Bayes estimate of the transition density with the true one. The DPM model is given in equation (1.2.5) and the center measure is specified in equation (3.3.1). In Section 3.3, we gave the details of the "no gaps" MCMC algorithm. In our computation of the Bayes estimate of the transition density , we will use equation (3.4.2) while the first part of the equation can be evaluated by equation (3.4.3).

When implementing the "no gaps" algorithm, one needs to specify the parameters of the base measure of the Dirichlet process. The role of the precision parameter α in the Dirichlet process controls the extent how the true P_0 is close to the center measure G_0, if it exists. Usually, the precision parameter α is set to be small since the functional form of the true P_0 is unknown. Even if it is known, it tends to deviate from G_0 greatly. Since we expect that the true P_0 tends to differ much from G_0, the center measure G_0 is mainly used to be a reference measure in developing a Bayesian framework, and could be chosen according to mathematical convenience. In our simulation, we try some distribution G_0 different from that given in equation (3.3.1) such that we set the components $(\beta_0, \beta_1, \sigma)$ to have a joint normal-inverse-gamma distribution. We find that if both G_0 are roughly the same diffuse, they yield close Bayes estimate. However, in practice, the center measure G_0 shall be moderately informative in the sense that it reflects the order of the magnitude of the sample data roughly. As seen in equation (1.2.8), the posterior consists of the prior part and the

likelihood part. The likelihood part indicates the information from the data. If there is enough data, the data will swamp the prior and the specification of the prior does not affect the inference much. However, in application, typically the sample size is not big enough. If the center measure G_0 puts most of its mass around a very small area of the parameter space, the random P from the prior information will concentrate on this small area. The prior will then affect the posterior heavily and it will be difficult for the few data to update the shape of P properly. When specifying the constant parameter for the base measure G_0 in equation (3.3.1), we suggest that the mean u_0 of β_0 should reflect the mean of the data, the standard deviation $\sqrt{V_\beta}$ of β_0, β_1 should reflect the range of plausible values for the data centered around their mean. For example, if most of the data are between -2 and 2, the prior will be informative if one specifies $u_0 = 0$, $\sqrt{V_\beta} = 2$. But if one set $\sqrt{V_\beta} = 0.000001$ and $u_0 = 0$, the random P will put a lot of mass around 0 in its components β_0, β_1 in the prior information. The posterior of P will tend to concentrate around the small area also in small sample size problem so that the estimate of the transition density may deviate from the true. A diffuse prior plays little role in the posterior so that the inference is critically determined by the data even in a small sample size problem. However, we do not suggest a prior which is too diffuse either. Chow * (1998) gave the reason that the multinomial sampling of the "configuration" vector s would run into problem as the likelihood will stay near 0 for long time, invalidating the samples generated from the posterior. Our simulation experiments does confirm the fact. In fact, if the prior is chosen to be moderately informative, the result is quite insensitive to the specification of the constant parameter in equation (3.3.1) over a wide range.

In Theorem 3.2.4, we have showed that the "no gaps" algorithm will converge for

*In that paper, he used Muller and MaCeachern's (1997) model mentioned in Section 1.1 to analyze the overnight Hong Kong Interbank Offer Rates (HIBOR).

almost any initial starting value. However, the theorem does not imply that the algorithm will certainly converge in reasonable steps. In the "no gap" algorithm, the number of the parameters increases linearly with the sample size. It usually takes more time for a high dimensional MCMC algorithm to converge. In above paragraph, we mention that a too diffuse prior may cause the likelihood gets locked at 0 for a long period of time which leads to improper samples. Furthermore, theoretical results may fail numerically. Tierney (1994) mentioned that rounding error arising in computations may introduce absorbing state which causes a non-convergence MCMC. Hence, before calculating our interested quantity based on the posterior sample, it is necessary to diagnose if the algorithm has reached convergence (approximately). Various methods for diagnosing the convergence of MCMC algorithm could be found in Brooks and Roberts (1997). In general, most of these methods are quite subjective and tell only that the algorithm has not reached convergence if it does not really. It is also difficult to implement these methods in high dimensional parameter problems. Instead, we use the informal graph method to judge if the "no gap" algorithm has converged by trying different starting values. If the algorithm reaches convergence, the posterior sample will show the same pattern whatever the initial values are. Fortunately, the "no gaps" algorithm will generally reach (approximate) convergence quickly in a few thousands of steps. In our numerical demonstration, when we use equation (3.4.2) and (3.4.7) to approximate the Bayes estimate, we set the $N = 15,000$ which indicate the "burn in" period (approximate convergence) has reached and pick up the posterior sample every $r = 10$ step, and set the posterior sample size $M = 5,000$. We set $r = 10$ in order to reduce the variance of the estimate. Also, even if the algorithm has not reached convergence at step $15,000$, the Bayes estimate will not affected much by the first few samples. Recall that we will check the convergence of the algorithm informally by the graphical method with different initial values in each example.

In the first example, 150 observations are sampled from a model which lies in the support of the prior.

$$X_{n+1} = \begin{cases} -2.5 - \frac{1}{1+\exp(-X_n+1)} + \varepsilon_{n+1} \text{ with probability } 0.3 \\ \frac{2}{1+\exp(-2X_n+1)} + 0.5\varepsilon_{n+1} \text{ with probability } 0.4 \\ 2.5 + \frac{1}{1+\exp(-X_n-1)} + 0.25\varepsilon_{n+1} \text{ with probability } 0.3 \end{cases} \tag{5.1.1}$$

where $X_0 = 3$ and $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \text{N}(0,1)$. The true transition density satisfies equation (1.2.6) with

$$P_0(\theta) = 0.3\delta_{(-2.5,-1,-1,1,1)}(\theta) + 0.4\delta_{(0,2,-1,2,0.5)}(\theta) + 0.3\delta_{(2.5,1,1,1,0.25)}(\theta)$$

and δ_ϕ is a degenerate distribution put mass 1 on ϕ. To complete the specification of the base measure in equation (3.3.1), we set

$$u_0 = 0,\ V_\beta = 3^2,\ V_{\gamma_1} = 5^2, V_{\gamma_2} = 5^2,\ a = b = 0.01,\ \alpha = 1.$$

The Bayes estimate of the transition density and the true transition density are presented in Figure 5.1.

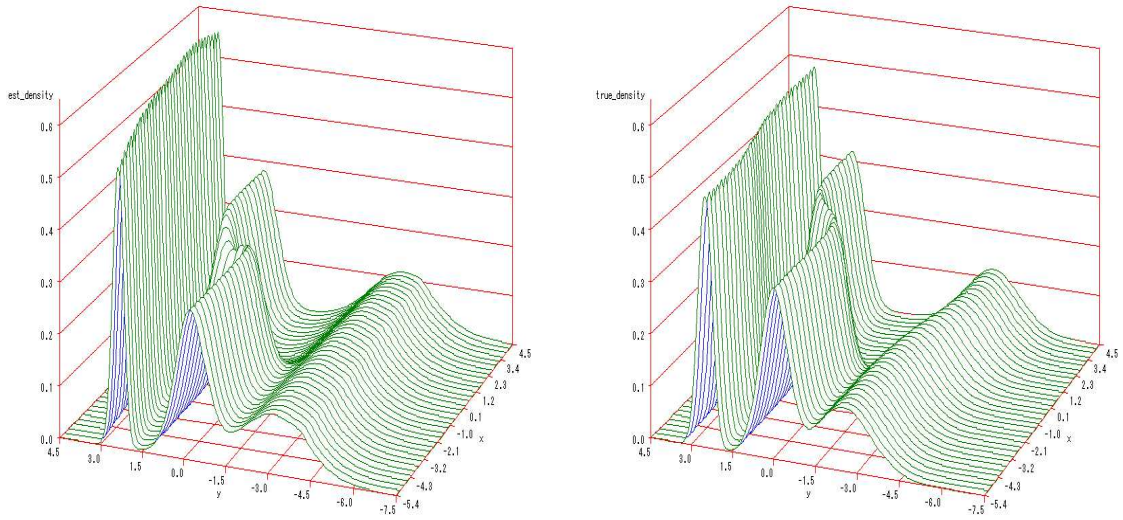
In the second example, 200 samples are simulated from a mixture of AR(1) as follows,

$$X_{n+1} = \begin{cases} 1.5 - 0.5X_n + \varepsilon_{n+1} \text{ with probability } 0.6 \\ -1.5 + 0.5X_n + 0.5\varepsilon_{n+1} \text{ with probability } 0.3 \end{cases} \tag{5.1.2}$$

where $X_0 = 3$ and $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \text{N}(0,1)$. In Section 1.2, we have shows that any AR(1) model could be approximated by a switching-regression model in equation (1.2.1). Hence the above finite mixture of AR(1) could be approximated arbitrarily well by a finite mixture of switching-regression models arbitrarily closely which lie in the support of the prior. We specify the constant parameter in equation (3.3.1) as follows.

$$u_0 = 0,\ V_\beta = 3^2,\ V_{\gamma_1} = 5^2, V_{\gamma_2} = 5^2,\ a = b = 0.01.\alpha = 1.$$

Figure 5.1: Bayes transition density estimate *vs* the true in example one



(a) Bayes estiamte $\hat{f}(y|x)$ based on 150 sample

(b) True $f(y|x)$

The Bayes estimate of the transition density and the true transition density are presented in Figure 5.2.

In the third example, 50 samples are simulated from a smooth threshold autoregressive (STAR) model as follows,

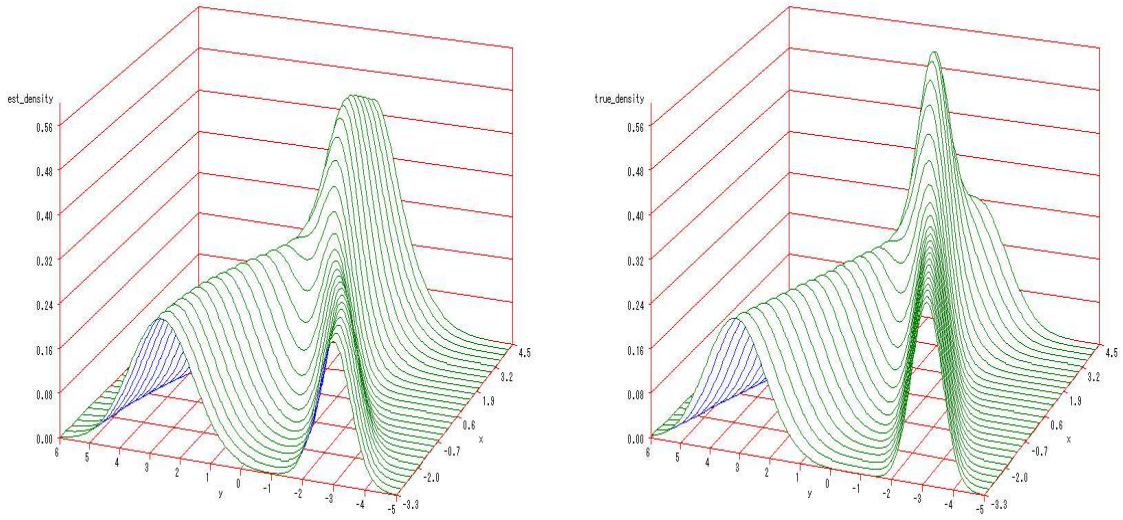
$$\begin{aligned} X_{n+1} =& (2 + 0.5X_n)\frac{1}{1+\exp(-1.5X_n)} \\ & - (1 + 0.5X_n)\frac{\exp(-1.5X_n)}{1+\exp(-1.5X_n)} + 0.5\varepsilon_{n+1} \end{aligned} \tag{5.1.3}$$

where $X_0 = 0$ and $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$. We specify the constant parameter in equation (3.3.1) as follows.

$$u_0 = 0,\ V_\beta = 2^2,\ V_{\gamma_1} = 5^2, V_{\gamma_2} = 5^2,\ a = b = 0.1.\alpha = 1.$$

The Bayes estimate of the transition density and the true transition density are presented in Figure 5.3.

Figure 5.2: Bayes transition density estimate *vs* the true in example two



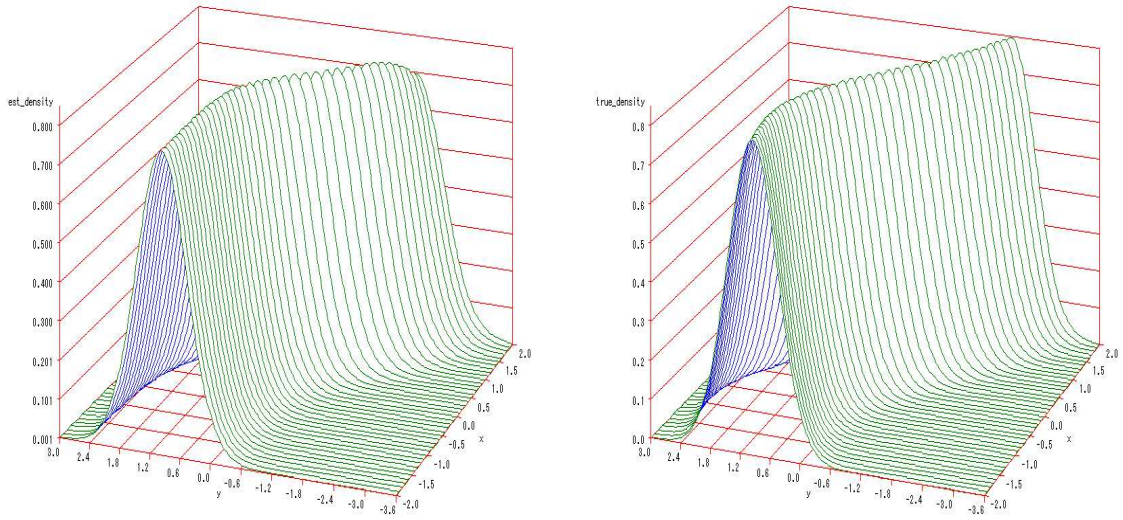
(a) Bayes estiamte $\hat{f}(y|x)$ based on 200 sample

(b) True $f(y|x)$

5.2 Comparing Dirichlet Mixture Model under the Markov and the i.i.d. Assumption

The i.i.d. assumption does not hold in many situations. Often, the observations show evidence of dependence on the past especially for time series data. it is expected that a model with i.i.d. assumption cannot capture the dependence structure among the data and naturally its performance will suffer. On the other hand, if Markov model is used to model the data which are i.i.d., its performance may not be much worse than that of the model with i.i.d. assumption since the latter is a special case of the former. We will demonstrate the fact by comparing the Dirichlet mixture model under Markov assumption and i.i.d. assumption. We will use the criterion of the prediction mean squared errors (PMSE). Suppose that there are n observations available. For any $l \leq i \leq n$ where n is the sample size and l is a given integer, we use the first

Figure 5.3: Bayes transition density estimate *vs* the true in example three



(a) Bayes estiamte $\hat{f}(y|x)$ based on 50 sample

(b) True $f(y|x)$

$i-1$ observations to predict the the next observation. The predictor is denoted by $\hat{u}_{i+1}$ for $l \leq i \leq n$. Then the prediction mean squred error (PMSE) is defined as

$$\text{PMSE} = \frac{1}{n-l} \sum_{i=l+1}^{n} (X_i - \hat{u}_i)^2. \tag{5.2.1}$$

we shall simulate two data sets from a known Markov process and a data set with i.i.d. observations. Both sample sizes are 250. We will evaluate the PMSE with the Dirichlet model under Markov assumption and i.i.d assumption. We set $l = 150$. the Bayes predictor is given (3.4.7) based on the posterior sample. Recall that under i.i.d. assumption, we only need to set $\beta_1 \equiv 0$ when implementing the "no gap" algorithm and using equation (3.4.7) to calculate the Bayes predictor.

The first data set is simulated from a AR(1) model as follows:

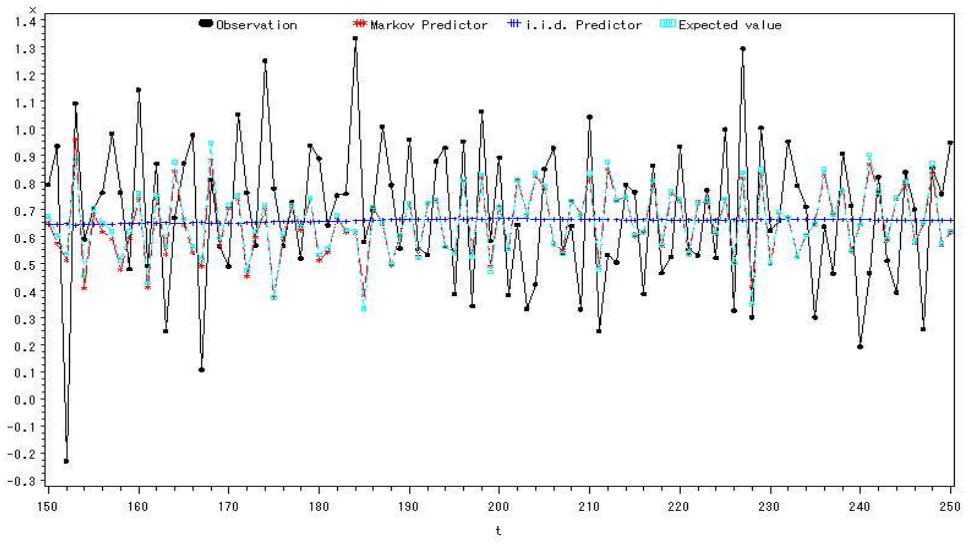
$$X_{n+1} = -0.5X_n + 0.25\varepsilon_{n+1} \tag{5.2.2}$$

We specify the constant parameter in equation (3.3.1) for this example as follows:

$$u_0 = 0.6,\ V_\beta = 0.5^2,\ V_{\gamma_1} = 3^2, V_{\gamma_2} = 3^2,\ a = b = 0.05, \alpha = 1.$$

Before starting the simulation, we find that at $N = 10000$, the "no gap" algorithm has approximately reached convergence. The posterior samples from step 10,001 to step 15,000 are used to calculated the Bayes predictors with equation (3.4.7). The observations, their expections conditioned on the past, and the corresponding Bayes predictors under both models between time 150–250 are plotted in Figure 5.4. The estimated PMSE for the Dirichlet mixture Markov model is 0.0649 and for the Dirichlet mixture model under i.i.d. assumption is 0.0731. The theoretically optimal value for the PMSE is equal to $\mathrm{Var}(0.25\varepsilon_i) = 0.0625$.

Figure 5.4: Predicting AR(1) with DPM under Markov and i.i.d. assumption



The second data set is simulated from a Threshold Autoregressive (TAR) model as follows:

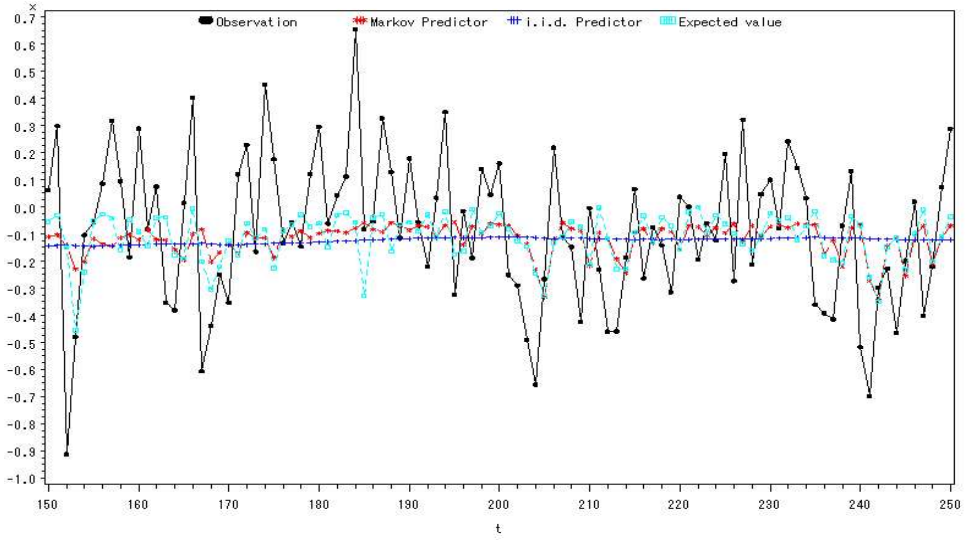
$$X_{n+1} = \begin{cases} -0.5X_n + 0.25\varepsilon_{n+1} \text{ when } X_n > 0 \\ 0.5X_n + 0.25\varepsilon_{n+1} \text{ else.} \end{cases} \tag{5.2.3}$$

We specify the constant parameter in equation (3.3.1) for this example as follows.

$$u_0 = 0,\ V_\beta = 0.5^2,\ V_{\gamma_1} = 3^2, V_{\gamma_2} = 3^2,\ a = b = 0.05, \alpha = 1.$$

Before starting the simulation, we find that at $N = 10000$ the "no gap" algorithm has approximately reached convergence. The posterior samples from step 10,001 to step 15,000 are used to calculated the Bayes predictors with equation (3.4.7). The observations, their expections conditioned on the past, and the corresponding Bayes predictors under both models between time 150–250 are plotted in Figure 5.5. The estimated PMSE for the Dirichlet mixture Markov model is 0.0696 and for the Dirichlet mixture model under i.i.d. assumption is 0.0768. The theoretically optimal value for the PMSE is equal to $\mathrm{Var}(0.25\varepsilon_i) = 0.0625$.

Figure 5.5: Predicting TAR(1) with DPM under Markov and i.i.d. assumption



The third data set composes of i.i.d. observations from the following model.

$$X_n = \beta_n + 0.1\varepsilon_n, \tag{5.2.4}$$

where $\beta_i \overset{\text{i.i.d.}}{\sim}$ from the gamma distribution $\mathcal{G}(0.1, 1)$, $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 1)$ and β_i's and ε_i's are independent. The constant parameter in equation (3.3.1) for this example is set

as follows:

$$u_0 = 0.1,\ V_\beta = 0.3^2,\ V_{\gamma_1} = 3^2, V_{\gamma_2} = 3^2,\ a = b = 0.05.\alpha = 1.$$

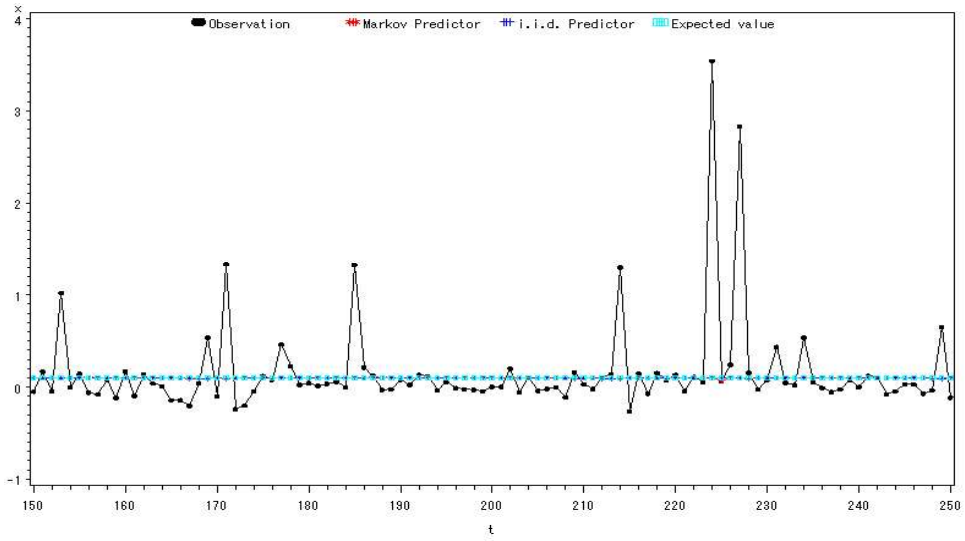
Again, in this example, the posterior samples from step 10,001 to step 15,000 are used to calculated the Bayes predictors with equation (3.4.7). The observations, their expections conditioned on the past, and the corresponding Bayes predictors under both models between time 150-250 are plotted in Figure 5.6. The estimated PMSE for the Dirichlet mixture Markov model is 0.271 and for the Dirichlet mixture model under i.i.d. assumption is 0.271. The theoretically optimal value for the PMSE is equal to $\mathrm{Var}(0.1\varepsilon_i) + \mathrm{Var}(\beta_i) = 0.11$. We shall note that the Bayes predictors from both models are close to the theoretically expected value 0.1 while the PMSE=0.271 from both models are much larger than the theoretically optimal PMSE 0.11. It is of interest to discuss this question. Simulation results show that it needs for more sample size from the the gamma distribution $\mathcal{G}(0.1, 1)$ for sample means and variance to be close to their expectation. For the specific simulation data, the sample mean is 0.15 and sample variance is 0.02693. So the predictors from the Dirichlet mixture models are much better than the predictor based on the sample mean.

5.3 Real Data Analysis

In this section, we analyze the famous Canadian lynx data. The lynx data set has been popular in the literature for testing new statistical methodology for time series analysis and is available in the R software. Our interest is focused on prediction. The result from the Dirichlet mixture models are very encouraging.

The Canadian lynx data set consists of annual Canadian lynx trappings around the Mackenzie River from 1821 to 1934 inclusive as recorded by the Hudson Bay company. It reflects to some extent the population size of the lynx in the Mackenzie

Figure 5.6: Predicting i.i.d. observations with DPM under Markov and i.i.d. assumption



River district. If the proportion of the number of lynx being caught to the population size remains approximately constant, after logarithmic transforms, the differences between the observed data and the population sizes remain approximately constant. Hence, it helps us to study the population dynamics of the ecological system in the system in that area. For further background about the data, refer to Tong (1983). Figure 5.7 depicts the time series plot of

$$X_t = \log_{10}(\text{number of lynx trapped in year } 1820 + t), t = 1, \ldots, 114.$$

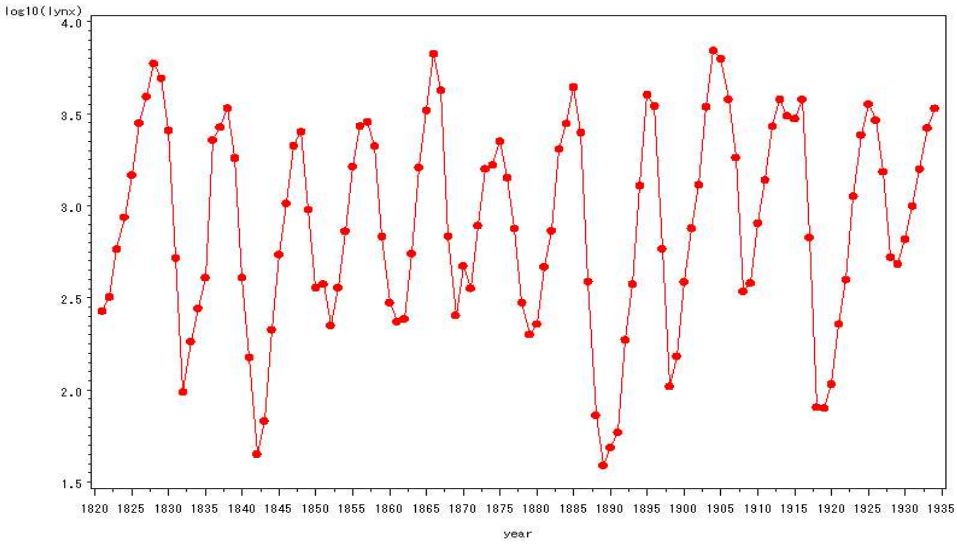
The periodic fluctuation displayed in this series has profoundly influenced the ecological theory (see below for discussion).

Model 5.3.1 was probably first time series model built for Canadian lynx data by Moran (1953).

$$X_t = 1.05 + 1.41X_{t-1} - 0.77X_{t-2} + \varepsilon_t, \tag{5.3.1}$$

where $\{\varepsilon_t\} \sim \text{IID}(0, 0.04591)$. Moran (1953) in his paper pointed out a "curious

Figure 5.7: Number of trapped lynx in the MacKenzie River district



feature" – the sum of squares of residuals corresponding to values of X_t greater than the mean is 1.781, where the sum of squares of residuals corresponding to values of X_t smaller than the mean is 4.007. The ratio the two sums is 2.250, which would be judged significant at the 1% level (F-test) against the null hypothesis that the two data sets of residuals are random samples from the same normal population.

The ARIMA models may be not proper for this data set. Please refer to Tong (page 8–31, 1983) for some arguments. Obviously, this series is not time-irreversible, which implies the existence of nonlinearity in lynx data. From Figure 5.7, we can see that there is an approximate ten-year cycle and that the ascent periods (around 6 years) tend to exceed the descent periods (around 4 years) by approximately 50%. Tong (1983) proposed to fit the data by TAR (threshold autoregressive) models. Model 5.3.2 is the simplest TAR model he fitted for the data.

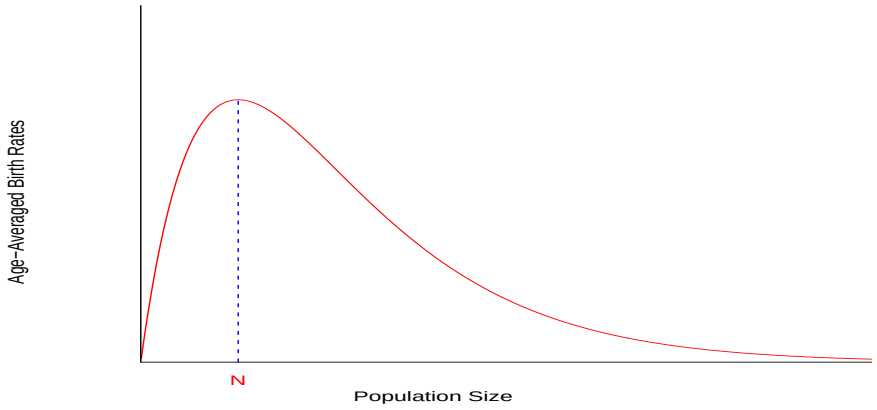
$$X_t = \begin{cases} 0.62 + 1.25X_{t-1} - 0.43X_{t-2} + \varepsilon_t \text{ if } X_{t-2} \leq 3.25, \\ 2.25 + 1.52X_{t-1} - 1.24X_{t-2} + \varepsilon'_t \text{ else }, \end{cases} \tag{5.3.2}$$

where $\{\varepsilon_t\}$ and $\{\varepsilon'_t\}$ are independent sequences or i.i.d. variables with $N(0, 0.0381)$ and $N(0, 0.0626)$ respectively. The TAR models admitted some nice ecological interpretation (Tong 1983). Some ecologists have hypothesized, on the basis of independent experiments, the following function form as a represntative of one standard type of birth curve:

$$\overline{b}(x) = b \exp(-x/k),$$

which is shown in Figure 5.8. The crucial feature of this type of birth curve is that when the adult population exceeds the critical point N, the competition for food so reduces the average adult fecundity that the birth rate of the entire population begins to fall off. The TAR model seems reasonable for analysis of the lynx data

Figure 5.8: A typical Birth Curve



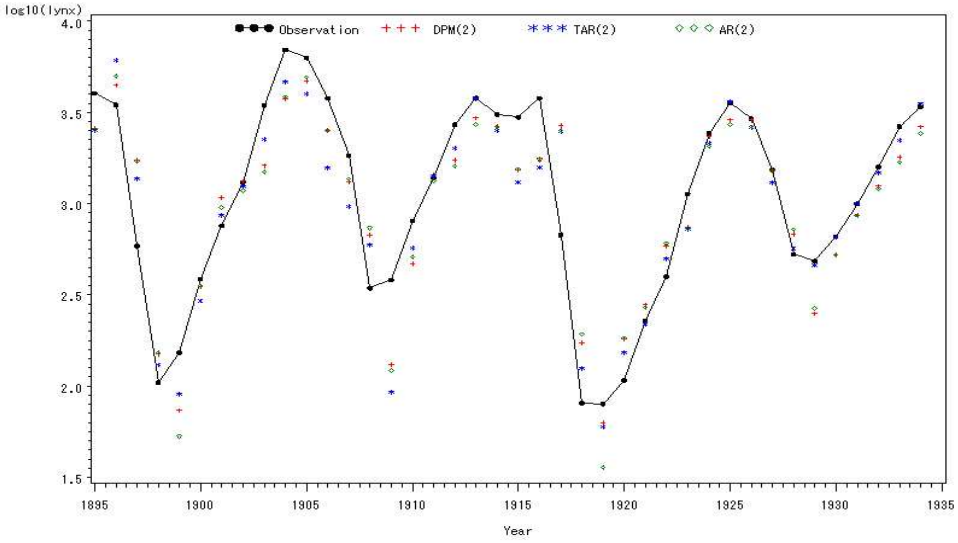
and provide a good fit of the data. There are some other models such as the STAR (smooth threshold autoregressive) models 5.3.3 fitted well for the data.

$$X_t = [0.62 + 1.25X_{t-1} - 0.43X_{t-2} + \varepsilon_t] \frac{1}{1 + \exp[\gamma(X_{t-2} - 3.25)]} + [2.25 + 1.52X_{t-1} - 1.24X_{t-2} + \varepsilon'_t] \frac{\exp[\gamma(X_{t-2} - 3.25)]}{1 + \exp[\gamma(X_{t-2} - 3.25)]}, \quad (5.3.3)$$

where γ is sufficiently large value, and ε_t's and ε'_t's are defined in equation (5.3.2). One key difference between models (5.3.2) and (5.3.3) is whether the coefficients should be smooth or radical in population density. This is of interpretation and belief. However, given the available data, the TAR and STAR models are statistically indistinguishable.

Here, we use the Dirichlet mixture models to predict the annual lynx numbers. We shall note that the conditional densities given by the Dirichlet mixture model and STAR model are smooth while that given by the TAR model is discontinuous at some critical points. Figure 5.9 plots the one-step predictions for the last 40 observations by AR(2), TAR(2) and the Dirichlet mixture models with two lagged terms. The PMSE is 0.0593 for AR(2), 0.0509 for Dirichlet mixture models and 0.0463 for TAR(2). The performance is better than AR(2) and a little worse than TAR(2), given the small sample size.

Figure 5.9: One-Step ahead Predictions of Trapped Lynx



Appendix

A Some Results on Markov Chains

This section gives a brief introduction to some concepts and results on the general state time-homogeneous Markov chain. Most of the material here is extracted from Meyn and Tweedie (1993) and Tierney (1994).

A.1. General Concepts

Let χ be a complete seperable space equipped with the Borel σ-algebra $\mathcal{B}(\chi)$. Usually χ will be a subset of $\mathbb{R}^k$. A "Markov transition kernel" on $(\chi, \mathcal{B}(\chi))$ is a map $P : \chi \times \mathcal{B}(\chi) \to [0, 1]$ such that:

(i) for any $B \in \mathcal{B}(\chi)$, the function $P(\cdot, B)$ is measurable,

(ii) for any $x \in \chi$, $P(x, \cdot)$ is a probability measure on $(\chi, \mathcal{B}(\chi))$.

A "time-homogeneous Markov chain" with the transion kernel $P(x, \cdot)$ is a sequence of χ-valued random values $\{X_n, n \geq 0\}$ such that for n and $A \in \mathcal{B}(\chi))$,

$$P_\mu(X_{n+1} \in A | X_0, \cdots, X_n) = P(X_n, X_{n+1} \in A),$$

where μ denote the initial distribution of X_0, and P_μ denotes the law of the overall process. In particular, P_x denotes the overall law if the chain starts at x. The critical aspect of a Markov chain is that it is forgetful of all but its most immediate past. This means that the future of the process is independent of the past given its present value. The overall law is completely determined given the initial distribution μ and the transition kernel P. For any n and any $A_i \in \mathcal{B}(\chi)\, i = 0, \cdots, n$,

$$P_\mu(X_0 \in A_0, \cdots, X_n \in A_n) = \int_{x_0 \in A_0} \cdots \int_{x_n \in A_n} \mu(dx_0) P(x_0, dx_1) \cdots P(x_{n-1}, dx_n)$$

Suppose that $P^n(x, \cdot)$ denotes the conditional probability of X_n given $X_0 = x$,

$$P^n(x, A) = P_u(X_n \in A | X_0 = x).$$

The celebrated "Chapman-Kolmogorov equation" states that for any $0 \leq m \leq n$,

$$P^n(x, A) = \int_\chi P^m(x, dy) P^{n-m}(y, A).$$

A measure π^* on $(\chi, \mathcal{B}(\chi))$ is said to be "invariant measure" for the chain if

$$\pi^*(A) = \int_\chi \pi^*(dx) P(x, A) \text{ for any } A \in \mathcal{B}(\chi).$$

A probability measure π is called "invariant (stationary) distribution" of the chain if

$$\pi(A) = \int_\chi \pi(dx) P(x, A) \text{ for any } A \in \mathcal{B}(\chi).$$

It means that if the initial distribution of X_0 is π, the distribution of any X_n is π since

$$\pi(A) = \int \pi(dw) P(w, A) = \cdots = \int \pi(dw) P^n(w, A) = P_\pi(X_n \in A).$$

The existence of invariant measure π^* does not imply the existence of the invariant probability measure unless π^* is finite. The invariant distribution π is an "equilibrium distribution" for the chain if for π-almost all x,

$$\lim_{n \to \infty} P^n(x, A) = \pi(A).$$

A chain with unique equilibrium distribution has good mathematical properties. However, the invariant distribution of a chain may not exist, or may not be unique.

Example A.1.1. Let χ be the real line $\mathbb{R}$, $P(x = 0, \cdot)$ is a degenerate measure puts mass 1 at 0, $P(x \neq 0, \cdot)$ is the standard Normal distribution. It is easy to verify that π_1, the measure degenerated at point 0, and π_2, the standard normal distribution are both invariant, equilibrium distribution of P.

"Irreducibilicity" ensures the uniqueness of the invariant distribution if it exists. In above example, the invariant distribution is not unique because the chain is not irrducible. A Markov process is called "φ-irreducible" if for a given measure φ on $\mathcal{B}(\chi)$, whenever $\varphi(A) > 0$, then for any x, there exists some integer n such that

$$P^n(x, A) > 0,$$

that is, the set A will be finally reached with positive probability starting from any point. Given the measure φ, we could always construct a measure ψ such that the chain is ψ-irreducible and for any A with $\psi(A) = 0$,

$$\psi\{\, y : \text{ The chain will finally reach } A \text{ staring from } y \text{ with positive probability } \} = 0.$$

Such a measure ψ is called a "maximal irreducibility measure" for the chain. All maximal irreducible measures are mutually absoultely continuous and dominate the irreducible measures. We denote by $\mathcal{B}^+(\chi)$ the sets of positive ψ-measure,

$$\mathcal{B}^+(\chi) = \{\, A \in \mathcal{B}(\chi) : \psi(A) > 0 \,\}.$$

The class $\mathcal{B}^+(\chi)$ is uniquely defined due to the equivalence of all maximal irreducibility measures. A chain is called "Recurrent" if it is irreducible and for any $A \in \mathcal{B}^+(\chi)$

$$U(x, A) = \sum_{n=1}^{\infty} P^n(x, A) = \infty \text{ for any } x \in \chi, \text{ and any } A \in \mathcal{B}^+(\chi).$$

A recurrent chain admits a unique invariant measure π^* (up to a constant). A chain is called "positive" if it is irreducible and admits an invariant distribution π. Recurrence does not imply positive unless π^* is finite. However positivity always implies recurrence. In a recurrent (positive) chain, the invariant measure π^* (the invariant distribution π) is a maximal irreducible measure. For this reason, sometimes it is convenient to denote the irreducibility property of a recurrent (positive) chain by π^* (π)- irreducible, and

$$A \in \mathcal{B}^+(\chi) \Longleftrightarrow \pi^*(A) > 0 \quad (\pi(A) > 0).$$

Recurrent is equivalent to the following conditions:
For any $A \in \mathcal{B}^+(\chi)$,

$$P_x(\{ \text{ The chain visits } A \text{ infinitely often } \}) > 0 \text{ for all } x \ ,$$

$$P_x(\{ \text{ The chain visits } A \text{ infinitely often } \}) = 1 \text{ for } \pi^*\text{-almost all } x \ .$$

The chain is said to be "Harris recurrent" if for any $A \in \mathcal{B}^+(\chi)$,

$$P_x(\{ \text{ the chain visits } A \text{ infinitely often } \}) = 1 \text{ for all } x \ .$$

Theorem A.1.1 (Theorem 9.1.5, Meyn and Tweedie (1993)). *For a recurrent chain with invariant measure π^*, the space χ could be decomposed into two disjoint sets H and N such that,*
(a) *N is null, $\pi^*(N) = 0$. Starting from N, there is positive probability for the chain to enter H, and positive probability for the chain to remain in N forever.*
(b) *H is "maximal absorbing" in the sense that $P(x, H) = 1$ for any $x \in H$ and $P(x, H) = 1$ implies $x \in H$. Starting from H, the chain will never enter N.*
(c) *The chain restricted to H is Harris recurrent.*

If a chain is Harris recurrent and positive, it is called "Harris positive". A Markov chain is "periodic" if there exists a measurable disjoint partition $B_0, B_1, \ldots, B_{d-1}$ of χ for some $d \geq 2$ such that for all $i = 0, \ldots, d-1$ and all $x \in B_i$, $P(x, B_j) = 1$ for $j = i + 1 \pmod{d}$. Otherwise, it is "aperiodic". A chain is called "ergodic" if it is positive Harris and aperiodic. A necessary and sufficient condition for a chain to be ergodic with invariant distribution π is,

$$\|P^n(x, \cdot) - \pi\| \to 0 \text{ for any } x,$$

where $\|\cdot\|$ denotes the total variation norm. Two stronger forms of ergodicity are called "geometric ergodicity" and "uniformly ergodicity", both ensuring that the

convergence happens at some rate. An ergodic chain is called geometrically ergodic if there exists some nonnegative, extended real-valued M with $\int |M| \, d\pi < \infty$ and a positive constant $r < 1$ such that

$$\|P^n(x, \cdot) - \pi\| \leq \mathbb{M}(x) r^n \text{ for any } x.$$

A chain is said to be "uniformly ergodic" if and only if

$$\sup_{x \in \chi} \|P^n(x, \cdot) - \pi\| \leq R r^n \text{ for some } R > 0, \, r < 1.$$

Above, we have considered a series of conditions in increasing order of strength: Recurrence → Positivity → Harris Positivity → Ergodicity → Geometric Ergodicity → Uniformly Ergodicity. In order to look at their implication we need some more notations. A function $h : \chi \to \mathbb{R}$ is called "harmonic" if for all $x \in \chi$,

$$\int P(x, dy) h(y) = h(x),$$

It is equivalent to saying that $\{h(x_i), i \geq 0\}$ is a martingale series for each initial condition. A set $C \in \mathcal{B}(\chi)$ is called a "small set" if there exists some m and measure v on $\mathcal{B}(\chi)$ such that for any $B \in \mathcal{B}(\chi)$,

$$\inf_{x \in C} P^m(x, B) \geq v(B) \text{ and } v(C) > 0.$$

In the general theory, small sets play similar roles to individual states in discrete chain theory. Given a function V on χ, denote by $P^n V(x) = P^n(x, V) = \int P^n(x, dy) V(y)$ and $\pi(V) = \int V \, d\pi$. The following theorem states the most important and fundamental result on ergodic chains.

Theorem A.1.2 (Theorem 13.0.1 and 17.1.7, Meyn and Tweedie (1993)). *The transition kernel P defines an ergodic chain with invariant distribution π if and only if*

$$\|P^n(x, \cdot) - \pi\| \to 0, \text{ for any } x.$$

Let P_π denote the overall law governing the ergodic chain with initial distribution $X_0 \sim \pi$. If the function $g(x_1, \ldots, x_k) : \chi^k \to \mathbb{R}$ is measurable and and P_π-integrable, then for any initial distribution of X_0,

$$\frac{1}{n}\sum_{i=1}^{n}[g(X_1, \ldots, X_k) + \cdots + g(X_n, \ldots, X_{n+k-1})] \to \int g dP_\pi = E_\pi[E(g(X_0, \ldots, X_{k-1})|X_0)] \text{ a.s..}$$

As noted in Theorem A.1.1, for a positive chain with invariant distribution π, the space χ could be decomposed into two parts H and N such that $\pi(N) = 0$ and the chain restricted to H is positive Harris and will never leave H once it enters H. These arguments together with Theorem A.1.2 leads to the following theorem.

Theorem A.1.3 (Theorem 1, Tierney (1994)). *Let P be a Markov transition kernel and π be one invariant distribution of P. Suppose that P is φ-irreducible (particularly π-irreducible), P is positive recurrent and π is the unique invariant distribution of P. If the chain is also aperiodic, then the following assertions hold.*
(a) *For π-almost all x,*

$$\|P^n(x, \cdot) - \pi\| \to 0,$$

(b) *If the function f is π-integrable, then for π-almost initial X_0,*

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \int f d\pi \text{ a.s.} \quad [\pi].$$

In an ergodic chain the chain will always converge irrespective of the initial state. In an aperiodic positive chain the converge are ensured except on a π-null set. If we could verify the aperiodic positive chain is Harris recurrent, then by definition, the chain is ergodic. Theorem A.1.4 and A.1.5 give some conditions for the chain to be Harris recurrent while Theorem A.1.5 connects the aperiodic positive chain and ergodic chain through the "drift condition".

Theorem A.1.4 (Theorem 1 and Corollary 1 and 1*, Tierney (1994), Theorem 17.1.5 Meyn and Tweedie (1993)). *For any bounded harmonic function h of an irreducible chain P with invariant distibution π, the following assertions hold*

(a) *If the chain is positive recurrent, $h = \int h d\pi$ a.s.*

(b) *The chain is Harris positive recurrent if and only if h is constant.*

(c) *If $P(x, \cdot)$ is absolutely continuous with respect to π for all x, P is Harris recurrent.*

(d) *Denote by $P_a^n(x, \cdot)$, the absolutely continuous part and by $P_s^n(x, \cdot)$, the singular part of the Lesbegue decomposition of $P(x, \cdot)$ with respect to π. If $P_a^n(x, \chi) \to 1$ for all x, then P is Harris recurrent.*

Theorem A.1.5 (Theorem 11.3.4 and 14.0.1, Meyn and Tweedie (1993)). *Suppose that the chain is irreducible and aperiodic, and $f \geq 1$ is a function on χ. Then* (a) *and* (b) *are equivalent:*

(a) *The chain is positive with invariant measure π, and $\int f \, d\pi < \infty$.*

(b) *There exist a function $V : \chi \to [0, \infty)$ which is finite at some point, small set C and $b < \infty$ such that for all $x \in \chi$,*

$$\triangle V(x) = \int P(x, dy)V(y) - V(x) \leq -f(x) + bI(x \in C).$$

Let $\|Q_1 - Q_2\|_f = \sup_{g \leq f} |\int g \, dQ_1 - \int g \, dQ_2|$ denote the f-norm distance between two measure. (Note the special case $f \equiv 1$). Let $S_v = \{ x : V(x) < \infty \}$. Then (a) *or* (b) *implies that $\pi(S_v) = 1$ and for any $x \in S_v$,*

$$P(x, S_v) = 1, \quad \textit{and} \quad \|P^n(x, \cdot) - \pi\| \leq \|P^n(x, \cdot) - \pi\|_f \to 0.$$

Moreover if $\pi(V) < \infty$, then exists $R < \infty$ such that for all $x \in S_v$,

$$\sum_{i=0}^{\infty} \|P^n(x, \cdot) - \pi\| \leq \sum_{i=0}^{\infty} \|P^n(x, \cdot) - \pi\| \leq R(V(x) + 1).$$

If $S_v = \chi$, that is, V is finite everywhere, the chain is positive Harris recurrent.

The condition in (b) of Theorem A.1.5 is called "drift condition". It means that $V(x_n)$ is expected to move monotonically downtowards at some uniform rate until the chain enters C. Drift condition could also be used to verify the geometric ergodicity of a chain.

Theorem A.1.6 (Theorem 15.0.1, Meyn and Tweedie (1993)). *For an irreducible and aperiodic chain, if there exists function $V : \chi \to [1, \infty)$ finite at some point, small set C, $b < \infty$ and $\beta > 0$ such that for all $x \in \chi$,*

$$\triangle V(x) = \int P(x, dy)V(y) - V(x) \leq -\beta V(x) + bI(x \in C),$$

then the chain is geometric ergodic.

Historically, one of the most significant conditions for uniformly ergodicity of Markov chains is the "Doeblin condition". If there exists a probability measure ϕ with the property that for any $0 < \varepsilon < 1$, there exists some m and $\delta > 0$,

$$\phi(A) > \varepsilon \Rightarrow \inf_{x \in \chi} P^m(x, A) \geq \delta.$$

then the Markov chain is said to satisfy Doeblin's condition.

Theorem A.1.7 (Theorem 16.0.2, Meyn and Tweedie (1993)). *For a ψ-irreducible chain, the following are equivalent:*

(a) *The chain is uniformly ergodic, there exists some R, $\rho < 1$ such that*

$$\sup_{x \in \chi} \|P^n(x, \cdot) - \pi\| \leq R\rho^n.$$

(b) *The chain is aperiodic and Doeblin's condition holds.*

(c) *There exist $\varepsilon > 0$ and a probability measure v such that for all $A \in \mathcal{B}(\chi)$,*

$$\inf_{x \in \chi} P^m(x, A) \geq \varepsilon v(A)$$

In particular, Condition (c) *implies a simple bound*

$$\sup_{x \in \chi} \|P^n(x, \cdot) - \pi\| \leq \rho^{\frac{n}{m}},$$

where $\rho = 1 - \varepsilon$.

A.2. Poisson Equation

Suppose that a Markov chain is positive, recurrent with transition probability P and invariant probability measure π. The "Poisson equation" associated with a π-integrable Borel function g is the equation

$$G - P(G) = g - \pi(g). \tag{A.1}$$

Due to the dependence structure of a Markov chain, it is somewhat difficult to study the statistical property of the sum $S_n(\overline{g}) = \sum_{i=1}^{n}(g(X_i) - \pi(g))$ directly. However, the series $S_n(\overline{g})$ can be written as

$$\begin{aligned} S_n(\overline{g}) &= \sum_{i=1}^{n}[G(X_i) - PG(X_i)] \\ &= \sum_{i=1}^{n}[G(X_i) - PG(X_{i-1})] + \sum_{i=1}^{n}[PG(X_{i-1}) - PG(X_i)] \\ &= M_n(g) + PG(X_0) - PG(X_n) \end{aligned} \tag{A.2}$$

where $M_n(g) = \sum_{i=1}^{n}[G(X_i) - PG(X_i)]$ is a martingale if a solution to the Poisson equation (A.1) exists. The martingale approach via solution to the Possion equation is introduced by Maigret (1978), Duflo (1997) and Meyn and Tweedie (1993) has been the key approach to (functional) central limit theorems and (functional) laws of the iterated logarithm for Markov chains. If $\sum_{k=0}^{\infty}|P^k(x,g) - \pi(g)|$ is finite everyehere, then

$$G(x) = \sum_{k=0}^{\infty}\{P^k(x,g) - \pi(g)\} \tag{A.3}$$

is a solution to the Poisson equation (up to an additive constant). This can be easily verified if we note that $PG(x) = \sum_{k=1}^{\infty}\{P^k(x,g) - \pi(g)\}$, so $G(x) - PG(x) = g - \pi(g)$. Also $\sum_{k=0}^{\infty}|P^k(x,g) - \pi(g)|$ is finite everyehere under the assumption of Theorem A.1.5 and A.1.7

Theorem A.2.8 (Theorem 17.4.2, Meyn and Tweedie (1993)). (a) *If the condition* (b) *in Theorem* A.1.5 *is satisfied for an everywhere finite function V, then for any $|g| \leq f$, equation* (A.3) *is a solution to the Poisson equation* (A.1) *which satisfies*

$$|G(x)| \leq \sum_{k=0}^{\infty} \|P^k(x,\cdot) - \pi\|_f \leq R(V(x)+1),$$

for some $R < \infty$.

(b) *Suppose that the chain is uniformly ergodic, that is, there exists some $\varepsilon > 0$ and a probability measure v such that for all $A \in \mathcal{B}(\chi)$,*

$$\inf_{x \in \chi} P^m(x, A) \geq \varepsilon v(A).$$

For any g bounded by the constant M, the solution (A.3) *to the Poisson equation is uniformly bounded by*

$$|G(x)| \leq M \sum_{k=0}^{\infty} \|P^k(x,\cdot) - \pi\| \leq \frac{M}{1-(1-\varepsilon)^{\frac{1}{m}}}.$$

B Some Techniques of Markov chain Monte Carlo

Let π be a probability measure on a measurable space $(\chi, \mathcal{B}(\chi))$. Suppose that we are interested in $\int f\, d\pi$ where f is a π-integrable measuarble function. If π is easily sampled, we could simulate i.i.d. sample $\theta_1, \ldots, \theta_n$ from π. By the strong law of large numeber, we could approximate $\int f\pi$ by $n^{-1}\sum_{i=1}^{n} f(\theta_i)$ for a large enough n. If the distribution function of π is too intractable to sample, another possible way is to construct a π-irreducible, aperiodic, time-homogeneous Markov chain $P(\theta, \cdot)$ with invariant distribution π. If we could sample the sequence $\{\theta_n, n \geq 1\}$ from the Markov model, Theorem A.1.2 and A.1.3) ensures that $n^{-1}\sum_{i=1}^{n} f(\theta_i)$ may still be used to approximate $\int f d\pi$. This leads to the idea of Markov chain Monto Carlo (MCMC) simulation.

The MCMC techniques are extremely useful in Bayesian Inference. The Bayes estimate of $f(\theta)$ under the squared error loss is the posterior conditional expectation of $f(\theta)$ given the data as follows,

$$\mathrm{E}(f(\theta)|X) = \int f\, d\pi(\theta|X).$$

The expression of $P(\theta|X)$, the posterior distribution of θ given the sample could be very complex. MCMC methods are generally used in which the posterior $P(\theta|X)$ is the stationary distribution of some Markov chain. The most popular MCMC algorithms are Gibbs sampling scheme, Metropolis-Hastings methods and their hybrids. Our review of MCMC algorithms and theories is based on Gamerman (1997), Tierney (1994) and others.

The Gibbs Sampler:

Gibbs sampling is a MCMC scheme where the transition kernel is formed by the full conditional distributions. Assume the distribution of interest is $\pi(\theta)$ where

$\theta = (\theta_1, \ldots, \theta_d)$. Each one of the components θ_i could be a scalar, vector or a matrix. Let $\theta_{-i} = (\theta_j; j \neq i), i = 1, \ldots, d$. Suppose that all the full conditional distributions $\pi(\theta_i|\theta_{-i})$, $i = 1, \ldots, d$ are available and we could easily sample from these distributions.

Gibbs Sampling Scheme:

1) Initialize the iteration counter of the chain to $j = 1$, and set initial values $\theta^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_d^{(0)}\right)$.
2) Obtain a new value $\theta^{(j)} = \left(\theta_1^{(j)}, \ldots, \theta_d^{(j)}\right)$ from $\theta^{(j-1)}$ through successive generation of values

$$\begin{aligned}
\theta_1^{(j)} &\sim \pi\left(\theta_1|\theta_2^{(j-1)}, \ldots, \theta_d^{(j-1)}\right) \\
\theta_2^{(j)} &\sim \pi\left(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_d^{(j-1)}\right) \\
&\vdots \\
\theta_d^{(j)} &\sim \pi\left(\theta_d|\theta_1^{(j)}, \ldots, \theta_{d-1}^{(j)}\right)
\end{aligned}$$

3) Change the counter j to $j + 1$ and return to step 2 until convergence is reached.

Obviously the Gibbs sampling scheme defines a time-homogeneous Markov chain since the values at j-th step depends on that at step $j - 1$ with the transition kernel

$$P\left(\theta^{(j-1)}, \theta^{(j)}\right) = \prod_{i=1}^{d} \pi\left(\theta_i^{(j)}|\theta_1^{(j)}, \ldots, \theta_1^{(j)}, \theta_{i+1}^{(j-1)}, \ldots, \theta_d^{(j-1)}\right).$$

Metropolis-Hastings Algorithm:

Assume that $\pi(\theta)$ has the density (still denoted by π) with respect to some finite measure ν. In Metropolis-Hastings sampler, a reversible Markov chains is formed in which the transition kernel density p with respect to ν satisfies

$$\pi(\theta)p(\theta, \phi) = \pi(\phi)p(\phi, \theta), \quad \text{for any} \quad \theta, \phi. \tag{B.1}$$

The above equation gives sufficient condition in order that $\pi(\theta)$ be the stationary distribution of the chain.

Let $E^+ = \{x : \pi(x) > 0\}$. Let $q(\theta, \phi)$ be an arbitrary Markov kernel density with respect to ν satisfying $Q(x, E^+) = 1$ for any $x \notin E^+$. The acceptance probability is defined as

$$\alpha(\theta, \phi) = \min \left\{ 1, \frac{\pi(\phi) q(\phi, \theta)}{\pi(\theta) q(\theta, \phi)} \right\}.$$

We note that the acceptance probability depends on the ratio $\pi(\phi)/\pi(\theta)$ only; the complete knowledge of π is not required. In particular, the proportionality constant is not needed. The kernel

$$p(\theta, \phi) = q(\theta, \phi)\alpha(\theta, \phi), \text{ if } \theta \neq \phi,$$

defines a density $p(\theta, \cdot)$ for every possible value of the parameter different from θ and satisfies the reversibility condition given in equation (B.1). There is a probability $r(\theta) = 1 - \int p(\theta, \phi)\, d\nu(\phi)$ for the chain to remain at θ. So the Metropolis kernel is defined as

$$P(\theta, d\phi) = p(\theta, \phi) d\upsilon(\phi) + r(\theta)\delta_\theta(d\phi).$$

Under the reversibility condition, π is the invariant distribution of the chain.

$$\pi(\phi) = \int \pi(d\theta)\, P(\theta, \phi).$$

Metropolis-Hastings Sampling scheme:

1) Initialize the iteration counter to $j = 1$ and set the initial values $\theta^{(0)}$.
2) Generate a ϕ from $q(\theta^{(j-1)}, \cdot)$ and U from the uniform distribution over $[0, 1]$. If $U \leq \alpha(\theta^{(j-1)}, \phi)$, the move is accepted and set $\theta^{(j)} = \phi$, otherwise $\theta^{(j)} = \theta^{(j-1)}$.
3) Change the counter j to $j + 1$ and return to step 2 until convergence is reached.

The following are two special cases of Metropolis-Hastings Algorithm:
(i): *Metropolis algorithm*: If $q(\theta, \phi) = q(\phi, \theta)$ for any ϕ, θ, the acceptance probability becomes

$$\alpha(\theta, \phi) = \min \left\{ 1, \frac{\pi(\phi)}{\pi(\theta)} \right\}.$$

(ii): "Independence chains": If $q(\theta, \phi) = f(\phi)$.

The Componentwise Metropolis-Hastings Algorithm:

In the Metropolis-Hastings algorithm, the θ vector is taken as a whole and updated in a single step. Similar to the Gibbs sampling, we could also divide the θ vector into d blocks $\theta_1, \ldots, \theta_d$ and update them one by one sequentially. Suppose $\pi(\theta_i|\theta_{-i})$ is dominated by v_i, $i = 1, \ldots, d$. If some of the components are continuous, the dominating measure v_i may be set to be the Lebesgue measure. If some of the components is discrete, the dominating measure v_i may be set to be the counting measure. For each of the components θ_i, we propose an arbitrary conditional transition kernel density $q_i(\theta_i, \phi_i)$ (with respect to v_i) which may well depend on the values of θ_{-i}. Consequently, the acceptance probability could be written as

$$\alpha_i(\theta_i, \phi_i) = \min \left\{ 1, \frac{\pi(\phi_i|\theta_{-i}) q_i(\phi_i, \theta_i)}{\pi(\theta_i|\theta_{-i}) q_i(\theta_i, \phi_i)} \right\}. \tag{B.2}$$

Similarly, the i-th component transition kernels is

$$P_i(\theta_i, d\phi_i) = q_i(\theta_i, \phi_i)\alpha_i(\theta_i, \phi_i) dv_i(\phi_i) + \left[1 - \int q_i(\theta_i, \phi_i)\alpha_i(\theta_i, \phi_i) dv_i(\phi_i) \right] \delta_{\theta_i}(d\phi_i), \tag{B.3}$$

which satisfies the reversibility condition and hence satisfies

$$\pi(\phi_i|\theta_{-i}) = \int \pi(\theta_i|\theta_{-i}) P_i(\theta_i, \phi_i).$$

Equivalently, we could define the i-th transition kernel as

$$P^{(i)}(\theta, d\phi) = P_i(\theta_i, d\phi_i)\delta_{\theta_{-i}}(d\phi_{-i}),$$

It is easy to prove that π be an invariant distribution of $P^{(i)}$, that is,

$$\pi(\phi) = \int \pi(\theta)P^{(i)}(\theta, \phi). \tag{B.4}$$

The componentise Metropolis-Hastings steps are as follows:

1) Initialize the iteration counter to $j = 1$ and set the initial values $\theta^{(0)}$.
2) Repeat the following sequentially from $i = 1$ to d.
   Generate a ϕ_i from $q_i(\theta_i^{(j-1)}, \cdot)$ and U_i from the uniform distribution over $[0, 1]$. If $U_i \leq \alpha_i(\theta_i^{(j-1)}, \phi_i)$, the move is accepted and set $\theta_i^{(j)} = \phi_i$, otherwise $\theta_i^{(j)} = \theta_i^{(j-1)}$.
3) Change the counter j to $j + 1$ and return to step 2 until convergence is reached.

Metropolis-within-Gibbs Algorithm:

In (componentwise) Metropolis-Hastings, the more similar is the proposed q (q_i) to π ($\pi(\theta_i|\theta_{-i})$), the closer to 1 will the acceptance probability α (α_i) be. This does not necessarily ensure fast convergence but may imply substantial computational savings.

In the ideal situation, if all $\pi(\theta_i|\theta_{-i})$ are easily sampled from, we may set $q_i = \pi(\theta_i|\theta_{-i})$. Then the acceptance probability $\alpha_i = 1$ for $i = 1, \ldots, d$. Hence the proposed θ_i is drawn from $\pi(\theta_i|\theta_{-i})$ and accepted with probability 1. This is just the Gibbs sampler!

But this situation that all full conditional distribution $\pi(\theta_i|\theta_{-i})$ could be easily sampled is rare in complex problems met in practice. A convenient way is to perform a Gibbs step by setting $q_i = \pi(\theta_i|\theta_{-i})$ if it is easily sampled and otherwise use

Metropolis-Hastings updating with a d-step cycle. This method is called *Metropolis-within-Gibbs algorithm.*

In practice, in order to use Theorem A.1.3, we have to verify its conditions:
(i) π is the invariant distribution of the designed Markov chain,
(ii) the chain is irreducible and aperiodic,
(iii) f is π integrable, that is, $\int |f|\, d\pi < \infty$.
The condition (i) is always satisfied if the components $\theta_1, \ldots, \theta_d$ are updated sequentially in a fixed order either by the Gibbs samper or the Metropolis -Hasting sampler. Note that by equation (B.4), each Metropolis-Hasting step (Gibbs step is a special case) produces a component kernel $P^{(i)}$, with invariant distribution π. Hence, after a cycle of updating all the d components, the cycle kernel P is given by

$$P(\theta, d\phi) = \int \cdots \int \prod_{i=1}^{d} P^{(i)}(\phi_{i-1}, d\phi_i),$$

where ϕ_i, $i = 1, \ldots, d$, are the intermediate states after updating d-th component, $\phi_0 = \theta$ is the starting point and the $\phi = \phi_d$ is the new state after a cycle. A use of the induction method leads to the conclusions that π be the invariant distribution of the cycle kernel P [Tierney 1994]. A sufficient condition for (ii) to hold is that each component transition kernel P_i in equation (B.3) is $\pi_i = \pi(\theta_i|\theta_{-i})$-irreducible and aperiodic.

C Some Other Useful Results

Theorem C.0.1. (Portmanteau) *The following are equivalent.*

1. $\{P_n\} \Rightarrow P$.

3. $\limsup P_n(F) \leq P(F)$ *for all* F *closed.*

4. $\liminf P_n(U) \geq P(U)$ *for all* U *open.*

5. $\lim P_n(B) = P(B)$ *for all* $B \in \mathcal{B}(\chi)$ *with* $P(\partial B) = 0$.

Lemma C.0.1 (Lyapounov's Inequality). *When* $0 \leq \alpha \leq \beta$,

$$\mathrm{E}^{1/\alpha}[|X|^{\alpha}] \leq \mathrm{E}^{1/\beta}[|X|^{\beta}]$$

Theorem C.0.2 (Azuma's Inequality†, Azuma (1967)). *Let* $X_1, X_2, \ldots$ *be a martingale difference sequence. If for each* i, $\alpha_i \leq X_i \leq \beta_i$ *and* $\gamma_i = \beta_i - \alpha_i$ *is a fixed constant (where* α_i, β_i *may depend on* $X_1, \ldots, X_{i-1}$*), then for any* n *and* $a > 0$,

$$\Pr\left(\sum_{i=1}^{n} X_n \geq na\right) \leq \exp\left[\frac{-2(na)^2}{\sum_{i=1}^{n}\gamma_i^2}\right],$$
$$\Pr\left(\sum_{i=1}^{n} X_n \leq -na\right) \leq \exp\left[\frac{-2(na)^2}{\sum_{i=1}^{n}\gamma_i^2}\right].$$

A collection of functions $\{f_\theta : \theta \in T\}$ is called "uniformly equicontinuous" if for each $\varepsilon > 0$, there exists $\delta > 0$ such that for any two $x, y \in \chi$ with $d(x, y) < \delta$,

$$\sup_{\theta \in T} |f_\theta(x) - f_\theta(y)| < \varepsilon.$$

The set of all bounded continuous on the locally compact and seperable metric space χ forms a metric space denoted by $\mathcal{C}(\chi)$, whose metric is given by

$$d(f_1, f_2) = \sup_{x \in \chi} |f_1(x) - f_2(x)|.$$

†When X_n's are independent, the inequality is called Hoeffeding's inequality

The set of functions

$$\{ f_\theta : \chi \to \mathbb{R} : \theta \in T, f_\theta \in \mathcal{C}(\chi) \},$$

is a subset of $\mathcal{C}(\chi)$. This subset is said to be "precompact" if for any sequence $\{\theta_i : i \geq 1, \theta_i \in T\}$, there exists a subsequence $\{\theta_{n_i}\}$ and a θ_0 (which is not necessary in T) such that

$$\sup_{x \in \chi} |f_{\theta_{n_i}}(x) - f_{\theta_0}(x)| \to 0 \text{ as } n_i \to \infty.$$

Theorem C.0.3 (Ascoli's Theorem). *Suppose that a topological space χ is compact. A collection of function $\{f_\theta : \chi \to \mathbb{R}, \theta \in T, f_\theta \in \mathcal{C}(\chi)\}$ is precompact if and only if both the following two conditions are satisfied:*

(a) *The set of functions is uniformly bounded, that is,*

$$\sup_{\theta \in T} \sup_{x \in \chi} |f_\theta(x)| < \infty,$$

(b) *The set of functions is equicontinuous.*

List of References

[1] Amewou-Atisso, M., Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (2003). Posterior consistency for semiparametric regression problems. *Bernoulli*.

[2] Antoniak, C. (1974). Mixtures of Dirichlet processes with application to Bayesian non-parametric problems. *Annals of Statistics* **2**: 1152-1174.

[3] Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tôhoku Mathematical Journal* **19(3)**: 357-367.

[4] Barron, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Unpublished manuscript.

[5] Barron, A. R. (1989). Uniformly powerful goodness of fit tests. *Annals of Statistics* **17**: 107-124.

[6] Barron, A. R., Schervish, M. and Wasserman, L. (1999). The consistency of posterior distributions in non parametric problems. *Annals of Statistics* **27**: 536-561.

[7] Bernstein, S. (1917). *Theory of Probability* (in Russian).

[8] Blackwell, D. (1973). Discreteness of Ferguson selections. *Annals of statistics* **1**: 356-358.

[9] Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Annals of statistics* **1**: 353-355.

[10] Blackwell, D. and Dubins, L. (1962). Merging of opinions with increasing information. *Annals of Mathematical Statistics* **33**: 882-886.

[11] Billingsley, P. (1968) *Convergence of Probability Measure.*

[12] Borwanker, G., Kallianpur, G. and Prakasa Rao, B. L. S. (1971). The Bernstein-von Mises theorem for Markov process, *Annals of Mathematical Statistics* **42**: 1241-1253.

[13] Bush, C. A. and MacEachern, S. N. (1996). A semi-parametric Bayesian model for randomized block designs. *Biometrika* **83**: 275-285.

[14] Dalal, S. R. (1978). A note on the adequacy of mixtures of Dirichlet processes *The Indian Journal of Statistics* **40**(A): 185-191.

[15] de Finetti, B. (1937). La prevision: see lois logiques, ses sources subjectives. *Ann. Instit. H. Poincare* **7**: 1-68.

[16] Diaconis, P. and Ylvisaker, D. (1985). Quantifying prior information. In: *Bayesian Statistics 2*, eds. J. M. Bernardo *et al.*, North-Holland, Amsterdam, 133-156.

[17] Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates (with discussion). *Annals of Statistics* **14**: 1-67.

[18] Diaconis, P. and Freedman, D. (1986). On inconsistent Bayes estimates of location. *Annals of Statistics* **14**: 68-87.

[19] Doskum, K. A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Annal of Probability* **2**: 183-201.

[20] Doss, H. J. (1985). Bayesian nonparametric estimation of the median; part I: Computation of the estimates *Annals of Statistics* **13**: 1432-1444.

[21] Doss, H. J. (1985). Bayesian nonparametric estimation of the median; part I: Asymptotic properties of the estimates *Annals of Statistics* **13**: 1445-1464.

[22] Duflo, M. (1997). *Random Iterative Models*. Springer.

[23] Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89**: 268-277.

[24] Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixture. *Journal of the American Statistical Association* **90**: 577-588.

[25] Escobar, M. D. and West, M. (1998). Computing Bayesian Nonparametric Hierarchical Models. In *Practical nonparametric and Semiparametric Bayesian Statistics, lecture Notes in Statistics* 133 (Dipak Dey *et al.*, eds,), Springer-Verlag, New-York.

[26] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**: 209-230.

[27] Ferguson, T. S. (1974). Prior distribution on the spaces of probability measures. *Annals of Statistics* **2**: 615-629.

[28] Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions, *Recent Advances in Statistics*, MH Rizvi, J. Rustagi & D. Siegmund, eds., Academic Press, New York, 287-302.

[29] Ferguson, T. S. and Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data , *Annal of Statistics* **1**: 163-186.

[30] Franses, P. H., and van Dijk, D. (2000). *Non-linear Time Series Models in Empirical Finance*. Cambridge university press.

[31] Freedman, D. (1963) On the asymptotic distribution of Bayes estimates in the Discrete case I. *Annals of Mathematical Statistics* **34**: 1386-1403.

[32] Gamerman, D. (1997). *Markov Chain Monte carlo–Stochastic Simulation for Bayesian Inference*, Chapman & Hall in statistical Science Series.

[33] Gaudard, M. and Hadwin, D. (1989). Sigma-algebras on spaces of probability measures. *Scand. J. Statist.* **16**: 169-175.

[34] Gelfand, A. E. (1999). Approaches for semiparametric Bayesian regression. In *Asymptotics, Nonparametrics and Time Series: A Tribu. to Madan Lal Puri* (Subir Ghosh, Ed.), Marcel Dekker, Inc. 615–638.

[35] Ghosal, S., Ghosh, J. K. and Samanta, T. (1995). On convergence of posterior distribution. *Annals of Statistics* **23**: 2145–2152.

[36] Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999a). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics* **27**: 143-158.

[37] Ghosal, S., Ghosh, J. K. and Ramamoorthi R. V. (1999b). Consistency issues in Bayesian nonparametrics. In *Asymptotics, Nonparametrics and Time Series: A Tribute to Madan Lal Puri* (Subir Ghosh, Ed.), Marcel Dekker, Inc. 639–667.

[38] Ghosal, S., Ghosh, J. K. and Ramamoorthi R. V. (1999). Consistent semiparametric Bayesian inference about a location parameter. *Journal of Statistical Planning and Inference* **77**: 181–193.

[39] Ghosal, S., Ghosh, J. K. and van der Vaart A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics* **28**: 500-531.

[40] Ghosal, S. and van der Vaart A. W. (2001), Entropies and rates of convergence for Bayes and maximum likelihood estimation for mixture of normal densities. *Annals of Statistics*, **29**: 1233–1263.

[41] Ghosh, J. K., Ghosal, S. and Samanta, T. (1994). Stability and convergence of the posterior in non-regular problems. In *Statistical Decision Theory and Related Fields V* (S. S. Gupta and J. O. Berger, eds), Springer-Verlay, 183-199.

[42] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of American Statistical Association* **58**: 13-30.

[43] Kim, Y. and Lee, J. (2001). On posterior consistency of survival models. *Annals of Statistics* **29**: 666-686.

[44] Laplace, P. S. (1774) Memoire sur la probabilite des causes par les evenements. Memoires de mathematique et de physique presentes a l'academie 'rougate des sciences, and dus dans ses assembies, 6:621-656. (Translated in *Statist. Sci.* **1**: 359-378.

[45] Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Annals of Statistics* **20**: 1222-1235.

[46] Lavine, M. (1994) More aspects of Polya tree distributions for statistical modeling. *Annals of Statistics* **22**: 1161-1176.

[47] Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related bayes estimates. *Univ. Calif. Publ. Statist.* **1**: 277-330.

[48] Le Cam, L. (1958). Les propertietes asymptotiques aes solutions de Bayes. *Publ. Inst. Statist. Univ. Paris* **7**: 17-35.

[49] Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Annals of Statistics* **1**: 38-53.

[50] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Annals of Statistics* **12**: 351-357.

[51] Maigret, N. (1978). Théorème de limite centrale pour une chaine de markov recurrenté Harris poitive. *Ann. inst. Henri Poincaré Ser B,* **14**: 425-440.

[52] MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation* **23**: 727-741.

[53] MacEachern, S. N. and Muller, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*, **7**: 223-228.

[54] Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*, Springer, New York.

[55] Moran, P.A.P. (1953). The statistical analysis of the Canadian lynx cycle, I: Structure and prediction. *Austrian Journal of Zoology,* **1**: 163-173.

[56] Muller, P., West, M. and MacEachern, S. (1997). Bayesian models for non-linear autoregressions. *Journal of Time Series Analysis* **18**: 593-614.

[57] Mauldin, R. D., Sudderth, W. D. and Williams, S. C. (1992). Polya trees and random distributions. *Annals of Statistics* **20**: 1203-1221.

[58] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics* **9**: 249-265.

[59] Prakasa Rao, B. L. S. (1978). Density Estimation for Markov processes using Delta-sequences, *Ann. Inst. Statist. Math.* **30**: 321-328.

[60] Roberts, G. O. and Smith, A. F. M. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-hastings algorithms. *Stochastic Processes and Their Applications* **49**: 207-216.

[61] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**: 639-650.

[62] Sethuraman, J. and Tiwari, R. (1982). Convergence of Dirichlet measures and interpretation of their parameters. In *Statistical Decision Theory and Related Topics III 2* (Gupta, S. S. and Berger, J. O., Eds), Academic Press, New York.

[63] Schwartz, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4**: 1–26.

[64] Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distribution. *Annals of Statistics* **29**: 687-714.

[65] Tierney, L. (1994). Markov chain for exploring posterior distributions *Annals of Statistics* **22**: 1701-1762.

[66] Tong, M. (1983). *Threshold Models in Non-linear Time Series Analysis*, Springer.

[67] von, Mises, R. (1931). Wahrscheinlich keitsrechnung. Springer, Berlin.