Universidad
Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Contributions to Bayesian nonparametrics

Autor:

YANYUN ZHAO

Director/es:

María Concepción Ausín

Michael Peter Wiper

DEPARTAMENTO DE ESTADÍSTICA

Leganés, May 2015

Universidad
Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Contributions to Bayesian nonparametrics

Autor: *YANYUN ZHAO*

Director/es: María Concepción Ausín & Michael Peter Wiper

Firma del Tribunal Calificador:

Firma

Presidente: (Nombre y apellidos)

Vocal: (Nombre y apellidos)

Secretario: (Nombre y apellidos)

Calificación:

Leganés, de de

Universidad Carlos III

PH.D. THESIS

Contributions to Bayesian nonparametrics

Author:

Yanyun Zhao

Advisor:

María Concepción Ausín & Michael Peter Wiper

DEPARTMENT OF STATISTICS

Leganés, Madrid, May 29th, 2015

In memory of my father

Acknowledgements

I would like to express my deepest gratitude to my supervisors Professor María Concepción Ausín and Professor Michael Peter Wiper for their consistent support, continued encouragement and infinite patience. During my years of Ph.D. study, they guide me from the rough idea towards the full paper for the second chapter of this dissertation, remain my central source of support at every darkness moment, steer me through difficult times and give me full freedom to select research topics for the third chapter in this thesis and especially provide quick and insightful feedback for any ideas I might have. They also have a significant impact on me with their great personality and generosity of spirit. I am extremely lucky to start my PhD under their supervision.

I am deeply indebted to Professor Bas Kleijn who arranged a short visit for me in University of Amsterdam, gave me an enthusiastic reception and suggested the topic for the fourth chapter. He was always willing to give insightful comments and prompt response for my draft proposal. Sometimes a detailed discussion in his office helped crystallize my views on the relevant topic.

I would also like to thank Jan van Waaij from University of Amsterdam for the extensive discussions about lectures on Bayesian nonparametrics taught by Professor Bas Kleijn. I am strongly inspired by his intuitive explanations for the asymptotic theory of posterior distribution and really admire his considerable expertise in dealing with highly difficult mathematical problems. Special thanks go to Evelien Wallet for her administrative support during this academic visit in University of Amsterdam.

My gratitude goes to Professor Jesús Gonzalo who helped me come to Spain and enter Universidad

Carlos III de Madrid and Professor Juan Pablo Rincón Zapatero who offered invaluable help and advice to sustain me through the tough moments of my postgraduate study.

I also thank Professor Qu for his friendship, kindness, extensive support and helpful suggestions at different stages of my academic pursuit. Thanks also go to Weining for his enthusiasm and delightful discussions from which I benefit a great deal. I am highly impressed and inspired by his persistence and constant optimism in research.

I'd owe a deep gratitude to Xiuping for her love, who accompanied me and shared all of my frustrations and struggles on this odyssey. She always reminded me to keep focused and insisted on my research, especially during the time I got stuck. Her full of optimism for the future inspires me to pursuit a career in academia.

Last but not the least, I would like to thank my family, especially my late father who set an example for me how to be a man of unbendable perseverance and unfaltering confidence in the face of difficulty, steadfast in adversity and unwavering loyalty to the family. I thank them for providing me the best quality of the life and the best education even then they financially struggled against the family budget constraints. I also thank my older brother's family for their constant support to me on this journey and particularly my daring little nephew Yaoyao and sweet niece Mengmeng in the other side of the globe who give me endless joy.

Resumen

Esta tesis se centra en las propiedades frecuentistas de los procedimientos Bayesianos dentro de un amplio espectro de modelos estadísticos de dimensión infinita a través de métodos no paramétricos Bayesianos. Se presentan tres ensayos sobre los aspectos asintóicos de la distribución a posteriori en varios modelos estadísticos introducidos en tres capítulos.

En el Capítulo 2, dentro de un contexto de estimación Bayesiana de densidades, se construye una distribución a priori del tipo Berstein-Dirichlet en el espacio de densidades multivariantes en hipercubo unidad y se obtiene la tasa de convengencia de la distribución a posteriori correspondiente. Se desarrolla un nuevo algoritmo de muestreo para este modelo basado en métodos de "slice sampling" y se ilustra mediante datos simulados y reales.

En el Capítulo 3, se considera un enfoque semiparamétrico Bayesiano para un modelo de regresión lineal con restricciones de momentos condicionales. La variable de error sigue una distribución Gaussiana cuya varianza depende de los predictores. Se desarrolla un procedimiento Bayesiano adaptativo en el que las distribuciones a priori sobre la función de desviación estándar condicional se construyen cuidadosamente.

El Capítulo 4 está dedicado a la cuestión de la tasa de convergencia de la distribución a posteriori en una amplia gama de modelos a priori. Teniendo en cuenta los problemas de estimación en la frontera del soporte donde ningún modelo a priori puede cumplir los criterios habituales para el análisis de la distribución a posteriori en muestras grandes, se desarrolla un nuevo criterio que permite la selección de modelos a priori flexibles para contrarrestar estos problemas con condiciones del modelo más fuertes, manteniendo además la propiedad de tasa óptima.

Abstract

This dissertation focuses on the frequentist properties of Bayesian procedures in a broad spectrum of infinite-dimensional statistical models via Bayesian nonparametric approaches. Three essays concern the asymptotic aspects of posterior distribution in various statistical models presented in the subsequent three chapters.

In the context of multivariate density estimation discussed in Chapter 2, a Bernstein-Dirichlet prior is constructed in the space of multivariate densities on hypercube and the corresponding posterior contraction rate is obtained. We implement this model through a novel sampling algorithm based on a slice sampling scheme for the simulated and real data.

In Chapter 3, we consider a Bayesian semiparametric approach for a linear regression model with conditional moment restrictions. The error variable follows a Gaussian distribution whose variance depends on the predictors. An adaptive Bayesian procedure is performed when the priors on the conditional standard deviation function are carefully constructed.

Chapter 4 is devoted to the issue of posterior convergence rate for a broad range of priors. Motivated by the boundary support estimation problems where any constructed prior could not meet the usual criteria for large sample analysis of the posterior distribution, we develop a new yardstick that allows flexible prior selections to counteract these problems by the stronger model conditions and meanwhile the rate optimality property is maintained.

Contents

List of Figures

Chapter 1

Introduction

A primary purpose of statistics is to make some inference on the unknown quantities related to a collection of measured data. This invites the assumption that the data is assumed to be generated from some unknown underlying probability distribution. Usually we employ statistical models or stochastic models to describe the random phenomenon and then analyze the data for the estimation of quantities of interest. The observed data is a measurement denoted by X, taking values in a sample space $(\mathcal{X}, \mathscr{X})$, where $\mathscr{X}$ is a Borel σ-algebra on a polish space $\mathcal{X}$.

More specifically, a statistical model $\mathscr{P}$ is a class of probability measures over the sample space $(\mathcal{X}, \mathscr{X})$. In most real problems, the collection of probability distributions is indexed by some quantity, which is universally called a parameter. Denote the parameter by θ and the parameter space, the class of all possible values of the parameter, by Θ. Given θ, the distribution of the observation X follows some distribution denoted by P_θ, or say $X|\theta \sim P_\theta$ in an abbreviated form. To put it briefly, the statistical model could be defined as,

$$\mathscr{P} = \{P_\theta : \theta \in \Theta\}.$$

The model $\mathscr{P}$ contains all the possible candidate distributions for the observation X, in which statisticians provides reasonable explanations for the uncertainty. As such, the data, that have been generated coinciding with some unknown distribution must be analyzed for statisticians to attempt to learn about this unknown probability distribution and make some type of inference about certain aspects of the distribution.

In many statistical problems, the distribution that generates the data is known except for the values of finite-dimensional parameters or infinite-dimensional components. Consequently, the statistical problem falls into two categories: namely parametric models and nonparametric models. When it comes to the inferential procedures, two main statistical schools have emerged along the way to treat the statistical problems in a different perspective during the history of statistics. These two schools of thought are known as frequentist (classical) and Bayesian respectively.

One of the underlying postulates in the frequetist paradigm is that the experimental data is generated in accordance with some probability distribution indexed by some unknown, fixed parameter θ_0. Frequentist statisticians does not permit any probability statement made about this true parameter θ_0. Classical inferential approaches have been well established to learn about this unknown true parameter, such as maximum likelihood estimation, hypotheses testing, confidence intervals and many other things. In this thesis, we employ the Bayesian nonparametric approach to the statistical estimation problems.

1.1 Bayesian nonparametrics inference

1.1.1 Priors, posterior and Bayes' rule

In the Bayesian paradigm, all unknown quantities are treated as random variables and a joint distribution for all of them must be specified for the subsequent statistical inference. To this aim, we need to formulate a so-called prior distribution Π for the parameter θ as well as the conditional distribution of the data X given the parameter θ that we denote by P_θ. Hence the joint distribution on the product space $\mathcal{X} \times \Theta$ is determined by the prior distribution and the conditional distribution P_θ. This is a uniform framework for parametric and nonparametric Bayesian statistics in which the unique difference lies completely in the dimensionality of the parameter. However, more attention should be paid to the nonparametric case since it is not easy to build a proper prior for infinite-dimensional parameters. In the following, we provide a broad overview of conditioning device and Bayes' rule.

Let the parameter space Θ be a polish space equipped with a Borel σ-algebra $\mathscr{B}$ and a triple $(\Theta, \mathscr{B}, \Pi)$ be a probability space. A regular conditional distribution P_θ on the sample space $(\mathcal{X}, \mathscr{X})$ is a Markov kernel, mapping from $(\Theta, \mathscr{B})$ into $(\mathcal{X}, \mathscr{X})$ satisfying,

(i) for each $\theta \in \Theta$, the mapping $A \to P_\theta(A)$ is a probability measure,

(ii) for each $A \in \mathscr{X}$, the mapping $\theta \to P_\theta(A)$ is measurable.

Then one could obtain a well-defined joint distribution for the pair (X, θ) on the product measurable space $(\mathcal{X} \times \Theta,\ \mathscr{X} \times \mathscr{B})$ as follows,

$$Pr(X \in A,\, \theta \in B) = \int_B P_\theta(A)\, d\Pi(\theta), \quad A \in \mathscr{X},\, B \in \mathscr{B}. \tag{1.1}$$

This gives rise to two types of distributions: namely the marginal distribution of the data and the so-called posterior distribution of the parameter.

Let B be equal to the full parameter set Θ, then the marginal distribution of the observation X is given by,

$$Pr(X \in A) = \int_\Theta P_\theta(A)\, d\Pi(\theta), \quad A \in \mathscr{X}. \tag{1.2}$$

Assume that our model $\mathscr{P}$ is dominated by some σ-finite measure μ, then it is possible to express the posterior distribution in terms of densities $p_\theta = \frac{dP_\theta}{d\mu}$. Bayes' rule then yields a version of the posterior distribution as follows,

$$\Pi\big(\,\theta \in B \mid X\big) = \frac{\displaystyle\int_B p_\theta(X)\, d\Pi(\theta)}{\displaystyle\int_\Theta p_\theta(X)\, d\Pi(\theta)}, \quad B \in \mathscr{B}. \tag{1.3}$$

Suppose that $X = (X_1, X_2 \ldots, X_n)$ and $X_1, X_2 \ldots, X_n$ is conditionally independent and identical distributed (hereafter abbreviated i.i.d) given θ, then the posterior distribution of the parameter θ given the data X is,

$$\Pi\big(\,\theta \in B \mid X_1, \ldots, X_n\big) = \frac{\displaystyle\int_B \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)}{\displaystyle\int_\Theta \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)}, \quad B \in \mathscr{B}. \tag{1.4}$$

The posterior distribution of θ is typically viewed as a data-unpdated version of the prior Π. In other words, the degrees of the belief for the unknown parameter can be altered by the observation X via the conditioning mechanism. The amount of the information about θ stems from two sources, namely the data and the prior. After observing the data, we could incorporate the observation into the knownledge of learning about the parameter. Here the changes in knowledge about the unknown quantity take place due to Bayes' theorem.

The posterior distribution provides all we need for the statistical inference. Estimation, credit regions and hypothesis testing, which consists of the Bayesian inferential problems, may then be carried out via the posterior distribution. For example, we could obtain a Bayesian point estimator of the parameter by using mean, median or mode of the posterior distribution. In particular, the posterior mean is justified as a minimizer over the model relative to the common squared error loss function. Other loss functions justify the use of the median or mode.

To apply the theory of Bayesian estimation, it is necessary to specify a prior for the parameter. For parametric models, there are many available choices to select a prior distribution, such as Beta, Gaussian and Gamma distributions, to name a few. Certain prior distributions in these finite-dimensional models enjoy the nice conjugate property. For instance, beta distribution is a conjugate prior distribution for samples drawn from a Bernoulli distribution. In this case, for all observational values, the posterior at each state of sampling belongs to the family of beta distribution.

However, the construction of a prior in infinite-dimensional models is more complicated. Generally speaking, there are two driving factors behind it. On the one hand, formulating a prior in infinite-dimensional cases always involves the topological aspects but infinite-dimensional space could support a great deal of diverse norm topologies so that it is difficult to accommodate a sufficiently large support for the prior. In contrast to this, there exists one unique norm topology in the finite-dimensional vector space where all norms are equivalent in a sense that they define the same topology. On the other hand, Lebesgue measure gives us a typical candidate measure that dominates a finite-dimensional model and then all the distributions in this setting could be expressed in terms of probability densities for convenient treatment. In an infinite-dimensional case, we could not find a default uniform measure to dominate the model.

1.1.2 Dirichlet process prior

In a seminar contribution Ferguson (1973) proposed a first nonparametric prior known as the Dirichlet process prior that has become a popular tool in the area of Bayesian nonparametrics. The Dirichlet process is thought of as the "normal" distribution of Bayesian nonparametrics as well as a basic building block for priors in other infinite-dimensional objects. Also, the Dirichlet process prior has become the default prior in the space of probability distributions. Prior to the demonstration of the Dirichlet process, we begin with the introduction of the Dirichlet distribution.

DEFINITION 1.1 (Dirichlet distribution) Let $Q = (Q_1, Q_2, \ldots, Q_k)$ be a random vector such that $\sum_{i=1}^{k} Q_i = 1$ and $Q_i \geq 0$ for $i = 1, 2, \ldots, k$. The k-dimensional random vector Q is said to possess the Dirichlet distribution with parameter $\alpha_1, \alpha_2, \ldots, \alpha_k > 0$ which we denote by $Q \sim Dir(\alpha_1, \alpha_2, \ldots, \alpha_k)$, if it has the following density,

$$\frac{\Gamma(\alpha_1 + \alpha_2 + \ldots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \ldots x_{k-1}^{\alpha_{k-1} - 1} x_k^{\alpha_k - 1},$$

where $x_1, \ldots, x_k > 0$ and $\sum_{i=1}^{k} x_i = 1$.

The Dirichlet distribution is actually the multivariate generalization of the beta distribution.

DEFINITION 1.2 (Dirichlet Process, Ferguson (1973)) A random probability measure P on the sample space $(\mathcal{X}, \mathscr{X})$ is said to follow a Dirichlet process $\mathcal{DP}(M, G_0)$ with a base distribution G_0

on $(\mathcal{X}, \mathscr{X})$ and concentration parameter M, if for each finite measurable partition $A_1, A_2, \ldots, A_k$ of $\mathcal{X}$,

$$(P(A_1), P(A_2), \ldots, P(A_k)) \sim Dir(MG_0(A_1), MG_0(A_2), \ldots, MG_0(A_k)).$$

The Dirichlet process can be widely described as an infinite-dimensional generalization of the Dirichlet distribution. It is a distribution over the space of all probability measures such that any marginal distribution on the finite partition possesses a Dirichlet distribution. These two parameters M, G_0 in the definition of Dirichlet Process play intuitive roles. Roughly speaking, the base distribution can be understood as the expectation of the Dirichlet process. That is to say, $E(P(A)) = G_0(A)$ for each measurable set $A \subset \mathcal{X}$. In addition, we could treat the concentration parameter M as a measure of precision of the Dirichlet process. Actually, $Var(P(A)) = G_0(A)(1 - G_0(A))/(M+1)$. More mass of the prior will be centered around the mean as the concentration parameter M is larger.

It has been shown that the posterior of the Dirichlet process prior maintains the conjugate property just as the finite-dimensional Dirichlet distribution. In other words, if $X_1, X_2, \ldots, X_n | P \sim P$ and $P \sim \mathcal{DP}(M, G_0)$, then the posterior distribution of P is given as follows,

$$P|X_1, X_2, \ldots, X_n \sim \mathcal{DP}\left(M+n, \frac{M}{M+n}G_0 + \frac{\sum_{i=1}^{n} \delta_{X_i}}{M+n}\right) \tag{1.5}$$

where δ_{X_i} is the Dirac function at X_i for $i = 1, 2, \ldots, n$.

1.1.3 Construction of a Dirichlet process

Dirichlet process could be constructed through several methods. In this Section, we describe the generalized Pólya urn scheme and stick-breaking representation that will be discussed in more details in Chapter 2.

One construction of a Dirichlet process is related to the Pólya urn scheme. Consider an i.i.d sequence $X_1, X_2, \ldots$ generated from some distribution P that possesses a Dirichlet process prior. Its associated sequence of predictive distribution, such as, $X_1, X_2|X_1, X_3|X_1, X_2$, can be used as a convenient tool to demonstrate the structure of the Dirichlet process. Using the conjugate property of the Dirichlet process, it can be easily shown that,

$$X_{N+1}|X_1, X_2, \ldots, X_N \sim \begin{cases} \delta_{X_1} & \text{with probability } \frac{1}{M+N}, \\ \delta_{X_2} & \text{with probability } \frac{1}{M+N}, \\ \vdots & \qquad \vdots \\ \delta_{X_N} & \text{with probability } \frac{1}{M+N}, \\ G_0 & \text{with probability } \frac{M}{M+N}. \end{cases} \tag{1.6}$$

This construction can be regarded as the continuous analog of the well-known Pólya urn scheme. More explanations and demonstration about it could be found in more details in Ghosh and Ramamoorthi (2003b).

Sethuraman (1994) described the Dirichlet process in another constructive way and proposed a so-called stick-breaking construction to generate a sample from Dirichlet process. Suppose an infinite sequence of weights $\{\pi_k\}_{k=1}^{\infty}$ is generated in accordance with the procedures given as follows,

$$\beta_k \sim beta(1, M),$$
$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)$$

Then it can be shown that the following random probability distribution G has a Dirichlet process $\mathcal{DP}(M, G_0)$,

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}, \quad \theta_k^* \sim G_0 \tag{1.7}$$

his construction makes clear the fact that all the random distributions drawn from the Dirichlet process are discrete. The mechanism can be understood as the successive breaking of a stick of length 1. We first break a random portion β of the stick and assign the weight π_1 to this part and recursively break the remaining stick in random to obtain π_2, π_3 and so on. This simple and neat description of Dirichlet process priors facilitates a wide range of extensions and novel implementation procedures for the Dirichlet process.

When we want to estimate a density function in a Bayesian paradigm, Dirichlet process could not be directly chosen as a desired prior in this case since it produces discrete probability distributions exclusively. This can be remedied by a modification of Dirichlet process. A prior on densities could be induced by assigning a Dirichlet process prior on the mixture distribution F. Antoniak (1974) introduced Dirichlet process mixtures as the powerful Bayesian model for density estimation. Given a class of kernels $\{\psi(\cdot, \mu) : \mu \in \tilde{\Theta}\}$ indexed by the parameters, a mixture density is defined by,

$$p_F(x) = \int \psi(x, \mu)\, dF(\psi), \tag{1.8}$$

where F is a probability distribution on $\tilde{\Theta}$.

The underlying idea springs from the facts that smooth densities could be obtained by convoluting a probability distribution with kernels in a frequentist perspective. The kernel ψ may depend on an additional parameter ϕ. This could induce a rich family of kernel mixture priors, for example, the location-scale mixture prior and the Dirichlet Bernstein prior to name a few. The latter has been used as a prior on the multivariate density discussed in Chapter 2 of this thesis.

1.2 Bayesian asymptotics

Any inference or any statement of beliefs about the parameter is conditionally dependent on the choice of the prior. Especially in the context of Bayesian nonparametrics, the assessment of the quality of the corresponding posterior should require special care since it is a challenging task to define and technically regulate the priors in large dimensional spaces. Also, assigning a prior on an infinite-dimensional parameter which could be some seemingly "correct" make the posteriors easily tend to be unreasonable. Thus it could be better if the posterior could not be greatly influenced by the specification of different priors. We then say that inference is robust if the posterior is not seriously affected by the prior choices on which it is based. A tractable task is to discuss the asymptotic form of robustness. This approach helps us address the issue more manageable. The investigation of the large sample properties of the posterior naturally arises to assess the impact of a priori knowledge about the parameter on the posterior inference.

Another motivation for studying the asymptotic aspects of the posterior emerges from the analogous treatment in the frequentist nonparametric inference where statistical procedures are required to gain certain properties such as consistency, rate of convergence and minimaxity. One would ideally like to expect the Bayesian procedures to maintain these desirable frequentist properties.

To this end, we make the following assumption: the data is a random sample from some unknown but fixed distribution governed by a true value of unknown parameter. We treat Bayesian inference for the estimation problems from a classical perspective or develop the frequentist asymptotic aspects of the posterior distribution.

In this Section, two main asymptotic properties of the posterior distribution, known as posterior consistency and posterior contraction rate, will be established to quantify the effect of the priors.

1.2.1 Posterior consistency

Loosely speaking, posterior consistency entails that the posterior concentrates in every small neighbourhood of the true parameter as the sample size of the data increases indefinitely. This property could be viewed as the minimum requirement for Bayesian nonparametric inference since it ensures that the truth could be extracted accurately from the posterior as long as a sufficiently large amount of observations are collected. Bayesian nonparametric procedures become extremely undesirable for lack of posterior consistency when an unlimited amount of information is available.

We begin with a formal definition for the notion of posterior consistency. Let the observation $X = (X_1, X_2, \dots, X_n)$ be a sample of size n drawn from an unknown true distribution denoted by P_{θ_0} and consider a prior Π on the parameter space $(\Theta, \mathscr{B})$ equipped with a metric d. Let $P_{\theta_0}^n$ and $P_{\theta_0}^\infty$ denote the n-fold and infinite-fold true probability distribution P_{θ_0} respectively. The total variance distance of two probability measures P_{θ_1} and P_{θ_2} on the sample space $(\mathcal{X}, \mathscr{X})$ is defined as,

$$d_{TV}(P_{\theta_1}, P_{\theta_2}) = \sup_{A \in \mathscr{X}} |P_{\theta_1}(A) - P_{\theta_2}(A)| . \tag{1.9}$$

If P_{θ_1} and P_{θ_2} admit the corresponding probability density functions $p_{\theta_1}(x)$ and $p_{\theta_2}(x)$ relative to the Lebesgue measure respectively, then the Hellinger distance of probability measures P_{θ_1} and P_{θ_2} is given as,

$$d_H(P_{\theta_1}, P_{\theta_2}) = \left\{ \int \left(\sqrt{p_{\theta_1}(x)} - \sqrt{p_{\theta_1}(x)} \right)^2 dx \right\}^{1/2} . \tag{1.10}$$

Sometimes $d_H(p_{\theta_1}, p_{\theta_2})$ is also used to denote the Hellinger distance.

DEFINITION 1.3 (Posterior consistency) A sequence of posterior distributions $\Pi(\cdot|X_1, X_2, \ldots, X_n)$ is said to be consistent at $\theta_0 \in \Theta$ with respective to some metric d if for every $\varepsilon > 0$,

$$\Pi(\theta : d(\theta, \theta_0) > \varepsilon | X_1, X_2, \ldots, X_n) \to 0, \tag{1.11}$$

either in $P_{\theta_0}^n$-probability or $P_{\theta_0}^\infty$-almost surely .

Note that the notion of posterior consistency relies not only on the unknown true parameter and the choice of the priors, but also on the version of the posterior distribution. We could neglect the influence of the version of posterior choices in dominated case where a single version matters. The metric d is often referred to the Hellinger distance or total variance metric. We then say the posterior is Hellinger consistent at the true parameter θ_0 if d is the Hellinger distance.

As early as the 1940's, Doob's pioneering work stated that consistency held except a null set measured by the prior. Unfortunately, it did not identify the domain of the parameters where posterior consistency takes place. Freedman (1963) showed that there was a huge null-set of the prior where inconsistency occurs. Schwartz (1965) pioneered a new route to gain posterior consistency by an appropriate control on the size of the model and mild conditions about a large weak support of the prior in terms of Kullback-Leibler divergence.

The first requirement about controlling the size of the model can be generally understood to recover the true parameter P_{θ_0} from the statistical model. There are several ways to satisfy this condition. One approach to make it accessible hinges upon the existence of a uniformly exponentially powerful hypothesis test. This can be exhibited in the testing problems of the form,

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \in \{\theta : \, d(\theta, \theta_0) > \varepsilon\}. \tag{1.12}$$

The null hypothesis should be testable against parameters in $\{\theta \, : \, d(\theta, \theta_0) > \varepsilon\}$. A uniformly exponentially powerful hypothesis test could be formulated as follows,

DEFINITION 1.4 (Uniformly exponentially powerful hypothesis test) We say a sequence of test sequences $\{\phi_n(X) : \, n = 1, 2, \ldots\}$ taking values in $[0, 1]$ is uniformly exponentially consistent for the testing (1.12) if there exists $c > 0$ such that for all $\varepsilon > 0$ and $n = 1, 2, \ldots,$

$$P_{\theta_0}(\phi_n(X)) \leq e^{-cn} \quad \text{and} \quad \sup_{\theta : \, d(\theta, \theta_0) > \varepsilon} P_\theta(1 - \phi_n(X)) \leq e^{-cn}. \tag{1.13}$$

This test shows that both the type I and type II error probabilities tend to zero at an exponential rate hence we could distinguish between θ_0 and parameters in $\{\theta \, : \, d(\theta, \theta_0) > \varepsilon\}$. The latter condition about the large weak support requires that for any $\varepsilon > 0$,

$$\Pi(\theta : \, K(P_\theta, P_{\theta_0}) < \varepsilon) > 0, \tag{1.14}$$

where the Kullback-Leibler divergence between P_θ and P_{θ_0} is given as,

$$K(P_\theta, P_{\theta_0}) = -\int \log \frac{dP_\theta}{dP_{\theta_0}} \, dP_{\theta_0}. \tag{1.15}$$

THEOREM 1.5 (Schwartz (1965)) *Assume that the true parameter θ_0 is included in the parameter space Θ equipped with a prior Π such that,*

(i) the prior Π satisfies (1.14),

(ii) there exists a uniformly exponentially consistent test for (1.12).

Then for every $\varepsilon > 0$,

$$\Pi(\theta : d(\theta, \theta_0) > \varepsilon | X_1, X_2, \ldots, X_n) \to 0, \text{ in } P_{\theta_0}^{\infty}\text{-almost surely}. \tag{1.16}$$

This first requirement states that the prior should put positive mass to each small neighbourhood of the unknown true parameter θ_0. We say the priors possess the Kullback-Leibler property or the parameter lies in the Kullback-Leibler support of the priors if Schwartz's prior positivity condition (1.14) holds. It plays a key role in the investigation of posterior consistency for a wide spectrum of statistical models. A general discussion about the existence of the exponentially powerful tests exhibited in the second condition can be found in Ghosh and Ramamoorthi (2003a).

Since a Kullback-Leibler neighbourhood always includes a Euclidean neighbourhood ball, then the Kullback-Leibler support condition is always fulfilled in the parametric models for which the priors have large support in the Euclidean topology. Also the weak neighbourhoods are still large in the infinite- dimensional models. Such procedures will be problematic in stronger topologies induced by the Hellinger distance or L_1-distance widely as the loss function in the context of density estimation. However, there are several possible remedies to this problem. One popular method to treat this issue has been established with the aid of a sequence of sieves that truncated the parameter space. Let the ε-covering number denoted by $N(\mathscr{F}, \varepsilon, d)$ that is the smallest number of d-balls with radius ε needed to cover the space $\mathscr{F}$. Consider a class of density function denoted by $\mathscr{F}$ equipped with the Hellinger distance d_H and a prior Π. The observations $X_1, X_2, \ldots, X_n$ are assumed to be drawn from a true density $p_0 \in \mathscr{F}$.

THEOREM 1.6 (Ghosal et al. (1999)) *Assume that for all* $\delta > 0$,

$$\Pi(p : K(p_0, p) < \delta) > 0, \tag{1.17}$$

Furthermore, suppose that there exists a sequence of sets $\mathscr{F}_1, \mathscr{F}_2, \ldots$ *such that for positive constants*

c_1, c_2, c_3 *and each sufficiently large* n,

$$\log N(\mathscr{F}_n, \varepsilon, d) \leq c_3 n, \tag{1.18}$$

$$\Pi(\mathscr{F} \backslash \mathscr{F}_n) \leq c_1 e^{-c_2 n}. \tag{1.19}$$

Then, for any $\varepsilon > 0$,

$$\Pi(p: \, d_H(p, p_0) > \varepsilon | X_1, X_2, \ldots, X_n) \to 0, \text{ in } P_{\theta_0}^{\infty}\text{-almost surely}. \tag{1.20}$$

A similar idea of constructing appropriate sieves to give rise to consistency was carried out by Barron et al. (1999) under slightly stronger conditions. Several comments on this result are in order. The entropy condition (1.18) is an important ingredient to show Hellinger consistency. The primary merit of using the sieve device is that it may be selected a compact set on which a powerful exponentially test is built. Condition (1.19) states that the prior assigns a negligible probability mass on the complement of the sieve. Notwithstanding, the prior should maintain the Kullback-Leibler property exhibited in (1.17). Ghosal et al. (1999) deployed Dirichlet-normal mixture priors on the class of density functions to illustrate this result. Alternative approaches to deal with posterior consistency have been developed in the literature. We refer to the readers for more details in Walker and Hjort (2001), Walker (2004) and Kleijn (2015) for the recent progress on this topic.

1.2.2 Posterior contraction rate

Many procedures turn out to meet the criteria of posterior consistency, so more effort is taken to differentiate between consistent procedures. A more finer asymptotic aspect than consistency is known as the notion of posterior contraction rate. We may regard the notion of the rate of posterior convergence as the natural extension or refinement of posterior consistency. Consistency requires that the posterior shrinks to within any small ball centered on the true parameter. Here the speed of the convergence of the posterior quantifies arbitrarily small. But posterior contraction rate is typically the speed that a shrinking ball centered around the true parameter while still capturing the probability mass that goes to one. We present the precise definition for posterior contraction rate as follows,

DEFINITION 1.7 (Posterior contraction rate) Consider a positive sequence (ε_n) such that $\varepsilon_n \downarrow 0$, a sequence of posterior distributions $\Pi(\cdot|X_1, X_2, \ldots, X_n)$ is said to contract at $\theta_0 \in \Theta$ at a rate ε_n with respective to some metric d if for each $M_n \to \infty$ as $n \to \infty$,

$$\Pi(\theta : \, d(\theta, \theta_0) > M_n \varepsilon_n | X_1, X_2, \ldots, X_n) \to 0 \text{ in } P_{\theta_0}^n\text{-probability}. \tag{1.21}$$

The main aspect of the definition above distinguishes from that of consistency is that the radius of the shrinking ball depends on the sample size n. Any sequence that shrinked to zero with a rate slower than ε_n is also a contraction rate. It would be appropriate to term this " a rate of posterior convergence".

"For each M_n going to infinity" could be understood as M_n tends to infinity no matter how slowly. Typically in large dimensional cases, choosing the constant sequence M_n also works well for a large positive constant M. "$M_n \to \infty$" matters specifically in the parametric models where the posterior is needed to be scaled to a probability distribution supported on the full space.

A seminar work on the posterior convergence rate was carried out by Ghosal et al. (2000) that developed a general theory under some weak conditions along the same spirit as Schwartz (1965). Alterative approaches to address this issue involving somewhat stronger conditions can be found in Shen and Wasserman (2001) and Walker et al. (2007).

We use $D(\varepsilon, \mathscr{P}, d)$ to stand for the ε-packing number of $\mathscr{P}$. This is the maximal cardinality of some subset in $\mathscr{P}$ where every pair is of distance at least ε. In view of (2.1) in Ghosal et al. (2000), we could estimate the ε-packing number $D(\varepsilon, \mathscr{P}, d)$ by ε-covering number $N(\varepsilon, \mathscr{P}, d)$. The following theorem about the posterior contraction rates marked a significant milestone in the development of asymptotic theory in the context of Bayesian nonparametrics.

THEOREM 1.8 (Ghosal et al. (2000)) *Suppose that for a sequence ε_n with $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$, a constant $C > 0$ and sets $\mathscr{P}_n \subset \mathscr{P}$, we have,*

$$\log D(\varepsilon_n, \mathscr{P}_n, d) \leq n\varepsilon_n^2, \tag{1.22}$$

$$\Pi(\mathscr{P} \setminus \mathscr{P}_n) \leq e^{-(C+4)n\varepsilon_n^2}, \tag{1.23}$$

$$\Pi\Big(P \in \mathscr{P} \ : \ -P_0 \log \frac{dP}{dP_0} < \varepsilon_n^2,\ P_0\Big(\log \frac{dP}{dP_0}\Big)^2 < \varepsilon_n^2 \Big) \geq e^{-Cn\varepsilon_n^2}. \tag{1.24}$$

Then for sufficiently large $M > 0$, we have that,

$$\Pi\big(P \in \mathscr{P} : d(P, P_0) > M\varepsilon_n \,\big|\, X_1, \ldots, X_n \big) \longrightarrow 0 \text{ in } P_0^n\text{-probability}. \tag{1.25}$$

These three sufficient conditions can be regarded as the quantitative refinement of those alluding consistency. We provide a brief discussion of the conditions stated in this theorem. A more detailed description has been given in Ghosal et al. (2000). The prior condition (1.24) plays a crucial role in determining the rate. This condition put on the priors apart from the Kullback-Leibler property in consistency requires a sufficiently large enough probability mass on the sharpened Kullback-Leibler neighbourhood involving the second moment of the log-likelihood ratio. Intuitively, the smaller variance $P_0\Big(\log \frac{dP}{dP_0}\Big)^2$ would regulate the random variable p to fluctuate mildly, then the prior mass on this shrinking ball could be controlled with an exponential low bound.

The entropy condition (1.22) reflects to some significant extent the model complexity and ensures the existence of the exponentially powerful tests in a similar manner as that in consistency. The condition (1.23) that assigns a negligibly tiny prior mass on the complement of the sieve is not necessary especially in the cases that the sieve was chosen as the full model set itself.

The theory of posterior contraction rates has been extended for the non-i.i.d statistical settings discussed in Ghosal and van der Vaart (2007a). A greater range of priors could satisfy these sufficient conditions in most statistical models. However, the prior condition on the shrinking Kullback-Leibler neighbourhood here is somewhat restrictive. Bound support statistical problem, for example, could not meet this prior requirement. This provides a strong motivation behind Chapter 4. We refer the reader to this Chapter for more details.

1.2.3 Optimality and adaptation

Some estimators that we shall use in this thesis achieve the optimal posterior convergence rate. Informally speaking, this implies that no point estimator whether Bayesian or classical, can contract at a faster rate. In fact, it could be shown that if the posterior contracts at the rate ε_n, the point estimator that summarizes the location information about the posterior converges to the true parameter at the same rate ε_n. It follows that the posterior contraction rate could not exceed the minimax rate in terms of its speed. Thus that the posterior achieves the minimax rate could be viewed as an ideal target.

In a frequentist paradigm the general theory of the so-called minimax rate has been well established. A detailed introduction about the minimax rate has been given in Tsybakov (2009). It is well known that the minimax rate is of the order $n^{-\alpha/(d+2\alpha)}$ in the context of α-regular function estimation with d arguments. Frequentist adaptive procedures have been extensively developed in the literature, see for example, Bickel (1982), Efroimovich (1986), a series of papers by Lepskii (1991, 1992, 1993) or Tsybakov (2009) for a textbook treatment on this topic.

A Bayesian procedure is said to converge at the minimax rate if the posterior contraction rate in some space agrees with the minimax rate. In most statistical models, this space is characterized by a few hyperparameters, for instance, that describes the features of the model, such as the regularity, the shape and the sparsity.

In general, the optimal Bayesian procedures rely heavily on the true value of these hyperparameters. For example, in the context of α-smooth densities, we need to build a prior via the use of B-splines where the choice of the number of the elements depends on the smoothness level α. A collection of priors indexed by the regularity level α should be constructed. But the exact value of the smoothness level of the unknown parameter is rarely known to us in a realistic situation. Therefore a prior constructed for this target class will lead to a suboptimal rate at the true density if it is actually coarser or smoother than the hypothesized class. The failure of the regularity of the prior matching that of the unknown true parameter prompts us to construct a prior that yields the minimax rate but does not hinge on the information about the unknown regularity.

The posterior is called adaptive to all regularity levels if such a prior exists. From a Bayesian view of point, adaptation received growing attention in the past decade. Early work on Bayesian adaptation in an infinite-dimensional model has been done by Belitser and Ghosal (2003) where the domain of unknown parameter is accountable. Later Bayesian adaptive procedures have been established for various statistical models. The minimax concentration rate without an additional logarithm item was obtained in Huang (2004) who treated nonparametric regression problem using a wavelet basis. A wide range of Gaussian process priors could yield adaption for all smoothness levels in the context of density estimation, nonparametric regression and classification settings in van der Vaart and van Zanten (2009) and de Jonge and van Zanten (2010, 2012). Other particular class of priors have been investigated by Scricciolo (2006), van der Vaart and van Zanten (2009), Rousseau (2010), Kruijer et al. (2010), Shen and Ghosal (2012), Shen et al. (2013), Norets and Pati (2014) and Belitser and Serra (2014), among others. Recent attention has switched to alternative adaptive Bayesian techniques known as empirical Bayesian methods in which these unknown hyperparameters are estimated from the data in a frequentist perspective. These data-driven choices for the hyperparameters could give rise to better statistical inference. More general discussions could be found in Szabo (2014) and Szabo et al. (2013).

1.3 Overview of thesis

In this thesis we focus on asymptotic aspects of the nonparametric Bayesian procedures. In Chapter 2 we examine the posterior concentration rate in the context of multivariate density estimation using Bernstein polynomial prior. Also an MCMC algorithm based on slice sampling is developed to sample from the posterior. An extensive illustration for the proposed approach is conducted using simulated data and real data. Chapter 3 centers on the adaptive Bayesian procedure in a linear regression model where the error variance depends on the covariates. The rate of posterior contraction is obtained without any priori knowledge of the regularity level of the true error standard deviation function. Chapter 4 is mainly based on a work jointly written with Professor Bas Kleijn at University of Amsterdam. We propose a general theory of the posterior convergence rate in the statistical settings where the condition that priors shall assign a large amount of mass on the shrinking balls fails. This result is illustrated in several statistical models.

1.3.1 Chapter 2

This Chapter introduces a new approach to Bayesian nonparametric inference for densities on the hypercube, based on the use of a multivariate Bernstein polynomial prior. Posterior convergence rates under the proposed prior are obtained. A novel sampling scheme for the estimation of the posterior predictive density is developed. The algorithm is based on a stick-breaking representation of the model as well as the use of slice sampling techniques. The approach is illustrated with both simulated and real data examples.

1.3.2 Chapter 3

In this chapter we consider adaptive Bayesian semiparametric analysis of the linear regression model in the presence of conditional heteroskedasticity. The distribution of the error term on predictors is modelled by a normal distribution with covariate-dependent variance. We show that a rate-adaptive procedure for all smoothness levels of this standard deviation function is performed if the prior is properly chosen. More specifically, we derive adaptive posterior distribution rate up to a logarithm factor for the conditional standard deviation based on a transformation of hierarchical Gaussian spline prior and log-spline prior respectively.

1.3.3 Chapter 4

Most existing work primarily concerning the properties of posterior contraction rates relies heavily on carefully chosen priors that meet some requirements. A typical condition imposed on the priors has been proposed in Ghosal et al. (2000), i.e.

$$\Pi\Big(P \in \mathscr{P} \,:\, -P_0 \log \frac{dP}{dP_0} < \varepsilon_n^2,\, P_0\Big(\log \frac{dP}{dP_0}\Big)^2 < \varepsilon_n^2 \Big) \geq e^{-n\varepsilon_n^2}, \tag{1.26}$$

where (ε_n) is a positive sequence such that $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$. We call priors satisfying condition (1.26) above *GGV priors*. GGV priors play a crucial role in exploring the rate of posterior contraction in a broad spread of statistical models.

In this Chapter, we try to relax this condition to accommodate a wide range of priors. To that end, we formulate an alternative rates-of-posterior-convergence theorem, based on an approach proposed in Kleijn (2015). The aim is to strengthen model conditions and gain flexibility in the choice for a prior, while maintaining optimality of the posterior rate of convergence.

Chapter 2

Bayesian multivariate Bernstein polynomial density estimation

2.1 Introduction

Many real data samples possess characteristics such as multimodality, high skewness and kurtosis which are not well modeled by standard parametric distributions. In such cases, nonparametric modeling techniques might be preferable.

Although kernel density estimation techniques are the most popular approaches from the classical viewpoint, see e.g. Silverman (1986), in certain situations, alternative approaches based on approximating polynomials have been considered. In particular, Vitale (1975) developed a Bernstein polynomial based density estimator for density functions on a closed interval and this was extended to bivariate densities in Tenbusch (1994).

In the Bayesian context, most nonparametric density estimation is based on the use of Dirichlet process or Dirichlet process mixture priors, see e.g. Hjort et al. (2010) for a general review of the area. However, in the case of univariate densities on a closed interval, Petrone (1999a,b) developed an alternative approach based on the use of a Bernstein polynomial based prior. Consistency properties of the derived posterior distribution were examined in Petrone and Wasserman (2002).

The convergence rate of the posterior was derived in Ghosal (2001) and further studied in Kruijer and van der Vaart (2008). An extended Bernstein polynomial prior model was examined in Trippa et al. (2011). Finally, software for Bayesian Bernstein polynomial density estimation was developed in Jara et al. (2011).

However, much less effort has been devoted to the generalization of Bernstein polynomial priors to the multivariate case. One exception is Zheng et al. (2010) where a multivariate Bernstein polynomial prior is assumed for the spectral density of a random field. Also, posterior convergence rates of certain bivariate Bernstein polynomial priors are derived in Kruijer and van der Vaart (2008). In this Chapter, we derive the convergence rate of posterior distribution of a multivariate Bernstein polynomial model under very general conditions. Nevertheless, the main contribution of this Chapter is to introduce a stick-breaking representation of the model and develop a new computational approach to implementing multivariate Bernstein polynomial density estimation. The proposed algorithm is based on the slice sampling techniques for Dirichlet process mixture models developed in Walker (2007) and Kalli et al. (2011). It is shown to be less sensitive in computational time to large sample sizes and high variable dimension than the multivariate version of the algorithm used in Petrone (1999a).

The rest of this Chapter is organized as follows. Firstly, in Section 2, we briefly outline the properties of univariate Bernstein polynomials. In Section 3, we introduce the multivariate Bernstein polynomial prior and derive the associated posterior convergence rates. In Section 4, we present the multivariate Bernstein-Dirichlet prior and provide an algorithm to sample from the posterior distribution, which is a direct generalization to the multivariate case of the approach developed in Petrone (1999a). We discuss some of the computational inconveniences of this procedure and then, we construct a different representation of the model, which allows us to define a much more efficient sampling algorithm. Section 5 then illustrates our approach with both simulated and real data examples and finally, some conclusions and extensions are provided in Section 6.

2.2 Univariate random Bernstein polynomials

Bernstein polynomials, introduced by Bernstein (1912), are well known to provide good approximations to continuous functions on a closed interval. Let $h(x)$ be a continuous and bounded, real function defined on $[0,1]$. Then, the Bernstein polynomial of degree k for $h(x)$ is defined by:

$$B(x;k,h) := \sum_{j=0}^{k} h\left(\frac{j}{k}\right)\binom{k}{j} x^j (1-x)^{k-j}. \tag{2.1}$$

It is well known that, letting k tend to infinity, the Bernstein polynomial approximations converge uniformly to h and, moreover, that their derivatives also converge to the corresponding derivatives of h. More details and further results on the approximation properties of Bernstein polynomials are provided in e.g. Lorentz (1986) and Phillips (2003).

Bernstein polynomials are particularly useful to approximate distribution and density functions for variables defined on a closed interval. Let G be a distribution function on $[0,1]$ with $G(0)=0$, then it is easy to show that the k-th order Bernstein polynomial approximation to the corresponding density function is given by a mixture of beta densities:

$$b(x|k,G) = \sum_{j=1}^{k} \omega_{j;k}\beta(x|j,k-j+1), \tag{2.2}$$

where $\omega_{j;k} = \left[G\left(\frac{j}{k}\right) - G\left(\frac{j-1}{k}\right)\right]$,$\beta(x|c,d) = \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)}x^{c-1}(1-x)^{d-1}$ is a beta density and in the sequel, we omit to mention the argument to denote the beta density by $\beta(\cdot,\cdot)$.

Petrone (1999a,b) proposed the use of Bernstein polynomials to define a prior on the class of densities on $[0,1]$, called the *Bernstein polynomial prior*. This assumes a random density of the form (2.2), where k follows a discrete probability distribution $p(k)$ and given k, $\boldsymbol{\omega}_k^1 = (\omega_{1;k}\ldots,\omega_{k;k})$

follows a distribution function H_k^1 on the k-dimensional simplex,

$$\Delta_k^1 = \left\{ (\omega_{1;k} \ldots, \omega_{k;k}) : 0 \leq \omega_{j;k} \leq 1, \; j = 1, 2, \ldots, k, \; \sum_{j=1}^{k} \omega_{j;k} = 1 \right\}.$$

Petrone (1999a,b) showed that if for all k, $p(k) > 0$ and the density of H_k^1 is positive for any point in Δ_k^1, then every distribution on $[0, 1]$ is in the weak support of the Bernstein polynomial prior. Also, Petrone and Wasserman (2002) showed that the posterior distribution corresponding to this prior is consistent at any continuous true density, f_0, on $[0, 1]$.

Ghosal (2001) obtained the rates of convergence of the posterior distribution for the Bernstein polynomial prior assuming that the weight $\boldsymbol{\omega}_k^1$ follows a Dirichlet distribution $Dir(\alpha_{1;1}, \ldots, \alpha_{k;k})$ where the parameters $\alpha_{j;k}$ are bounded by some number M for all j and k. If the true density is itself a Bernstein density, then the rate is close to the parametric rate $n^{-1/2} \log n$. Otherwise, the rate of convergence is $n^{-1/3}(\log n)^{5/6}$, provided that the true density is strictly positive together with bounded second derivative.

One important case of the Bernstein polynomial prior is the so-called *Bernstein-Dirichlet prior*, which is defined by letting $\alpha_{j;k} := M\left(G_0(\frac{j}{k}) - G_0(\frac{j-1}{k})\right)$, where G_0 is a probability distribution function on $[0, 1]$. This is equivalent to assuming that G follows a Dirichlet process prior, $\mathcal{DP}(M, G_0)$, which is independent from the prior probability $p(k)$. Kruijer and van der Vaart (2008) have obtained the adaptive rate of convergence under this prior for strictly positive and α-smooth true densities when $\alpha \in (0, 2]$. They also extend their results to the multivariate case that we will consider in the next Section.

Regarding inference algorithms, the literature is much less developed. Petrone (1999a,b) propose an MCMC algorithm to approximate the predictive density for the Bernstein-Dirichlet prior. The results obtained are adequate but sometimes the algorithm can be very slow, especially if the sample size is large. Alternatively, Petrone and Wasserman (2002) consider a different approximation by computing the maximum likelihood density estimates for each k and the averaging with respect to some weights derived from the BIC or AIC. All these algorithms are based on the introduction of auxiliary variables, Y_i, independent and identically distributed according to G. More specifically, for a sequence of exchangeable variables, $\{X_1, X_2, \ldots\}$ with values in $[0,1]$, the Bernstein-Dirichlet prior is written hierarchically as follows:

$$\begin{aligned} X_i \mid k, Y_i &\sim \beta\left(z_k\left(Y_i\right), k - z_k\left(Y_i\right) + 1\right), \\ Y_i \mid G &\sim G, \\ G &\sim \mathcal{DP}(M, G_0), \\ k &\sim p(k). \end{aligned}$$

where

$$z_k\left(Y_i\right) := \sum_{j=1}^{k} j\, \mathbb{I}\left(\frac{j-1}{k} < Y_i \leq \frac{j}{k}\right), \tag{2.3}$$

and where $\mathbb{I}(\cdot)$ is the indicator function. Note that the auxiliary variables Y_i, provide the labels of the components of the beta mixture, for any value of k.

In this Chapter, we will use similar auxiliary variables, but we will consider an alternative specification for the prior model based on the stick-breaking representation of the Dirichlet process by Sethuraman (1994). Observe that, using this approach, we can write the Bernstein-Dirichlet prior as an infinite mixture of beta densities as follows:

$$f(x \mid k, \boldsymbol{\rho}, \mathbf{y}) := \sum_{s=1}^{\infty} \rho_s \beta\left(x | z_k(y_s), k - z_k(y_s) + 1\right), \tag{2.4}$$

where $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots)$ such that $\rho_1 = v_1$ and $\rho_s = v_s \prod_{l=1}^{s-1}(1 - v_l)$,$s = 2, 3, \ldots$, where a Beta prior distribution is assumed for $v_l \sim \beta(1, M)$, for $l = 1, 2, \ldots$, and $\mathbf{y} = (y_1, y_2, \ldots)$ such that the baseline prior distribution G_0 is assumed for y_s, $s = 1, 2, \ldots$.

Based on the stick-breaking representation of the Dirichlet process, Walker (2007) and Kalli et al. (2011) propose MCMC schemes using slice sampling techniques to deal with the infiniteness in Dirichlet process mixture models. These procedures compared to traditional approaches based on the original algorithm by Escobar and West (1995) produce better, faster and easier to implement algorithms. Therefore, it seems reasonable to use these ideas for approximating the predictive density under the Bernstein-Dirichlet prior.

In order to provide intuition for the reduction of computational cost, observe that using the MCMC approaches of Petrone (1999a,b) and Petrone and Wasserman (2002), it is necessary to sample n auxiliary variables y_i for $i = 1, \ldots, n$, where n is the sample size, and k weights, $(\omega_{1,k}, \ldots, \omega_{k,k})$ at each MCMC iteration given k. On the contrary, using the representation (2.4) and the slice sampling ideas based on Walker (2007), in practice it is only required to sample a finite number of mixture parameters (ρ_s, y_s) for $s = 1, \ldots, s^*$ at each MCMC iteration, where s^* will be in general much smaller than the sample size n.

2.3 Multivariate random Bernstein polynomials

Let the m-dimensional unit hypercube be denoted by $[0,1]^m$. Then, the associated m-dimensional Bernstein polynomial approximation at $\mathbf{x} = (x_1, \ldots, x_m)$, for a continuous, bounded function h on $[0,1]^m$ is defined by:

$$B(\mathbf{x}; k, h) := \sum_{j_1=0}^{k} \cdots \sum_{j_m=0}^{k} h\left(\frac{j_1}{k}, \ldots, \frac{j_m}{k}\right) \left(\prod_{r=1}^{m} \binom{k}{j_r} x_r^{j_r} (1 - x_r)^{k-j_r}\right). \tag{2.5}$$

As in the univariate case, the Bernstein polynomials and their derivatives converge uniformly to h and their corresponding derivatives as $k \to \infty$. In the case that $h = G$ is a multivariate distribution function, then analogous to (2.2), the corresponding density approximation is given by:

$$b(\mathbf{x}; k, G) = \sum_{j_1=1}^{k} \cdots \sum_{j_m=1}^{k} w_{j_1 j_2 \ldots j_m; k} \prod_{r=1}^{m} \beta(x_r | j_r, k - j_r + 1), \tag{2.6}$$

where $w_{j_1 j_2 \ldots j_m; k} = \int_{(j_1-1)/k}^{j_1/k} \cdots \int_{(j_m-1)/k}^{j_m/k} g(x_1, x_2, \ldots, x_m) dx_1 dx_2 \ldots dx_m$ and where g is the corresponding density function.

In particular, the multivariate Bernstein density of order k in 2.6 could uniformly approximate the smooth multivariate density functions with the error bounded by $1/k$.

LEMMA 2.1 *Let the probability density g be continuously differentiable on $[0,1]^m$ with bounded determinant of its associated Hessian matrix, then*

$$\sup_{0<x_1,x_2,\ldots,x_m\leq 1} |g(\mathbf{x}) - b(\mathbf{x};k,G)| = O(k^{-1}). \tag{2.7}$$

This property can easily be shown by observing that,

$$b(\mathbf{x};k,G) = k^m \mathbb{E}\left(\int_{J_1/k}^{(J_1+1)/k} \ldots \int_{J_m/k}^{(J_m+1)/k} g(\mathbf{z})\,\mathrm{d}z_1 \ldots \mathrm{d}z_m\right), \tag{2.8}$$

where $\mathbf{z} := (z_1, z_2, \ldots, z_m)$ and $J_r \sim$ Binomial$(k-1, x_r)$ for $r = 1, 2, \ldots, m$. Also the proof of this lemma is relegated to the Appendix.

It is straightforward to extend the results by Petrone (1999a,b) to the multivariate case to define a *multivariate Bernstein polynomial prior*, see Zheng et al. (2010). This consists of a multivariate random density defined on $[0,1]^m$ given by (2.6), where k has probability mass function $p(k)$, and given k, the weights:

$$\boldsymbol{\omega}_k^m := \{\omega_{j_1\ldots j_m;k} : \, j_r = 1, \ldots, k; \; r = 1, \ldots, m\}, \tag{2.9}$$

follow a distribution H_k on the k^m-dimensional simplex:

$$\Delta_k^m := \left\{\boldsymbol{\omega}_k^m : \, 0 \leq \omega_{j_1\ldots j_m;k} \leq 1, j_r = 1, \ldots, k, \; r = 1, \ldots, m, \sum_{j_1=1}^{k} \ldots \sum_{j_m=1}^{k} w_{j_1 j_2 \ldots j_m;k} = 1\right\}. \tag{2.10}$$

Similarly to the univariate case, it can be shown that every probability distribution on $[0,1]^m$ lies in the topology of weak convergence of the m-variate Bernstein polynomial prior provided that $p(k) > 0$, for all k, and the density of H_k^m is positive for any point in Δ_k^m, see Zheng et al. (2010).

2.3.1 The convergence rate of the posterior distribution

In this Section, we derive the convergence rate of the posterior distribution of the multivariate Bernstein polynomial prior, extending Theorem 2.3 of Ghosal (2001) to the multivariate case. Similar to Ghosal (2001), we will assume that the k^m weights, given in (2.9), follow a Dirichlet distribution $Dir(\alpha_{11\ldots1;k}, \alpha_{11\ldots2;k}, \ldots, \alpha_{kk\ldots k;k})$, where the parameters, $\alpha_{j_1,\ldots,j_m;k}$, are bounded above by some finite number, M, for all j_r and k, for $r = 1, \ldots, m$.

Assume that we observe an independent and identically distributed sample, $\mathbf{X}_1, \ldots, \mathbf{X}_n$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{im})$ are generated from a true m-variate density $f_0 \in \mathscr{P}$, where $\mathscr{P}$ denotes the class of all probability density functions supported on $[0,1]^m$. Suppose $\mathscr{P}$ is equipped with a Borel algebra $\mathscr{F}$. Let P_0 be the probability measure with density f_0 and let $d_H(\cdot,\cdot)$ and $\|\cdot\|_1$ stand for the Hellinger distance and L_1-norm respectively. Then, following Ghosal (2001), given a prior, Π on a set of m-variate densities, the posterior distribution is a random measure:

$$\Pi\left(f \in A \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right) = \int_A \prod\nolimits_{i=1}^n f\left(\mathbf{X}_i\right) \Pi(df) \Big/ \int \prod\nolimits_{i=1}^n f\left(\mathbf{X}_i\right) \Pi(df), \tag{2.11}$$

where $A \in \mathscr{F}$. Assume a multivariate Bernstein prior and therefore, the prior on the density of the observations is concentrated on the space $\bigcup_{k=1}^{\infty} \mathscr{B}_k^m$ where $\mathscr{B}_k^m$ is the set of multivariate Bernstein densities of order k given by:

$$\mathscr{B}_k^m := \left\{ \sum_{j_1=1}^{k} \cdots \sum_{j_m=1}^{k} w_{j_1 j_2 \ldots j_m;k} \prod_{r=1}^{m} \beta(x_r | j_r, k - j_r + 1) : \boldsymbol{\omega}_k^m \in \Delta_k^m \right\}.$$

Then, under some conditions the following theorem establishes the corresponding posterior distribution converges to some rate and the proof is given in the Appendix.

THEOREM 2.2 *Let the true density f_0 be bounded away from 0 and satisfy the assumptions stated in lemma 2.1. Consider a Bernstein polynomial prior for f satisfying the condition* $B_1 e^{-\beta_1 k^m} \le p(k) \le B_2 e^{-\beta_2 k^m}$

for all k and some constants $B_1, B_2, \beta_1, \beta_2 > 0$. Then for a sufficiently large positive constant M,

$$\Pi\left(f : d_H(f, f_0) > Mn^{-1/(m+2)}(\log n)^{(m+4)/(2m+4)} \middle| \mathbf{X}_1, \ldots, \mathbf{X}_n \right) \to 0 \textit{ in } P_0^n\textit{-probability.}$$

2.4 The multivariate Bernstein-Dirichlet prior

In this Section, we introduce the *multivariate Bernstein-Dirichlet prior* which, similar to the univariate case, is defined as a multivariate Bernstein prior where the unknown distribution function G in (2.6) follows an m-dimensional Dirichlet process prior. Therefore, for a sequence of exchangeable multivariate variables, $\{\mathbf{X}_1, \mathbf{X}_2, \ldots,\}$, with values on $[0,1]^m$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{im})$, this prior model can be written hierarchically as follows:

$$\begin{aligned} \mathbf{X}_i \mid k, G &\sim b(\cdot; k, G) \quad \text{are conditionally i.i.d.,} \\ k &\sim p(k), \\ G &\sim \mathcal{DP}(M, G_0), \end{aligned} \tag{2.12}$$

where $b(\mathbf{x}; k, G)$ is as in (2.6) and the base measure, G_0, is a multivariate distribution function defined on $[0,1]^m$. This implies that the k^m weights, $\boldsymbol{\omega}_k^m$, defined in (2.9), follow a Dirichlet distribution, $Dir(\alpha_{11\ldots1;k}, \alpha_{11\ldots2;k}, \ldots, \alpha_{kk\ldots k;k})$, with parameters given by,

$$\alpha_{j_1 j_2 \ldots j_m;k} := M \int_{(j_1-1)/k}^{j_1/k} \cdots \int_{(j_m-1)/k}^{j_m/k} g_0(\mathbf{x})\, dx_1 dx_2 \ldots dx_m, \tag{2.13}$$

where $j_r = 1, \ldots, k$, and $r = 1, \ldots, m$, and g_0 is the probability density corresponding to the Dirichlet base measure, G_0.

Given this multivariate Bernstein-Dirichlet prior, we are now interested in sampling from the predictive density of a new observation. The first attempt consists in extending the algorithm by Petrone (1999a,b) to the multivariate case. Then, as in the univariate case, we introduce a set of auxiliary variables, $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots,\}$, with values on $[0,1]^m$, which are i.i.d. according to the multivariate Dirichlet process G. Therefore, the model structure can be represented in a hierarchical way as follows:

$$\begin{aligned} X_{ir} \mid k, Y_{ir} &\sim \beta\left(z_k\left(Y_{ir}\right), k - z_k\left(Y_{ir}\right) + 1\right), \quad \text{for } r = 1, \ldots, m, && (2.14)\\ \mathbf{Y}_i \mid G &\sim G, && (2.15)\\ G &\sim \mathcal{DP}(M, G_0), && (2.16)\\ k &\sim p(k), && (2.17) \end{aligned}$$

where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})$ and the function $z_k(\cdot)$ is defined in (2.3). Observe that the hidden variable, $\mathbf{Y}_i$, provides information about which of the k^m components of the mixture (2.6) the $\mathbf{X}_i$ comes from. Therefore,

$$\mathbf{X}_i \mid \mathbf{Y}_i \sim \prod_{r=1}^{m} \beta\left(x_{ir} \mid j_r, k - j_r + 1\right) \quad \text{if } \mathbf{Y}_i \in \left(\frac{j_1 - 1}{k}, \frac{j_1}{k}\right] \times \ldots \times \left(\frac{j_m - 1}{k}, \frac{j_m}{k}\right].$$

Given an observed sample, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$, it is straightforward to extend the hybrid Monte Carlo algorithm of Petrone (1999a,b) to the multivariate case as follows. Firstly, select a starting value $(\mathbf{y}_1^{(0)}, \ldots, \mathbf{y}_n^{(0)})$. Then, repeat iteratively the following steps.

```
(i) Generate a value from the conditional posterior distribution of $k$ which
    is proportional to:
```

$$p(k) \prod_{i=1}^{n} \prod_{r=1}^{m} \beta(x_{ir}; z_k(y_{ir}), k - z_k(y_{ir}) + 1).$$

```
(ii) For $\nu = 1, \ldots, n$:

    (a) With probability $q(\nu) \propto Mb(\mathbf{x}_\nu; k, G_0)$, sample $\mathbf{y}_\nu$ from the density:
```

$$\psi(y_1, \ldots, y_m) \propto g_0(\mathbf{y}) \prod_{r=1}^{m} \beta(x_{\nu r}; z_k(y_r), k - z_k(y_r) + 1). \qquad (2.18)$$

(b) With probability,

$$q(l) \propto \prod_{i=1}^{n} \prod_{r=1}^{m} \beta(x_{\nu r}; z_k(y_{ir}), k - z_k(y_{ir}) + 1),$$

for $l = 1, \ldots, \nu - 1, \nu + 1, \ldots, n$, and set $\mathbf{y}_\nu = \mathbf{y}_i$.

(iii) Sample the weights, $\omega_{j_1 \ldots j_m; k}$, for $j_r = 1, \ldots, k$, and $r = 1, \ldots, m$, from a Dirichlet distribution with parameters, $(\alpha_{j_1 \ldots j_m; k} + N_{j_1 \ldots j_m; k})$, where $\alpha_{j_1 \ldots j_m; k}$ is defined in (2.13) and

$$N_{j_1 \ldots j_m; k} := \sum_{i=1}^{n} \prod_{r=1}^{m} \mathbb{I}\left(\frac{j_r - 1}{k} < y_{ir} \leq \frac{j_r}{k} \right). \tag{2.19}$$

Given a posterior sample from this algorithm, the predictive density can be approximated by,

$$\frac{1}{T} \sum_{t=1}^{T} \left(\sum_{j_1=1}^{k^{(t)}} \cdots \sum_{j_m=1}^{k^{(t)}} w^{(t)}_{j_1 j_2 \ldots j_m, k^{(t)}} \prod_{r=1}^{m} \beta(x_r | j_r, k^{(t)} - j_r + 1) \right), \tag{2.20}$$

where T denotes the size of the posterior sample and $k^{(t)}$ and $w^{(t)}_{j_1 j_2 \ldots j_m; k^{(t)}}$ denote, respectively, the values of the polynomial order and the weights at each stage of the algorithm.

Some comments about the algorithm are in order. Firstly, step 1 above is straightforward, as the distribution of k is defined on the positive integers. Then, we may compute the conditional posterior probabilities on a large range, $[1, k_{\text{max}}]$, or alternatively, define a Metropolis-Hastings algorithm where a candidate, $\tilde{k} = k \pm 1$, is generated with 0.5 probability. Step 2 is also easy to implement since the product of betas in (2.18) is stepwise constant in each hypercube,

$$\left(\frac{j_1 - 1}{k}, \frac{j_1}{k} \right] \times \ldots \times \left(\frac{j_m - 1}{k}, \frac{j_m}{k} \right]. \tag{2.21}$$

However, this step can be computationally intensive, especially if the sample size is large since it is required to sample a missing value from $(\mathbf{y}_\nu \mid \mathbf{y}_1, \ldots, \mathbf{y}_{\nu-1}, \mathbf{y}_{\nu+1}, \ldots, \mathbf{y}_n)$ for each $\nu = 1, \ldots, n$.

Finally, step 3 is also simple to carry out, but it may be very costly when the dimension, m, is large since the number of weights to sample is k^m. Furthermore, observe that $N_{j_1\ldots j_m;k}$, in (2.19), counts the number of $\mathbf{y}_i$'s in each hypercube, (2.21), and this can be frequently zero if m is large due to data sparsity, which can be viewed as a curse of dimensionality.

In the next Subsection, we will introduce a stick-breaking representation of the model which will allow for the construction of a more efficient MCMC algorithm. This will be based on the slice sampling techniques for Dirichlet process mixtures proposed in Walker (2007), which considers the introduction of further auxiliary variables to handle with the infiniteness of the model. The proposed algorithm will be shown to be less sensitive in computational time to large sample sizes and/or variable dimension and also quite easy to implement. Further, it will not require the sampling of k^m weights at each stage of the algorithm so that problems with data sparsity will be avoided.

2.4.1 Stick-breaking representation

In this Section, we consider a different representation of the Bernstein-Dirichlet prior model introduced in (2.14)-(2.17). Firstly, observe that given k, the model in (2.14)-(2.16) is a Dirichlet process mixture, as introduced in Antoniak (1974), of independent beta densities driven by certain parameters, $\mathbf{Y}_i$. Therefore, using the stick-breaking representation of Sethuraman (1994), we may rewrite the multivariate Bernstein-Dirichlet prior as a countably infinite mixture model of independent beta densities on the unit hypercube as follows:

$$f(\mathbf{x}_i|k,\boldsymbol{\Omega},\boldsymbol{\Theta}) = \sum_{s=1}^{\infty} \rho_s \prod_{r=1}^{m} \beta\left(x_{ir}|z_k(y_{sr}), k - z_k(y_{sr}) + 1\right), \tag{2.22}$$

where $\boldsymbol{\Omega} = (\rho_1, \rho_2, \ldots)$ is an infinite set of weights such that $\rho_1 = v_1$ and $\rho_s = v_s \prod_{l=1}^{s-1}(1 - v_l)$, where a Beta prior distribution is assumed for $v_s \sim \beta(1, M)$, for $s = 1, 2, \ldots$, and where $\boldsymbol{\Theta} = (\mathbf{y}_1, \mathbf{y}_2, \ldots)$ is an infinite set of multivariate parameters, where $\mathbf{y}_s = (y_{s1}, \ldots, y_{sm})$, for $s = 1, 2, \ldots$, such that each $\mathbf{y}_s$ follows the baseline multivariate prior distribution G_0 and $z_k(\cdot)$ is defined in

(2.3).

Note that there is a difference between how missing variables, $\mathbf{y}_s$, are introduced here and in the previous algorithm. Observe that above, given a sample of multivariate observations, $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, there were a set of associated missing data, $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, which were i.i.d. from G. On the contrary, using the specification in (2.22), these missing variables are viewed as parameters such that each observation, $\mathbf{x}_i$, is governed the same set of infinite parameters, $\mathbf{\Theta} = (\mathbf{y}_1, \mathbf{y}_2, \ldots)$, for $i = 1, \ldots, n$.

Following Walker (2007), we now introduce a uniform latent variable u_i over $[0, 1]$ to convert the infinite mixture representation, (2.22), into a finite mixture representation as follows:

$$\begin{aligned} f(\mathbf{x}_i, u_i | k, \mathbf{\Omega}, \mathbf{\Theta}) &= \sum_{s=1}^{\infty} \mathbb{I}(u_i < \rho_s) \prod_{r=1}^{m} \beta\left(x_{ir} | z_k(y_{sr}), k - z_k(y_{sr}) + 1\right), \\ &= \sum_{s \in A(u_i)} \prod_{r=1}^{m} \beta\left(x_{ir} | z_k(y_{sr}), k - z_k(y_{sr}) + 1\right), \end{aligned}$$

where the set $A(u_i) := \{s : u_i < \rho_s\}$, which is clearly a finite set. Observe that integrating over u_i, we obtain the original infinite mixture density, (2.22). Also, given u_i, the number of mixture components is finite,

$$f(\mathbf{x}_i | u_i, k, \mathbf{\Omega}, \mathbf{\Theta}) = \frac{1}{R_i} \sum_{s \in A(u_i)} \prod_{r=1}^{m} \beta\left(x_{ir} | z_k(y_{sr}), k - z_k(y_{sr}) + 1\right),$$

where $R_i := \sum_{s=1}^{\infty} \mathbb{I}(u_i < \rho_s)$, is the number of elements in the set $A(u_i)$. Finally, the marginal distribution of u_i is uniform on the interval $[0, \rho_s]$, with probability ρ_s.

Then, we may introduce a further latent label variable, d_i, indicating to which of these finite mixture components belongs each observation:

$$f(\mathbf{x}_i, u_i, d_i | k, \mathbf{\Omega}, \mathbf{\Theta}) = \mathbb{I}(u_i < \rho_{d_i}) \prod_{r=1}^{m} \beta\left(x_{ir} | z_k(y_{d_i r}), k - z_k(y_{d_i r}) + 1\right).$$

Therefore, the complete likelihood function based on the sample $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is:

$$l(k, \mathbf{\Omega}, \mathbf{\Theta}|\mathbf{X}, \mathbf{u}, \mathbf{d}) = \prod_{i=1}^{n} \prod_{r=1}^{m} \mathbb{I}(u_i < \rho_{d_i}) \beta\left(x_{ir}|z_k(y_{d_i r}), k - z_k(y_{d_i r}) + 1\right).$$

where $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ and $\mathbf{d} = (d_1, d_2, \ldots, d_n)$.

Now, we can construct an MCMC algorithm to sample the posterior parameter distribution as follows. Firstly, select a starting allocation, $\mathbf{d}^{(0)} = (d_1^{(0)}, \ldots, d_n^{(0)})$. Then, simulate a Markov chain by repeating iteratively the following steps.

(i) Sample a finite number of weights $(\rho_1, \ldots, \rho_{s^*})$ jointly with $(u_1, \ldots, u_n)$.

(a) Sample from $v_s \sim \beta(n_s + 1, n - \sum_{l=1}^{s} n_l + M)$ for $s = 1, \ldots, d^*$, where $d^* = \max\{d_1, \ldots, d_n\}$ and $n_s = \sum_{i=1}^{n} \mathbb{I}(d_i = s)$ and set $\rho_s = v_s \prod_{l=1}^{s-1}(1 - v_l)$.

(b) Sample u_i by simulating from $U(0, \rho_{d_i})$ for $i = 1, \ldots, n$.

(c) If necessary, generate more weights, ρ_s, from the prior, by simulating from $v_s \sim \beta(1, M)$, until $\sum_{s=1}^{s^*} \rho_s > 1 - u^*$, where $u^* = \min\{u_1, \ldots, u_n\}$.

(ii) Sample the mixture parameters, $\mathbf{y}_s$, for $s = 1, \ldots, s^*$, by simulating from:

$$f(y_{sr}|\cdots) \propto f_0(y_{sr}) \prod_{i:d_i=s} \beta(x_{ir}; z_k(y_{sr}), k - z_k(y_{sr}) + 1). \qquad (2.23)$$

independently for $r = 1, \ldots, m$. If there is no d_i equal to s then sample $\mathbf{y}_s$ from G_0.

(iii) Sample the allocation variables, d_i, for $i = 1, \ldots, n$, by simulating from:

$$P(d_i = s|\cdots) \propto \mathbb{I}(u_i < \rho_s) \prod_{r=1}^{m} \beta(x_{ir}; z_k(y_{sr}), k - z_k(y_{sr}) + 1). \qquad (2.24)$$

`(iv) Sample` k `by simulating from the following full conditional distribution:`

$$P(k|\cdots) \propto p(k)\prod_{i=1}^{n}\prod_{r=1}^{m}\beta(x_{ir};z_k(y_{d_ir}),k-z_k(y_{d_ir})+1). \tag{2.25}$$

A few comments on this algorithm are in order. Firstly, step 1 is from Walker (2007) and Papaspiliopoulos and Roberts (2008) and only requires to sample from beta and uniform distributions. Step 2 is easy to sample as the product of betas is a stepwise function on $[0,1]$. Note that, the value of s^* will be in general smaller than the sample size n, and therefore, this step will be usually faster than step 2 in the algorithm based on Petrone (1999b) introduced before. Step 3 is the standard sampling approach for indicator variables in mixture models and it consists of sampling from which of the finite mixture components verifying $(u_i < \rho_s)$ comes each observation. Finally, step 4 is similar to step 1 from the previous algorithm and can be undertaken similarly.

Finally, given a posterior sample from this algorithm, the predictive density can be approximated using the sampling ideas by Walker (2007) as follows. At each iteration, sample a uniform $U(0,1)$ variable, v, and select the mixture component, s', such that, $\sum_{s=1}^{s'-1}\rho_s < v < \sum_{s=1}^{s'}\rho_s$. Then, consider the parameters, $\mathbf{y}_{s'}$, for this mixture component. Thus, the predictive density is approximated with:

$$\frac{1}{T}\sum_{t=1}^{T}\rho_{s'}^{(t)}\prod_{r=1}^{m}\beta\left(x_r|z_{k^{(t)}}\left(y_{s'r}^{(t)}\right),k^{(t)}-z_{k^{(t)}}\left(y_{s'r}^{(t)}\right)+1\right), \tag{2.26}$$

where T denotes the size of the posterior sample, $\rho_{s'}^{(t)}$ the weight of the sampled mixture component at the t-th iteration, $y_{s'r}^{(t)}$ are the elements of $\mathbf{y}_{s'}^{(t)}$, which are the parameter of the sampled mixture component at the t-th iteration and, finally, $k^{(t)}$ is the value of the polynomial order at the t-th iteration.

Observe that the main advantage in the estimation of this predictive density compared to (2.20) is that it is not required to evaluate the posterior distribution for the k^m weights, $\boldsymbol{\omega}_{k^m}$, in the Bernstein polynomial which, besides being a very large number of elements to sample, there may be a lot of weights with very small values when the data is sparse, leading to numerical problems with the algorithm.

2.5 Simulations and empirical applications

In this Section, we undertake several simulation studies and a real data example to illustrate the performance of the proposed nonparametric Bayesian approach. For simplicity in the visualization, we only consider examples in the two-dimensional case in order to better illustrate the accuracy in density estimation.

In all cases, we impose the following noninformative prior assumptions. For the baseline distribution, F_0, we assume a uniform distribution on $[0,1]^2$. We also set the smoothing parameter to be $M = 1$, as suggested in Petrone (1999b), in order to express a small degree of belief in the prior guess. Finally, we assume the following hierarchical prior structure for the Bernstein polynomial degree k:

$$\begin{aligned} k-1 \mid \lambda &\sim Poisson(\lambda), \\ \lambda &\sim Gamma(a,b). \end{aligned}$$

Observe that this prior structure is consistent with the assumptions in theorems 3.1, where it is required a strictly positive prior probability for all possible values of k. Also, in order to avoid a prior sensitivity to the choice of λ, we further assume a Gamma hyperprior where we might set for example $a = 1$ and $b = (n^{\frac{1}{3}} - 1)$, where n is the sample size. This implies that the prior mean for k is $n^{\frac{1}{3}}$, which is the value suggested for k in Sancetta and Satchell (2004) in the bivariate case.

The proposed MCMC algorithm described in Section 4 is run using $100\,000$ iterations and discarding the first $50\,000$ as burn-in iterations.

2.5.1 Simulated data

Firstly, we consider simulated data from the bivariate beta distribution proposed in Olkin and Liu (2003). This is a continuous variable with support on the unit square and it is a generalization of the univariate beta distribution function to the bivariate case. The bivariate beta distribution, $\beta_2(x, y; a, b, c)$, is derived by considering the joint distribution of two random variables:

$$X := \frac{U}{U+V}, \quad Y := \frac{V}{V+W},$$

where U, V and W are three independent standard gamma distributions with respective shape parameters, a, b and c. Clearly, the marginal distributions of X and Y are beta distributions, $\beta(x; a, c)$ and $\beta(y; b, c)$, respectively. This model can describe a wide range of densities on the unit square and can be easily generalized to the multivariate case.

Figure 2.1 shows the estimated predictive density, using (2.26), and true density for 200 data points simulated from a bivariate beta distribution $\beta_2(x, y; 5, 10, 10)$. The predictive and true marginal densities are also shown. We can see that the proposed Bayesian density estimation method based on Bernstein polynomials and using slice sampling provide a good fit to the data.

We have also compared these results with those obtained using the multivariate extension of the hybrid Monte Carlo algorithm of Petrone (1999a,b) described at the beginning of Section 4. As expected, the predictive distributions are almost identical to that shown in Figure 2.1. However, we have observed some differences in the mixing performance of the two algorithms as illustrated in Figure 2.2, where the trace plots and histograms of the posterior sample for k obtained with both algorithms are shown. Observe that the proposed slice sampling method seems to better explore the state space and, in particular, the tails of the posterior distribution. Nevertheless, the main difference between these two approaches is the computational cost. For this example, the computing time was of one hour and a half using the proposed slice sampling algorithm, while the hybrid Monte Carlo method took four hours using in both cases self programmed code in R 2.15.2 (R Development Core Team 2011) on a computer with a 3.4 Ghz core.

In order to illustrate the flexibility of the model, we now consider 200 simulated data from a mixture of bivariate beta densities $\beta_2(x, y; 5, 10, 10)$ and $\beta_2(x, y; 5, 1, 5)$ with equal weights. Using the same prior specifications as before, the proposed slice sampling algorithm is run for these

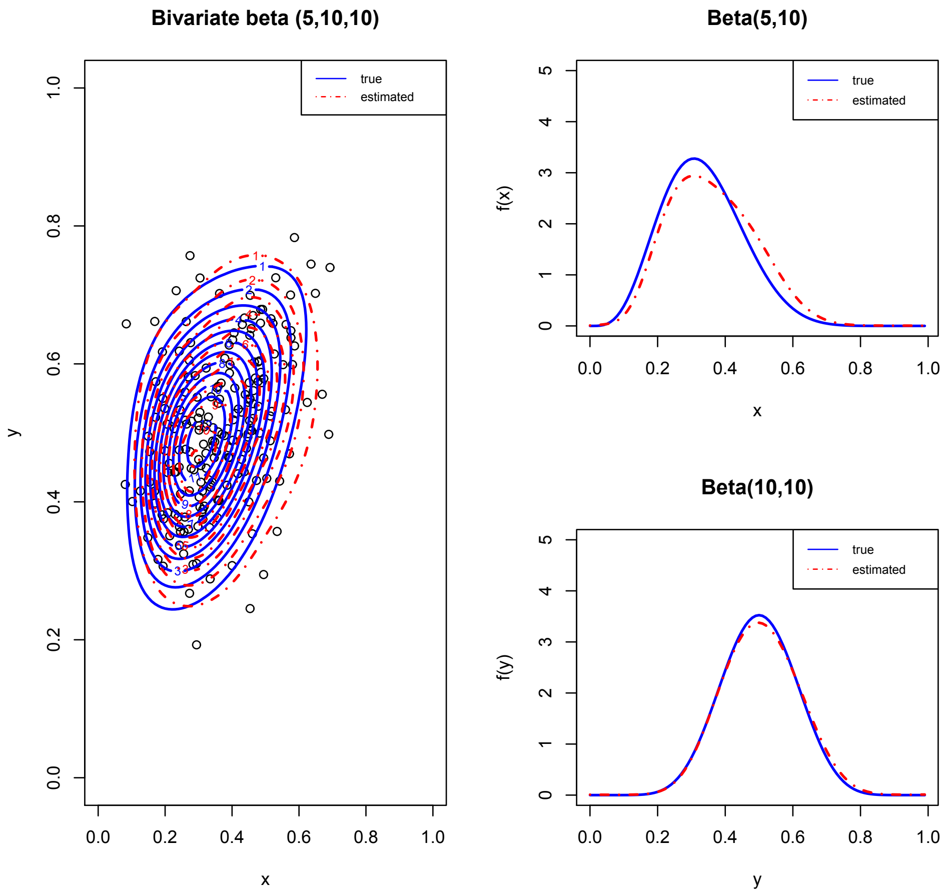

Figure 2.1: Predictive and true densities obtained from 200 simulated data from a bivariate beta distribution $\beta_2(x, y; 5, 10, 10)$ (left) and marginal distributions (right) using the proposed slice sampling algorithm.

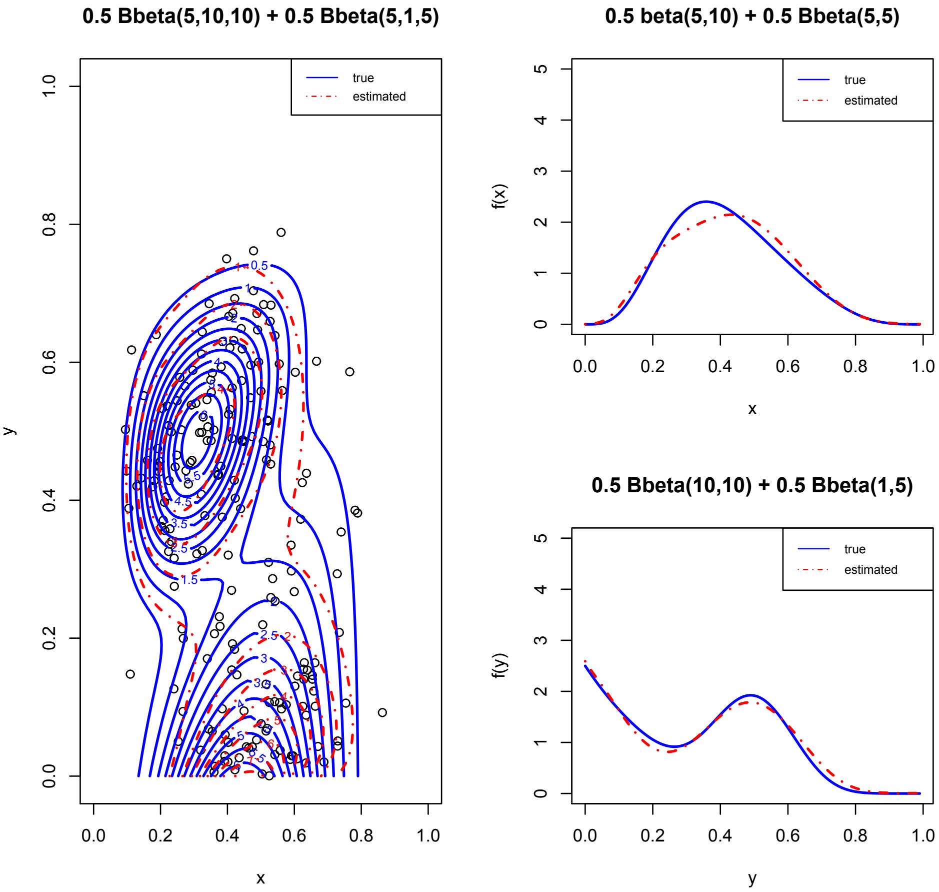

Figure 2.2: Trace plots and histograms of the posterior sample for k using the proposed slice sampling algorithm and the multivariate extension of the hybrid Monte Carlo algorithm by Petrone (1999a,b) obtained from simulated data from a bivariate beta distribution.

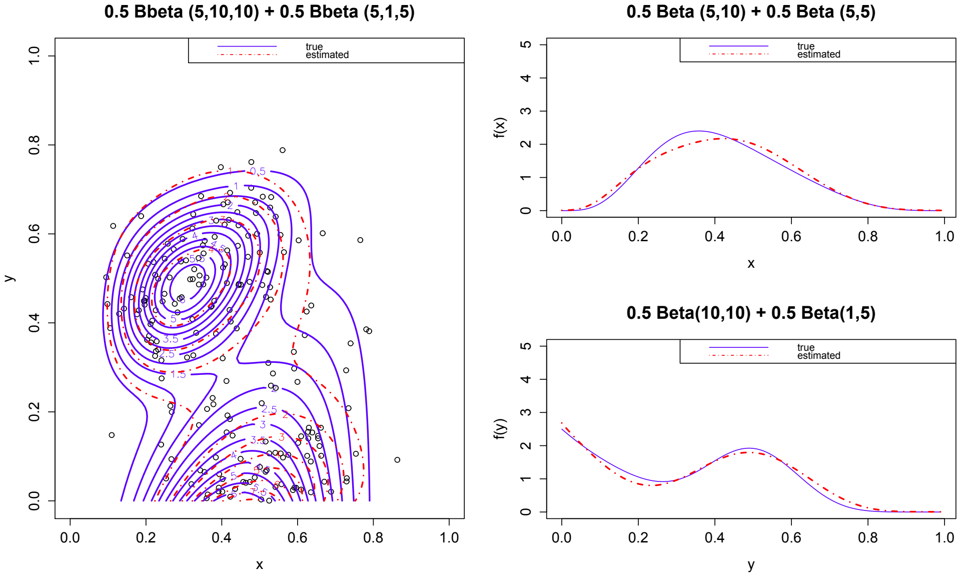

Figure 2.3: Predictive and true densities obtained from 200 simulated data from a mixture of bivariate beta distributions $\beta_2(x, y; 5, 10, 10)$ and $\beta_2(x, y; 5, 1, 5)$ with equal weights (left) and marginal distributions (right) using the proposed slice sampling algorithm.

data using the same number of iterations. The true and predictive joint densities are illustrated in Figure 2.3. The marginal predictive and true densities corresponding to mixtures of univariate beta distributions are also shown. We can observe that also in this mixture case there is a good fit to the true densities.

As before, we have compared these results with those obtained using the hybrid Monte Carlo method leading to similar predictive densities. However, in this case the differences in the mixing performance of the algorithm are more pronounced as shown in Figure 2.4. Also, while the slice sampling algorithm required less than two hours and a half, the hybrid Monte Carlo took more than three hours and a half.

Finally, we have also tried alternative prior specifications. In general, there is little sensitivity of the density estimations to the choice of the concentration parameter M. Similar to the univariate case in Petrone (1999b), the predictive densities get somewhat closer to the uniform prior distribution for larger values of M. On the other hand, as we would expect, there is slightly more sensitivity

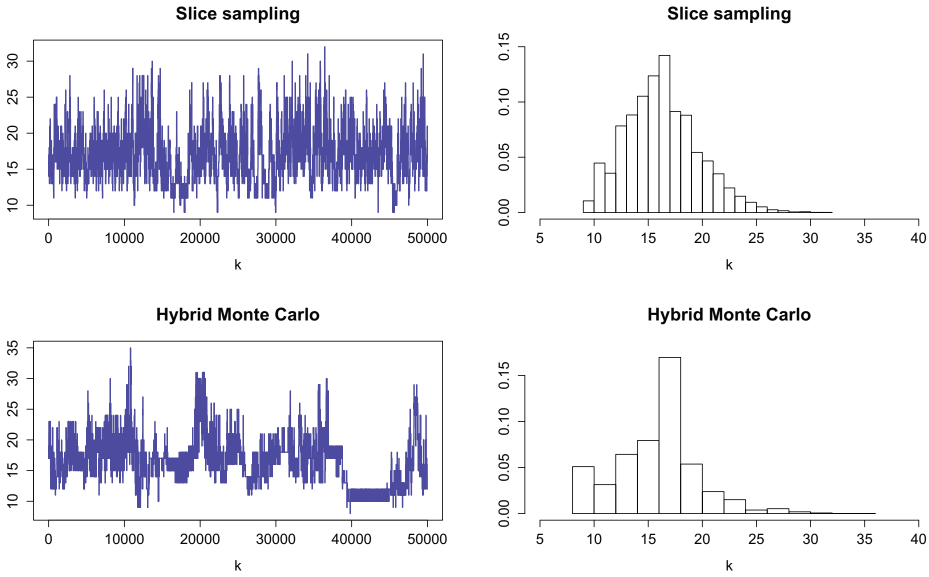

Figure 2.4: Trace plots and histograms of the posterior sample for k using the proposed slice sampling algorithm and the multivariate extension of the hybrid Monte Carlo algorithm by Petrone (1999a,b) obtained from simulated data from a mixture of two bivariate beta distributions.

to the prior specification for k. We have observed that using prior distributions for k concentrated on small values, such as a Poisson prior distribution with small mean, lead to smoother predictive densities than using a uniform prior on a closed interval, as observed for the univariate Bernstein model in Petrone (1999b). As noted earlier, k plays a similar role in the Bernstein polynomial to the bandwidth in kernel density estimation. This is the main reason to define a hierarchical prior structure for k as introduced at the beginning of this Section.

2.5.2 Real data example

In this Section, we illustrate the proposed Bayesian density estimation method based on Bernstein polynomials and using slice sampling to examine the relationship between the percentage of forest area (% of land area) and percentage of agricultural nitrous oxide emissions (% of total) in 127 countries in 2010. The data are available from http://data.worldbank.org/. Nitrous oxide is naturally present in the atmosphere, however, human activities in agriculture such as fertilizer use and waste and savannah burning are increasing the amount of this gas in the atmosphere. The impact of nitrous oxide emissions on warming the atmosphere is over 300 times that of carbon dioxide per unit weight. Therefore, it is interesting to examine the influence of the percentage forest area in the reduction of these emissions.

Figure 2.5 shows the scatter plot of these data together with the estimated joint density using the proposed Bernstein polynomial model with the same prior assumptions and MCMC iterations as in the simulation examples. We can observe that the model identifies three main clusters of countries. Firstly, there is a large group corresponding to those countries with more than 10% of forest area where there is a clear negative relationship between the percentage of forest area and the nitrous oxide emissions. Secondly, there is a fairly large group with less than 10% of forest area but comparatively a large percentage of nitrous oxide emissions. Finally, there is a small group of countries with a low percentage of forest area and a low percentage of nitrous oxide emissions.

Finally, Figure 2.6 shows the estimated marginal distributions of the percentage of forest area and the percentage of nitrous oxide emissions. We can observe that the distribution of the percentage of forest area has two modes, one is zero and the other is close to 0.4. The nitrous emissions percentage distribution is left-skewed with a mode close to 0.8. It seems that the model is flexible enough to capture adequately the different shapes of these distributions.

2.6 Conclusions and extensions

In this Chapter, we have extended the Bernstein-Dirichlet prior introduced in Petrone (1999a,b) for densities on a closed interval to the multivariate case and have obtained the convergence rate of the associated posterior distribution. Moreover, we have introduced a new algorithm for sampling from the posterior distribution. Various extensions are possible.

Firstly, although here we have defined the multivariate Bernstein polynomial using a single k, in principle it is possible to consider different values $k_1, \ldots, k_m$ for the different components of $\mathbf{x}$. This might be useful from a practical viewpoint if some variables are more spread than others.

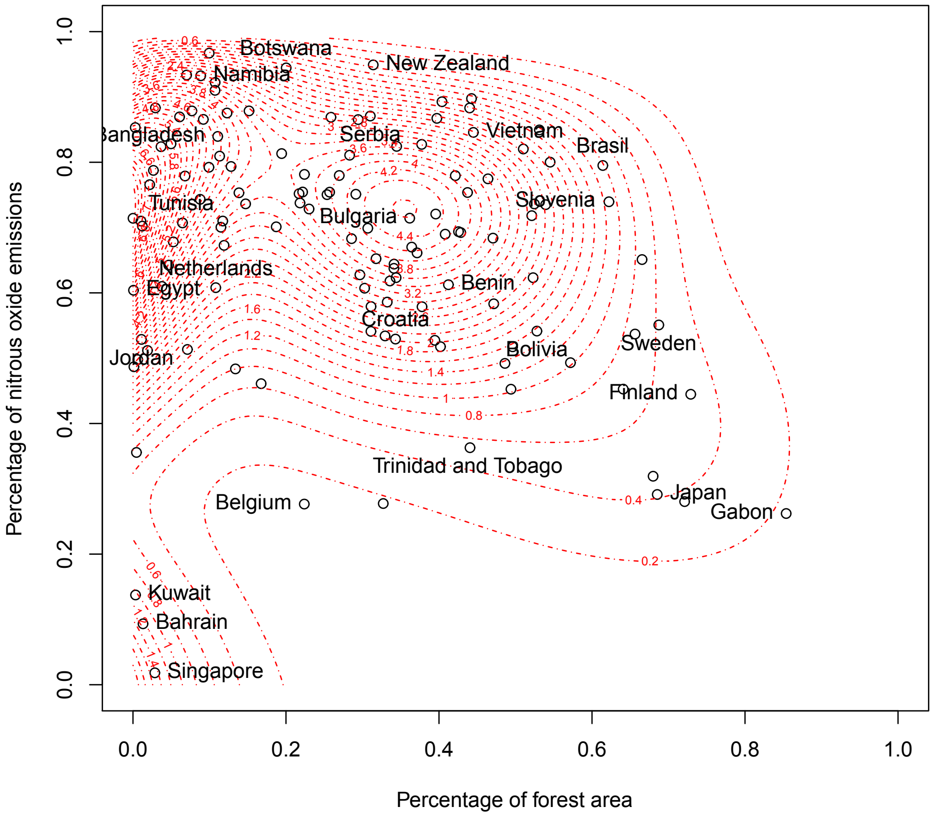

Figure 2.5: Predictive joint density of the percentage of forest area and percentage of agricultural nitrous oxide emissions obtained from a data base of 127 countries in 2010.

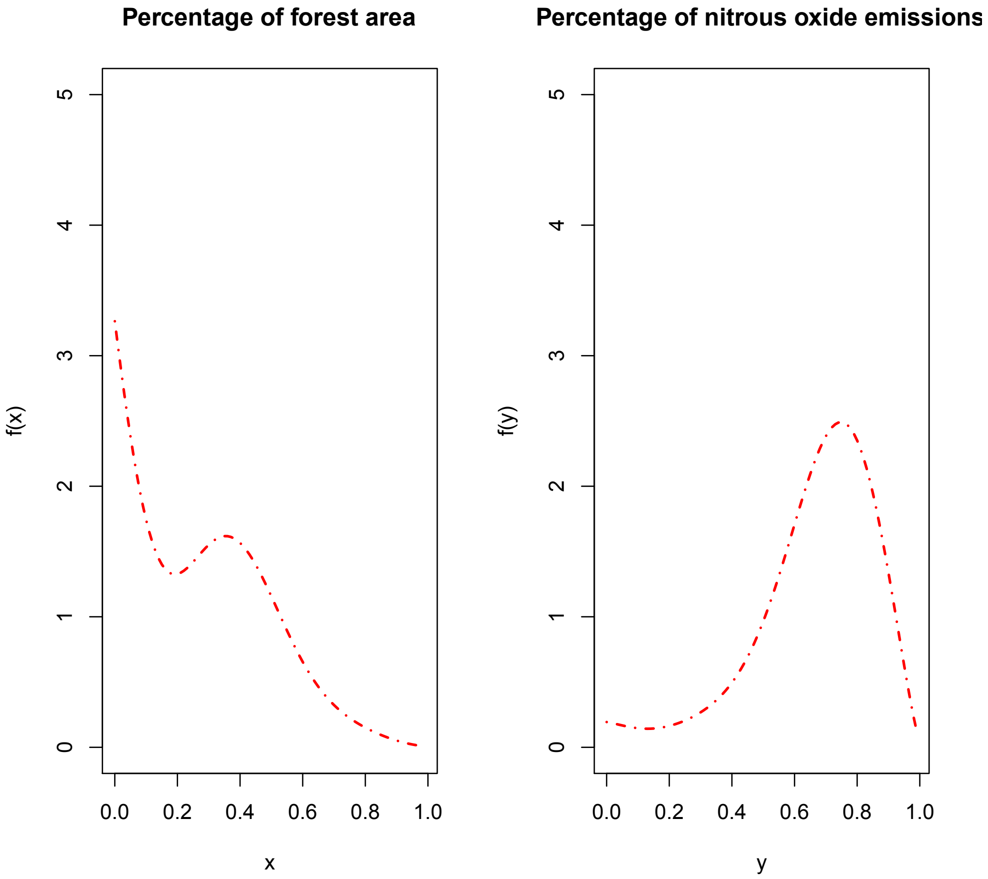

Figure 2.6: Marginal estimated distribution of the percentage of forest area (left) and percentage of agricultural nitrous oxide emissions (right) obtained from a data base of 127 countries in 2010.

Secondly, following Tenbusch (1994), it would be interesting to consider multivariate Bernstein densities on the triangle which might be more appropriate for modeling the joint density of various proportions of the same quantity. Finally, the multivariate Bernstein polynomial provides an asymptotic model for a copula, see e.g. Sancetta and Satchell (2004) so that it can be used to model the dependence structure of a multivariate distribution. Then the use of a Bernstein-Dirichlet prior for the copula could be combined with standard, Bayesian nonparametric priors for the marginals to provide a general, nonparametric approach to multivariate data modeling. Work is in progress on these problems.

Chapter 3

Bayesian linear regression with conditional heteroskedasticity

3.1 Introduction

We consider Bayesian estimation of the linear regression model that imposes conditional moment restrictions. A useful framework like $E(Y|X) = X'\beta_0$ or $Y = X'\beta_0 + \varepsilon$, $E(\varepsilon|X) = 0$ is widely formulated to analyze a number of statistical and econometric models. It is well-known that the procedure of estimating the parameters of interest could be expected to be efficient provided more information about the conditional error distribution is known. In this Chapter, we propose a Bayesian semiparametric method for consistent estimation of the regression coefficients and the standard deviation function when the error term is subject to a normal distribution with associated variance that is dependent on covariates.

The primary purpose of this Chapter is to investigate the asymptotic frequentist properties of the corresponding posterior distribution by putting a prior on the regression coefficients and the standard deviation in this linear model. An analysis of the asymptotic behavior of Bayesian methods in infinite-dimensional statistical models is important, such as posterior consistency, rate of posterior convergence, rate-optimality and adaptation properties and Bernstein-von Mises phenomenons, which reflect a sense of Bayesian robustness, namely that the prior does not have an impact on the posterior distribution too much when the amount of information collected in the data or the number of observations grows indefinitely.

In recent years, there has been substantial research in Bayesian nonparametrics on the development of this mathematical, asymptotical theory for a wide range of statistical models, see, for example, Ghosal et al. (1999, 2000); Ghosal and van der Vaart (2001, 2007b,a), to name a few. However, it has been studied very little in the linear models with predictor-dependent error densities. Norets (2015) established a semiparametric version of Bernstein-von Mises theorem under misspecification: the posterior credible regions of the regression coefficients are asymptotically equivalent to the frequentist ones and also this posterior inference is efficient even though the data generating process is not normal. Pelenis (2014) considered the kernel stick-breaking mixtures to model the conditional error distribution and demonstrated posterior consistency of the conditional error density and the finite regression coefficients for these kernel mixture priors. Also, Wang (2013) studied posterior consistency for the heteroscedastic nonparametric regression models by relaxing the assumptions of linearity in the model, with a substitution of an unknown, smooth regression function. There is a noticeable absence of rate adaptation results in these regression settings.

In this Chapter, we plug this gap and take up the investigation of this rate adaptive procedure, in order to provide a theoretical underpinning of the Bayesian approach to explore the possible accuracy at maximum capacity and assess the well-balanced spread of the underlying prior distribution across a continuum of regularities of the functions considered. Adaptive convergence rates for Bayesian nonparametric estimation in various statistical models have been established by Huang (2004), Scricciolo (2006), Belitser and Ghosal (2003), van der Vaart and van Zanten (2009), Rousseau (2010), Kruijer et al. (2010), de Jonge and van Zanten (2010, 2012), Shen and Ghosal (2012), Shen et al. (2013), Norets and Pati (2014) and Belitser and Serra (2014), among others.

A broad class of priors have been explored to yield adaptation across all smoothness levels. Recently, priors based on splines have received much attention to the construction of probability distribution on infinite-dimensional spaces. Various groups of researchers have worked with univariate splines or its corresponding tensor-product splines in the multivariate case as a useful block to construct a prior. For example, Huang (2004) built a prior on the discrete mixture of splines to develop a theorem on adaptive convergence rates in the context of regression and density estimation. de Jonge and van Zanten (2012) discussed priors on multivariate functions by choosing an appropriate probability distribution on the partition size and Gaussian prior on B-spline coefficients in the tensor-product B-spline expansions. Shen and Ghosal (2012) constructed a prior using finite random splines with a prior distribution on the number of terms. Belitser and Serra (2014) investigated an extension of these results involving spline-based priors by endowing a probability distribution on the location of the knots instead of assuming them to be equally spaced. This enables us to build a wide spectrum of priors on the conditional standard deviation of the regression error terms. It is widely known that the posterior distribution contracts at a rate of the order $n^{-\alpha/(2\alpha+d)}$ (up to an additional logarithm factor) for a α-smooth functions of d-variables, which agrees with the optimal rate of the estimators in the frequentist context. In other words, a fully rate-adaptive procedure can be obtained across all smoothness levels if that holds. One possible explanation of this phenomenon is that there is a sufficiently large amount of prior mass around the function of interest with total smoothness levels. We will show that the corresponding posterior converges at the optimal rate up to a logarithm factor without a priori knowledge of the smoothness levels of the conditional standard deviation.

From the practical point of view, diverse algorithms for normal linear regression where the error variance depends on the predictors have been exhibited in Yau and Kohn (2003) and Chib and Greenberg (2013) which considered transformed splines to model the variance and Goldberg et al. (1997) where a transformed Gaussian process prior was considered. Markov Chain Monte Carlo simulations carried out in these papers performed well in these models with flexible specifications for error variance. Here we center on the theoretical aspects in Bayesian normal regression models.

The Chapter is organized as follows. In Section 2 we give a general overview of the notation and a brief outline of the model we are studing. In Section 3 we provide a preliminary review on the notions of spline functions, univariate B-splines and tensor-product B-splines as well as its associated approximation properties. In Section 4, we show that the optimal posterior convergence rate can be achieved using two types of spline priors: one based on conditional Gaussian tensor-product spline prior or a hierarchical Gaussian spline prior, and the other built on log-spline prior that stems from finite random spline expansion with a random number of terms. We conclude with a brief discussion and some technical lemmas, all containing proofs as well as auxiliary theorems are delegated to the Appendix.

3.2 General model setup

In this Section, we take a detailed description of the notation and then describe our model.

3.2.1 Notation

For any $a,\, b \in \mathbb{R}$, denote $\lfloor a \rfloor$ to be the largest integer strictly smaller than a. Similarly, define $\lceil a \rceil$ to be the smallest integer which is strictly greater than a. We write $a \vee b,\ a \wedge b$ to stand for the maximum and the minimum between a and b respectively. Set $a_+ = a \vee 0$.

Let $\eta = (\beta, \sigma)$ and the true value $\eta_0 = (\beta_0, \sigma_0)$. Denote the conditional density function $N(\beta, \sigma^2(\boldsymbol{x}))$ by $f_{\boldsymbol{x}\eta}$ and let $f_{\boldsymbol{x}\eta_0}$ be the true conditional density function $N(\beta_0, \sigma_0^2(\boldsymbol{x}))$. The Kullback-Leibler divergence between η and η_0 in this case is then defined as,

$$K(\eta, \eta_0) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f_{\boldsymbol{x}\eta_0}(y) \log \frac{f_{\boldsymbol{x}\eta_0}(y)}{f_{\boldsymbol{x}\eta}(y)}\, dy\, dG_0(\boldsymbol{x}), \tag{3.1}$$

$$V(\eta, \eta_0) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f_{\boldsymbol{x}\eta_0}(y) \left(\log \frac{f_{\boldsymbol{x}\eta_0}(y)}{f_{\boldsymbol{x}\eta}(y)} \right)^2 dy\, dG_0(\boldsymbol{x}), \tag{3.2}$$

where $\mathcal{X}, \mathcal{Y}$ are the domains that will be specified later and $G_0(\cdot)$ is a general distribution function. The ε-Kullback-Leibler neighborhood around η_0 for any $\varepsilon > 0$ is expressed as,

$$K_\varepsilon(\eta_0) = \{\eta : K(\eta, \eta_0) < \varepsilon\}. \tag{3.3}$$

We define the Hellinger metric between η and η_0 in this case as,

$$d_H(\eta,\eta_0) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \left(\sqrt{f_{\boldsymbol{x}\eta}(y)} - \sqrt{f_{\boldsymbol{x}\eta_0}(y)} \right)^2 \, dy \, dG_0(\boldsymbol{x}). \tag{3.4}$$

We use the natural L^2-norm with respect to the distribution function $G_0(\cdot)$ to measure the distance between η and η_0:

$$d_2(\eta,\eta_0) = \left\{ \int_a^b \left([(\beta-\beta_0)^T \boldsymbol{x}]^2 + [\sigma(\boldsymbol{x}) - \sigma_0(\boldsymbol{x})]^2 \right) \, dG_0(\boldsymbol{x}) \right\}^{1/2}, \tag{3.5}$$

and denote the neighborhood of η_0 with respect to the distance function $d_2(\eta,\eta_0)$ as follows:

$$U_\varepsilon(\eta_0) = \left\{ (\beta,\sigma) : \int_a^b \left([(\beta-\beta_0)^T \boldsymbol{x}]^2 + [\sigma(\boldsymbol{x}) - \sigma_0(\boldsymbol{x})]^2 \right) \, dG_0(\boldsymbol{x}) > \varepsilon \right\}. \tag{3.6}$$

We use the notation $\lesssim$ to stand for somewhat inequality up to a constant. To compare two functions, for example, g_1, g_2, we denote $g_1 \lesssim g_2 \lesssim g_1$ by $g_1 \asymp g_2$. Let $\|\cdot\|_2$ and $\|\cdot\|_\infty$ denote the Euclidean norm and supremum norm respectively.

We now take a brief account of the definitions in the context of multivariate functions, especially describe the appropriate notions of smoothness in this multivariate case. Let's denote the space of continuous functions f on $[0,1]^d$ by $C\left([0,1]^d\right)$, equipped with the supremum norm $\|f\|_\infty$. For a multi-index $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$, let the sum $|\alpha| = \sum_{i=1}^d \alpha_i$ and the mixed partial derivative operator is defined as,

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}. \tag{3.7}$$

For $\alpha > 0$, the Hölder space $C^\alpha\left([0,1]^d\right)$ stands for the collection of functions f on $[0,1]^d$ with mixed partial derivative $D^r f \in C\left([0,1]^d\right)$ of all orders up to $|r| \leq \lfloor \alpha \rfloor$ satisfying,

$$|D^r f(\boldsymbol{x}) - D^r f(\boldsymbol{z})| \leq C \|\boldsymbol{x} - \boldsymbol{z}\|_2^{\alpha - \lfloor r \rfloor}, \tag{3.8}$$

for some positive constant C, each $\boldsymbol{x}, \boldsymbol{z} \in [0,1]^d$. Meanwhile, denote the norm on the Hölder class

$C^\alpha\left([0,1]^d\right)$ by,

$$\|f\|_{C^\alpha\left([0,1]^d\right)} = \|f\|_\infty + \sum_{r:\,|r|=\lfloor\alpha\rfloor} \|D^r f\|_\infty. \tag{3.9}$$

3.2.2 Restricted moment models

Suppose we observe a real-valued sample $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ where X_i is a d-dimensional covariate, Y_i is the response variable and $(X_i, Y_i) \sim P_0$ for $i = 1, 2, \dots, n$. The data generating process satisfies $Y|X = \boldsymbol{x} \sim N(\boldsymbol{x}'\beta_0,\, \sigma_0^2(\boldsymbol{x}))$ for some unknown true parameter $\beta_0 \in \Theta \subset \mathbb{R}^d$, unknown true conditional variance function σ_0^2 and all $\boldsymbol{x} \in \mathscr{X} = [0,1]^d$. In other words, this linear model could be described as,

$$Y_i = X_i'\beta_0 + \varepsilon_i, \quad i = 1, 2, \dots, n. \tag{3.10}$$

where error variables $\varepsilon_i|X_i = \boldsymbol{x}_i \sim N(0,\, \sigma_0^2(\boldsymbol{x}_i))$ for all $\boldsymbol{x}_i \in [0,1]^d$, $i = 1, 2, \dots, n$. In this semiparametric model, the unknown parameters are $(\beta,\, \sigma(\cdot))$ where the finite-dimensional parameter β is of interest and $\sigma(\cdot)$ is the infinite-dimensional nuisance parameter. Crainiceanu et al. (2007) studied a more general model with heteroscedastic errors using the penalized splines for the regression part. Our model could be rewritten as $(\Theta \times \mathcal{M},\, \mathscr{B} \times \mathscr{F})$ equipped with Borel σ-algebras $\mathscr{B}$ and $\mathscr{F}$ on Θ and $\mathcal{M}$ respectively, where,

$$\mathcal{M} = \{\sigma(\cdot) :\, [0,1]^d \to (\underline{\sigma}, \overline{\sigma})\}. \tag{3.11}$$

is a polish space on $\mathscr{X}$ and also is assumed to contain the true conditional standard deviation σ_0, $0 < \underline{\sigma} < \overline{\sigma} < \infty$. Let Π denote the total prior for the pair $(\beta,\, \sigma)$ on $(\Theta,\, \mathcal{M})$ which is defined by $\Pi(d\beta, d\sigma) = \Pi_\beta(d\beta) \times \Pi_\sigma(d\sigma)$ where Π_β and Π_σ are corresponding independent priors on β and σ respectively. Here we leave the distribution of covariates denoted by $G_0(\cdot)$ unspecified since it is ancillary and also of our interest is to focus on the conditional distribution. The corresponding posterior distribution for (β, σ) given the data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is denoted by,

$$\Pi(\cdot|(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)).$$

In view of Bayes' theorem, the posterior is given by,

$$\Pi(B|(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)) = \frac{\int_B L(\beta, \sigma; (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))\, \Pi(d\beta, d\sigma)}{\int L(\beta, \sigma; (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))\, \Pi(d\beta, d\sigma)}, \tag{3.12}$$

where the likelihood function $L(\beta, \sigma; (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$ could be written as,

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma(X_i)} \exp\left(-\frac{(Y_i - X_i'\beta)^2}{2\sigma^2(X_i)}\right). \tag{3.13}$$

Usually the posterior mean can be regarded as a Bayesian estimator of the unknown pair (β_0, σ_0). If this Bayesian estimator is consistent, the further concern is then of interest to consider the finer aspects of this posterior distribution or quantify the rate at which it contracts around the true unknown parameter, namely, posterior convergence rate. More precisely, for a given positive sequence (ε_n) going to zero, the posterior distribution is said to converge to the Dirac-mass at (β_0, σ_0) at the rate ε_n, if, as $n \to \infty$,

$$\Pi\big\{(\beta, \sigma) : \, d_H((\beta, \sigma), (\beta_0, \sigma_0)) > M\varepsilon_n \,\big|\, (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\big\} \longrightarrow 0 \text{ in } P_0^n\text{-probability}, \tag{3.14}$$

for a sufficiently large $M > 0$. Here this assertion of the definition is in-probability statement that holds under the true distribution typically governed by the true parameter pair (β_0, σ_0).

The main objective is to construct some priors for $\Theta \times \mathcal{M}$ to show that the corresponding posterior converges at an optimal rate at $(\beta_0, \sigma_0(\cdot)) \in \Theta \times \mathcal{M}$. Here the prior does not depend on the information about the unknown smoothness levels of the true conditional standard deviation function $\sigma_0(\cdot)$. So the so-called rate-adaptive procedure is obtained across all the regularity levels.

3.3 A preliminary introduction to Splines

In this Section, we will provide a general overview on spline function supported on unit hypercube following by a brief introduction on the splines defined on the unit interval $[0, 1]$. More extensive treatment on this subject could be found in Schumaker (2007).

3.3.1 Spline function on the unit interval

A spline function on $[0, 1]$ is essentially viewed as a generalization of the polynomial function on the unit interval. It is a piece polynomial function but enjoy the properties of global smoothness on its domain.

More specifically, let q, K be two fixed natural numbers and partition the unit interval $[0, 1]$ into K equally spaced subintervals $[(k-1)/K, k/K]$ for $k = 1, 2, \ldots, K$. Consider a spline function with the order q greater than 2, that is, all polynomials with its domain coinciding with one of those subintervals are of the degree smaller than $q - 1$ and this spline function is globally $q - 2$ times continuously differentiable on $[0, 1]$.

Let S_K be the collection of all splines of order q with simple knots at the points $\{k/K : k = 1, \ldots, K - 1\}$. It can be seen that S_K forms a $J = (q + K - 1)$-dimensional linear space. The so-called B-splines $B_1^K, B_2^K, \ldots, B_J^K$, which can be found in de Boor (2001), are used to give a convenient basis in this space. The concrete function forms of these B-splines are negligible to us. The primary properties of these B-splines closely used in this Chapter are that B-splines are always nonnegative, each basis function is supported on a tiny interval with its length at most q/K and the sum of all B-splines evaluated at any given point in the domain is equal to one. In other words, they constitute a partition of unity, i.e.

$$\sum_{i=1}^{J} B_i^K(x) = 1,$$

for each $x \in [0, 1]$.

3.3.2 Tensor-product spline on $[0, 1]^d$

In this Subsection we introduce spline functions on multi-dimensional domains with the help of multivariate polynomials. The construction of the linear space of such multivariate splines relies heavily on the spline space S_K in the unit interval described above. In fact, this linear space on $[0, 1]^d$ is a tensor-product of the univariate linear space on $[0, 1]$. More precisely, a unique direction denoted by a variable is assigned to each linear space in the tensor-product and then we obtain the multivariate polynomials supported on some tiny rectangles by taking the multiplication of polynomials with respect to one single variable defined on some small intervals.

Accordingly, the convenient basis for the linear space of tensor-product splines is the tensor-product B-splines, which equal to the products of the corresponding B-splines on $[0, 1]$. Hence the tensor-product space has dimension $(q+K-1)^d$, for example, in the construction of the space S_K defined above. The advantage of introducing the tensor-product B-splines is that they inherit the nice properties that univariate B-splines have as we shall see below.

In what follows, we consider a d-fold tensor-product space $\mathcal{S}_K = S_K \otimes \cdots \otimes S_K (d \text{ times})$ of tensor-product splines defined on the unit cube $[0, 1]^d$, that is partitioned equally into m^d cubes $I_{k_1} \times \cdots \times I_{k_d}$. A function $s : [0, 1]^d \to \mathbb{R}$ is defined to be a tensor-product spline in $\mathcal{S}_K$ if for each such tiny cube, s possesses the following multivariate polynomial form,

$$\sum_{k_1=0}^{q-1} \cdots \sum_{k_d=0}^{q-1} c_{k_1 \dots k_d} \, x_1^{k_1} \cdots x_d^{k_d}. \tag{3.15}$$

As was the case in the univariate spline space, the basis in $\mathcal{S}_K$ is provided by the so-called tensor-product B-splines as follows,

$$B^K_{j_1 \dots j_d}(x_1, \dots, x_d) = B^K_{j_1}(x_1) B^K_{j_2}(x_2) \cdots B^K_{j_d}(x_d). \tag{3.16}$$

It can be shown that $\mathcal{S}_K$ has dimension $(q + K - 1)^d$ and these multivariate B-splines also form a partition of unity,

$$\sum_{j_1=1}^{J} \cdots \sum_{j_d=1}^{J} B^K_{j_1 \ldots j_d}(x_1, \ldots, x_d) = 1, \tag{3.17}$$

for all $x_i \in [0, 1]$, $i = 1, 2, \ldots, d$.

3.3.3 Approximation properties of tensor-product B-splines

It is well-known that the univariate B-splines in the space S_K could approximate any function of interest in $C^\alpha[0, 1]$, for example, at the rate $J^{-\alpha}$ where $J = q + K - 1$. In other words, any function with a smoothness level α in $C^\alpha[0, 1]$ could be approximated by a couple of B-splines, $B^K_1, B^K_2, \ldots, B^K_J$ with its associated approximation error controlled by the order $J^{-\alpha}$.

This idea also works in the multivariate case. How well tensor-product B-splines approximate the generic function is uniquely determined by the target function's smoothness level α and the dimension of the linear space $\mathcal{S}_K$ induced by the tensor-product B-splines if the order q of the splines is chosen to be larger than the smoothness level α. The approximation ability in terms of tensor-product B-splines is stated in the following lemma which provides an upper bound of the approximation error with respective to the uniform distance.

LEMMA 3.1 (Shen and Ghosal (2014)) *Let q, d, $K \in \mathbb{N}$, $\alpha \in \mathbb{R}$, $\alpha \leq q$, $J = q + K - 1$. For any function $f \in C^\alpha\left([0, 1]^d\right)$, there exist $\theta = (\theta_{00\ldots0}, \ldots, \theta_{JJ\ldots J}) \in \mathbb{R}^{J^d}$ and a positive constant C_1 that only depends on q, d and α such that,*

$$\left\| f - \sum_{j_1=1}^{J} \cdots \sum_{j_d=1}^{J} \theta_{j_1 \ldots j_d} B^K_{j_1 \ldots j_d}(x_1, \ldots, x_d) \right\|_\infty \leq C_1 J^{-\alpha} \|D^\alpha f\|_\infty. \tag{3.18}$$

Furthermore, if $f > 0$, then each element of θ could be chosen to be positive for a sufficiently large J.

3.4 Adaptive posterior contraction results

Splines possess excellent approximation capabilities for smooth functions in the previous Section, where the approximation error is completely controlled by the dimension of the spline space and the smoothness level. More precisely, the error becomes smaller if the dimension grows and the objective function is smoother. From the frequentist view of point, Stone (1994) showed that the maximum likelihood estimator of the function in $C^{\alpha}([0,1]^d)$ achieves the rate of convergence $n^{-\alpha/(2\alpha+d)}$. As indicated in de Jonge and van Zanten (2012), a Bayesian estimator for probability densities or the regression functions in multivariate domains under weaker conditions also attained the optimal contraction rate $n^{-\alpha/(2\alpha+d)}$. Simultaneously, they established that a type of Gaussian process prior could yield the near-optimal adaptive posterior convergence rate, up to an additional logarithmic factor when α is unknown.

In the next two Subsections, we consider spline-based priors for $\sigma(\cdot)$ in a variety of means. In Subsection 3.4.1, we build a hierarchical Gaussian spline prior by putting Gaussian prior weights on the coefficient and adding another hierarchical layer for the partition size involved in the tensor-product B-splines. It follows that this hierarchical procedure achieves a near-optimal adaptive contraction rate. Alternative log-spline priors with finite random tensor-product splines and a random number of terms that also achieve the optimal adaptive rate of convergence will be demonstrated in Subsection 3.4.2.

Throughout this Section, we consider the following condition on Π_{β} :

(A1) Its support is $[\underline{\beta}, \overline{\beta}]$, where $\underline{\beta} < \overline{\beta}$ and $\underline{\beta}, \overline{\beta} \in (-\infty, \infty)$. For all $\varepsilon > 0$, there exists $m_1 > 0$ such that,

$$\Pi(\|\beta - \beta_0\|_2 \leq \varepsilon) \geq \exp(-m_1 d \log(1/\varepsilon)). \tag{3.19}$$

In fact, this is a mild assumption on the prior of β. And several ordinary distribution examples satisfy (3.19). More detailed and similar examples could be found in the discussion of the prior for weights vector θ in Subsection 3.4.2.

3.4.1 Hierarchical Gaussian spline prior

In this Subsection, a class of Gaussian process, whose sample path is defined by tensor-product splines extensively discussed in the preceding Section, will be used for the construction of priors on the conditional standard deviation in the linear model.

Let $Z_{00\dots0},\dots,Z_{JJ\dots J}$ be a series of i.i.d standard normal random variables, the random process W^K on $[0,1]^d$ is given by

$$W^K(x_1,\dots,x_d)=\sum_{j_1=1}^{J}\cdots\sum_{j_d=1}^{J}Z_{j_1\dots j_d}B^K_{j_1\dots j_d}(x_1,\dots,x_d),\quad x_i\in[0,1],\, i=1,2,\dots,d. \tag{3.20}$$

where $\{B^K_{j_1\dots j_d}(x_1,\dots,x_d):\ j_i=1,\dots,J,\, i=1,2,\dots,d\}$ is a group of tensor-product B-spline basis of $\mathcal{S}_K$, $J=q+K-1$, K is the partition size of the knots. de Jonge and van Zanten (2012) has shown that $\{B^K_{j_1\dots j_d}(x_1,\dots,x_d):\ j_i=1,\dots,J,\, i=1,2,\dots,d\}$ formed an orthonormal basis of the reproducing kernel Hilbert space (RKHS) $\mathbb{H}^K$ associated with this Gaussian process W^K and also extensively exhibited the properties of the concentration function, which plays a crucial role in determining the posterior convergence rate regarding to this Gaussian process prior induced by the stochastic process W^K.

In order that the corresponding posterior could be guaranteed to take on the asymptotic properties, posterior consistency for example, the prior should have large enough support. The tuning parameter K then should be required to vary with the sample size as well as the regularity of the function of interest and the number of observations should also go to infinity. This prior, the law of the Gaussian spline prior W^K, depends explicitly on the unknown smoothness level of the object. So this is not a desired rate-adaptive procedure.

We could remedy this problem if this partition size K is viewed as the so-called hyperparameter and itself is endowed with a separate prior. In other words, we assign a probability distribution on such an unknown tuning parameter and let the partition size be carefully selected through its posterior distribution. In a Bayesian perspective, it is natural to treat this parameter as one type of hyperparameter and let it be estimated from the data via its posterior mean.

Let $\tilde{K}$ be an independent $\mathbb{N}$-valued random variable, the hierarchical Gaussian process prior is denoted by $W^{\tilde{K}}$, where $W^{\tilde{K}}|_{\tilde{K}=K}$ is described in (3.20). As prior on the standard deviation, we employ the law Π_σ of the process $\tilde{\Psi}(W^{\tilde{K}})$, that is a transformation of the stochastic process $W^{\tilde{K}}$, where the link function $\tilde{\Psi}: \mathbb{R} \to (\underline{\sigma}, \overline{\sigma})$ is given by,

$$\tilde{\Psi}(W^{\tilde{K}}) = \Psi(W^{\tilde{K}})(\overline{\sigma} - \underline{\sigma}) + \underline{\sigma}, \tag{3.21}$$

for the logistic or normal function distribution Ψ.

The following theorem follows from Theorem 4.2 in de Jonge and van Zanten (2012) that presents the general rate of contraction results for Bayesian multivariate function estimation.

THEOREM 3.2 *Assume that $w_0 = \tilde{\Psi}^{-1}(\sigma_0) \in C^\alpha([0,1]^d)$ for some integer α less than q. Let the prior Π_σ be induced by the law of the stochastic process $\tilde{\Psi}(W^{\tilde{K}})$, where the probability mass of this hyperparameter $\tilde{K}$ for each $K \geq 1$ satisfies:*

$$C_1 \exp(-D_1 K^d \log^t K) \leq P(\tilde{K} = K) \leq C_2 \exp(-D_2 K^d \log^t K), \tag{3.22}$$

for some constants $C_1, C_2, D_1, D_2, t \geq 0$. Suppose that for any $\varepsilon > 0$, $\log\left\{\left[\frac{\overline{\beta}-\underline{\beta}}{2\varepsilon}\right]+1\right\} \leq n\varepsilon^2$ and also the prior for the regression coefficient Π_β satisfies $(A1)$. Let the largest eigenvalue of $E(X_i X_i')$ denoted by $\lambda_{max}(E(X_i X_i'))$ be bounded for $i = 1, 2, \ldots, n$. Then, for a sufficiently large constant $M > 0$,

$$\Pi\{\eta : d_H(\eta, \eta_0) > M\varepsilon_n | (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\} \longrightarrow 0 \textit{ in } P_0^n\textit{-probability},$$

where,

$$\varepsilon_n = c(n/\log^{1\vee t} n)^{-\frac{\alpha}{d+2\alpha}} \vee n^{-\frac{\alpha}{d+2\alpha}} (\log n)^{\frac{(1\vee t)\alpha}{d+2\alpha} + (\frac{1-t}{2})+},$$

for a large enough positive constant c.

Note that if $\tilde{K}^d$ follows a geometric distribution with $t = 0$, then condition (3.22) is satisfied.

Here the stochastic process prior $\tilde{\Psi}(W^{\tilde{K}})$ implies a posterior rate of concentration on the space of the standard deviation functions provided the true standard deviation has regularity level α less than q. As indicated in de Jonge and van Zanten (2012), we keep the order q involved in the splines fixed so that the prior could become simpler as well as easier for simulations computationally. A common choice for q is 4 in practice.

The prior does not depend on the smoothness level α so our procedure is adaptive. If t is chosen to be equivalent to one, the the rate ε_n becomes $(n/\log n)^{-\alpha/(d+2\alpha)}$, which coincides with the optimal posterior convergence rate, up to an additional logarithm item, since the rate $n^{-\alpha/(d+2\alpha)}$ for each $\alpha > 0$ is the minimax convergence rate in the function class $C^{\alpha}([0,1]^d)$.

3.4.2 Log-spline prior

We consider a prior, in this Subsection, induced by a random series expansion in terms of tensor-product B-splines as follows:

$$W^{J,\theta}(\boldsymbol{x}) = \sum_{j_1=1}^{J} \cdots \sum_{j_d=1}^{J} \theta_{j_1 \ldots j_d} B^{K}_{j_1 \ldots j_d}(x_1, \ldots, x_d), \tag{3.23}$$

where $\theta = (\theta_{00\ldots0}, \ldots, \theta_{JJ\ldots J})$ is a J^d-dimensional vector. A prior on h could be obtained by assigning a probability distribution on the number of items J and the associated coefficient vector θ of tensor-product B-splines discussed in Shen and Ghosal (2012) as follows:

(A2) We consider a prior for J satisfying,

$$\exp(-c_1 j \log^{t_1} j) \leq \Pi(J = j) \leq \exp(-c_2 j \log^{t_2} j), \quad j = 1, 2, \ldots, \tag{3.24}$$

for some positive constants c_1, c_2 and $0 \leq t_1 \leq t_2 \leq 1$.

(A3) Given J, the prior for J^d-dimensional vector θ satisfies for each $\|\theta_0\|_\infty \leq H$ and a sufficiently small $\varepsilon > 0$,

$$\Pi(\|\theta - \theta_0\|_2 \leq \varepsilon) \geq \exp(-c_3 J^d \log(1/\varepsilon)), \tag{3.25}$$

$$\Pi(\theta \notin [-M, M]^{J^d}) \leq J^d \exp(-c_4 M^{t_3}), \tag{3.26}$$

for some positive constants c_3, c_4, t_3 and a sufficiently large $M > 0$.

Note that (A2) holds for geometric, Poisson and negative distributions when t_1, t_2 are carefully chosen. And (A3) is fulfilled if we put independent gamma and exponential distributions on each element of the vector θ. If the support of θ is a bounded and closed set, then multivariate normal and Dirichlet distributions also meet (A3). We take the law of the following stochastic process as the prior on the standard deviation σ:

$$\tilde{\Phi}(W^{J,\theta}(\boldsymbol{x})) = \frac{e^{W^{J,\theta}(\boldsymbol{x})}}{\int_0^1 e^{W^{J,\theta}(\boldsymbol{x})}\, d\boldsymbol{x}}(\overline{\sigma} - \underline{\sigma}) + \underline{\sigma}, \tag{3.27}$$

where $W^{J,\theta}(\boldsymbol{x})$ is defined in (3.23). The law of the process $\tilde{\Phi}$ gives the so-called log-spline prior for the infinite-dimensional parameter σ.

We now present the result about the posterior contraction rate based on the product prior defined by Π_β and this log-spline prior.

THEOREM 3.3 *Let $w_0 = \tilde{\Phi}^{-1}(\sigma_0) \in C^\alpha([0,1]^d)$ and the prior for the regression coefficient β, the number of items J and the associated coefficients θ satisfy (A1), (A2) and (A3) respectively. Suppose that the maximal eigenvalue of $E(X_i X_i')$ is bounded for $i = 1, 2, \ldots, n$. Assume that we endow a prior on σ by the law of the process $\tilde{\Phi}(W^{J,\theta})$, then the corresponding posterior of $\eta = (\sigma, \beta)$ contracts at the rate,*

$$\varepsilon_n = n^{-\alpha/(2\alpha+d)}(\log n)^{\alpha/(2\alpha+d)-(t_2-1)/2}, \tag{3.28}$$

in terms of the Hellinger distance d_H.

In fact, we apply Theorem 2 in Shen and Ghosal (2012) to our linear model in the presence of the heteroscedasticity with this prior Π_η to get this result. The optimal posterior convergence rate relative to the Hellinger distance is obtained by carefully selecting some sequences $\bar{J}_n,\ J_n,\ M_n,\ \bar{\varepsilon}_n$ that satisfy the conditions stated in Theorem 2 of Shen and Ghosal (2012) in order to balance bias and model complexity in our semiparametric model.

3.5 Conclusions

To summarise, we obtain an adaptive procedure in a flexible linear model with heteroscedastic normally distributed errors. More specifically, under mild restrictions on the model and priors, the posteriors of the conditional standard deviation and of the finite regression coefficients adapt to the smoothness level of the underlying standard deviation function, which is assumed to be contained in a nonparametric model. This result indicates that we could implement this Bayesian procedure as if the regularity of the underlying function were known.

The alternative asymptotic property concerning in our normal linear regression model, the Bernstein-von Mises theorem, has been developed in Norets (2015). Further research is warranted for the investigation of the existence of the Bernstein-von Mises phenomenon in this semiparametric model where the parameter of interest is the finite-dimensional regression coefficients, by directly assigning a prior on the conditional error distribution with a zero mean restriction. The estimation of the coefficients of interest in this setting that avoid the potential model misspecifications would be efficient. Particularly challenging is how to model this conditional error density with the imposition of moment restrictions. Moreover, the problem is compounded by the fact that the appropriate constructions of the priors put on these conditional error densities, making it difficult to obtain the semiparametric efficiency bound.

It would be interesting to extend the adaptive posterior concentration rate and Bernstein-von Mises theorem in our model to that in the weakly dependent data. In infinite-dimensional models, there are few results concerning these two important asymptotic properties in the weakly dependent cases. Maybe we could establish this asymptotic result under appropriate conditions on the prior, an interesting future direction.

Chapter 4

Alternatives for Ghosal Ghosh and Vaart priors

4.1 Introduction

In Bayesian nonparametrics, the important aspects of the asymptotic behavior of the posterior distribution are well established in a wide spectrum of statistical models. An early seminar work on the weak consistency properties was carried out by Schwartz (1965) under some mild conditions. Posterior consistency in stronger metrics was studied by Barron et al. (1999), Ghosal et al. (1999) and Walker (2004). The general theme about the posterior contraction rate has been developed by Ghosal et al. (2000) and Shen and Wasserman (2001) that demonstrated a finer characterization of the posterior distribution concentrating around the true parameter provided the prior is suitably chosen in the model.

In recent years, much efforts have been made on these long-standing topics of interest and there has been remarkable progress on the asymptotic analysis of nonparametric Bayesian methods. We refer to Ghosal and van der Vaart (2007a,b), Walker et al. (2007), van der Vaart and van Zanten (2008, 2009) and Rousseau (2010) for further investigations in this area as well as Ghosal (2010) for a concise review on these topics.

Most existing work primarily concerning the properties of posterior contraction rates relies heavily on the carefully chosen priors that meet some assumptions. A typical condition to impose on the priors is proposed in Ghosal et al. (2000), i.e.

$$\Pi\Big(P \in \mathscr{P} \,:\, -P_0 \log \frac{dP}{dP_0} < \varepsilon_n^2,\, P_0\Big(\log \frac{dP}{dP_0}\Big)^2 < \varepsilon_n^2\Big) \geq e^{-n\varepsilon_n^2}, \tag{4.1}$$

where (ε_n) is a positive sequence such that $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$. We call the priors satisfying this Kullback-Leibler property described in (4.1) above as *GGV priors*. Condition (4.1) requires that the prior charges some specialized Kullback-Leibler neighborhood with a large amount of prior mass. GGV priors play a crucial role in exploring the rate of posterior contraction in a broad swath of statistical models. In other words, it suggests that in order to obtain the posterior contraction rate, the prior shall put a sufficiently enough amount of probability mass around the true probability distribution or spread on the model uniformly at some discretization level.

The lower bound for this sharpened Kullback-Leibler neighborhood was designed to bound from below the denominator of the expression of the posterior distribution. Accordingly, one could tackle the posterior distribution by means of a separate examination of its numerator and denominator. Then one could formulate a rates-of-posterior-convergence theorem by additional introduction of exponentially powerful test sequences mainly driven by the metric entropy number.

Hence a natural question arises. Could we deal with the numerator and denominator of the posterior distribution simultaneously? What happens if the prior fails to meet this condition? For instance, as argued in Kleijn (2015), any prior in the support boundary estimation could not satisfy this requirement. Can we relax this standing criteria for prior choices to explore a rates-of-posterior-convergence theorem?

The goal of this present Chapter is geared to address these issues on the basis of the approach in Kleijn (2015) and provide an answer to these research questions by establishing some conditions under which a rate-of-convergence-theorem can be explored for a greater class of priors. Here we do not generalize conditions involving GGV priors or sharpen some assertion, but rather investigate a probable formulation of a new standard as a yardstick for the prior selection that encompasses a number of well-chosen applied priors as our special cases and apply to a broad class of statistical settings and then demonstrate that these flexible constraints put on the prior can be slightly complemented by the stringent restrictions on the model. Other method to deal with this similar problem in the context of a nonincreasing density on $\mathbb{R}^+$ where the Kullback-Leibler property (4.1) fails has to be pointed out. The posterior contraction rate in this case has been explored in Salomond (2013) by applying a truncated version of probability density function to a mildly modified main result in Ghosal et al. (2000). But his reasoning is based on the key premise that any monotone nonincreasing density on $\mathbb{R}^+$ admits a representation of a mixture of uniform densities. We will discuss this case in our framework later.

The remainder of this Chapter proceeds as follows. In the next Section, we first present a general result for the posterior convergence rate based on some unconventional minimax test sequences which are related to a greater extent on the prior and then provide a concise description of the existence and power of these tests. In Section 3, we show that GGV priors fall in our formulation as special cases and we re-derive the GGV theorem in Ghosal et al. (2000) under some assumptions on the model and also demonstrate the illustrations through several examples in various statistical settings. In Section 4, we consider two cases in the separable models outside the scope of the application of the main results stated in Section 2. The first case is formulated with the help of sieves involving finite covers and the second formulation includes a similar summability condition imposed on priors stated in Walker et al. (2007). The rate of convergence results for the semi-parametric estimation of boundary support points for some density on $\mathbb{R}$ are given in Section 5. We conclude with a discussion and the containing proofs and some complementary lemmas are relegated to the Appendix.

4.2 Main result

4.2.1 Notation

Some notations shall be geared up for the illustration and analysis in the subsequent Sections. Suppose there is a sequence of observations $X_1, X_2, \dots, X_n$ drawn from an unknown distribution function P_0 associated with a density function p_0, each taking values in a sample space $(\mathcal{X}, \mathscr{X})$, where $\mathscr{X}$ is a Borel σ-algebra on $\mathcal{X}$. Let Π be a prior probability measure defined on some Borel σ-algebra $\mathscr{B}$ generated by a collection of probability measures in $\mathscr{P}$ that is dominated by a σ-finite measure. The posterior distribution is the random probability measure given by,

$$\Pi\big(\, A \mid X_1, \dots, X_n\big) = \int_A \prod_{i=1}^{n} p(X_i)\, d\Pi(P) \Bigg/ \int_{\mathscr{P}} \prod_{i=1}^{n} p(X_i)\, d\Pi(P), \quad A \in \mathscr{B}. \tag{4.2}$$

The expression above makes sense only if the denominator is non-zero P_0^n-almost surely. This constraint on the denominator is equivalent to require that the following holds,

$$P_0^n \ll P_n^{\Pi}, \quad \text{for each } n \geq 1, \tag{4.3}$$

where P_n^{Π} stands for the n-fold *prior predictive distribution*,

$$P_n^{\Pi}(A) = \int_{\mathscr{P}} P^n(A)\, d\Pi(P), \quad A \in \sigma(X_1, \dots, X_n). \tag{4.4}$$

Lemma 2.1 and Corollary 2.2 in Kleijn (2015) suggests that (4.3) is satisfied if some condition imposed on the prior mass holds.

Let $\varepsilon > 0$ and d be a metric on $\mathscr{P}$, the ε-bracketing number of $\mathscr{P}$ relative to metric d, $N_{[]}(\varepsilon, \mathscr{P}, d)$ is regarded as the minimal brackets to cover $\mathscr{P}$.

In the sequel, the computation of the integral $P_0(p/q)^{\alpha}$ will be throughout this Chapter, where $0 \leq \alpha \leq 1$ and p and q are associated probability density functions of the probability measures P and Q in $\mathscr{P}$. We assume that the indicator functions $1_{\{p_0>0\}}(x)$, $1_{\{p>0\}}(x)$ and $1_{\{q>0\}}(x)$ are considered as implicit factors throughout all calculations related to the density supports if necessary.

4.2.2 Main result

The following theorem as the main result of this Chapter, a generalization of Theorem 1.2 of Kleijn (2015) for the posterior consistency property, concerns the rate of posterior contraction in the model $\mathscr{P}$.

THEOREM 4.1 *Assume that there is a prior* Π *on* $(\mathscr{P}, \mathscr{B})$ *in which* $P_0^n \ll P_n^{\Pi}$ *holds for each* $n \geq 1$. *Consider a sequence* (ε_n) *such that* $\varepsilon_n \to 0$ *and* $n\varepsilon_n^2 \to \infty$ *as* $n \to \infty$. *Let* $V_n = \{P \in \mathscr{P} : d(P, P_0) > M\varepsilon_n\}$ *for* $M > 0$. *Suppose that for each* $n \geq 1$, $\{V_{n,m}\}_{m=1}^{N_n}$ *is a finite cover of* V_n *such that,*

$$N_n \leq e^{Ln\varepsilon_n^2}, \tag{4.5}$$

for some positive constant L. *If there exist a positive constant* $\tilde{C}$ *and some model subset* B_n *such that for each* $1 \leq m \leq N_n$, $\sup_{P \in V_{n,m}} \sup_{Q \in B_n} P_0(dP/dQ) < \infty$ *and,*

$$\inf_{0 \leq \alpha \leq 1} \sup_{Q \in B_n} \sup_{P \in \text{co}(V_{n,m})} \Pi(B_n)^{-\alpha/n} P_0 \left(\frac{dP}{dQ}\right)^{\alpha} \leq e^{-\tilde{C}\varepsilon_n^2}, \tag{4.6}$$

then for a sufficiently large $M > 0$, *we arrive at,*

$$\Pi\big(\, P \in \mathscr{P} : d(P, P_0) > M\varepsilon_n \,\big|\, X_1, \ldots, X_n \,\big) \longrightarrow 0 \text{ in } P_0^n\text{-probability}. \tag{4.7}$$

A few illustrations about the assumptions made in this theorem are in order. Condition (4.5) requires that the number of model subsets that cover the complement of some metric ball should not be too large. In fact, this requirement reflects the complexity of the model in some sense. Specifically, in infinite-dimensional statistical cases, the packing number or the associated metric entropy is considered as a concise expression of the model complexity. Broadly speaking, we employ a sequence of small covering sets to partition the parameter space of interest to explore exponentially powerful test sequences for the model, see theorem 4.2 for more details.

Condition (4.6) is an appropriate uniform control of the Hellinger transform $P_0\left(\frac{dP}{dQ}\right)^{\alpha}$ over some covering set and any neighborhood B_n of P_0 as well as the prior factor $\Pi(B_n)^{-\alpha/n}$ with a sufficiently small exponential bound. Given a collection of covering sets on V_n, this condition leaves a room for the flexible choices of B_n as well as prior Π and also strikes a balance between lower bound on the prior mass on B_n and upper bound of the Hellinger transform.

We put the same model integrability condition just as stated in Kleijn (2015). This restriction on the model may be thought of as a small price for the freedom to choose priors. Therefore theorem 4.1 is exactly regarded as an extension of his results about posterior consistency to the posterior contraction rates.

4.2.3 Posterior concentration

As argued in the previous Subsection, to obtain the rate of posterior contraction given a prior, one needs to show that the corresponding posterior concentrates on small balls around P_0 with a radius of some order. Generally speaking, this order, a positive sequence decaying to zero, could be viewed as a rate but the posterior can still capture most of the probability mass asymptotically. The following theorem, on the basis of the techniques proposed in Kleijn (2015), which was developed for establishing posterior consistency for a variety of statistical cases such as a prior charging Hellinger balls as well as the scenario that violates the requirement of prior mass on the Kullback-Leibler ball, asserts that the posterior contraction rate is determined provided a sequence of tests on the covering sets of the complement exist.

THEOREM 4.2 *Let $X_1, X_2, \ldots, X_n$ be i.i.d sample distributed from $P_0 \in \mathscr{P}$ which is a metric space equipped with some metric d and is also dominated by a σ-finite measure. Let Π be a prior on $\mathscr{P}$ such that $P_0^n \ll P_n^{\Pi}$ holds for all $n \geq 1$. Consider $V_n = \{P \in \mathscr{P} : d(P, P_0) > M\varepsilon_n\}$ with $M > 0$ as well as a positive sequence (ε_n) such that $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$ as $n \to \infty$. Assume that for each $n \geq 1$, $\{V_{n,m}\}_{m=1}^{N_n}$ is a finite cover of V_n such that,*

$$N_n \leq e^{Ln\varepsilon_n^2}, \tag{4.8}$$

for some positive constant L. If for each $n \geq 1$ and $1 \leq m \leq N_n$, there exists a test $\phi_{n,m}$ such that for a universal positive constant K,

$$P_0^n \phi_{n,m} + \sup_{P \in V_{n,m}} P_0^n \frac{dP^n}{dP_n^\Pi}(1 - \phi_{n,m}) \leq e^{-KM^2 n \varepsilon_n^2}, \tag{4.9}$$

and for all $P \in V_{n,m}$,

$$P_0^n (dP^n / dP_n^\Pi) < \infty, \tag{4.10}$$

then for a sufficiently large $M > 0$, we have that,

$$\Pi\big(P \in \mathscr{P} : d(P, P_0) > M\varepsilon_n \bigm| X_1, \ldots, X_n \big) \longrightarrow 0 \textit{ in } P_0^n\textit{-probability.} \tag{4.11}$$

Condition (4.9) plays a key role in determining the rate of posterior convergence by view of this theorem. The main feature that this test distinguishes from the traditional one discussed in Ghosal et al. (2000) is that the probability operator taken on in the type II error involves the prior predictive distribution, hence the prior is naturally regarded as an important factor to build up this test.

4.2.4 Existence and power of test sequences

A main element in the proof of theorem 4.2 is to construct some kind of nonparametric tests of P_0 versus some alternative d-balls that have sufficiently small exponential bound on the probability of the type I and type II errors, here d could be some general loss-functions besides the regular Hellinger metric or total variation distance. Early theorization of the construction of exponentially powerful hypothesis tests is traced back to Le Cam (1973, 1975, 1986) and Birgé (1983, 1984) who applied the minimax theorem with the Hellinger entropy number to formulate some test for any two pair convex set $\mathscr{P}_0$ and $\mathscr{P}_1$ of probability measures. To this aim, we explore an alternative version on this theme in terms of the so-called Hellinger transform described extensively in Kleijn (2003, 2015) and Kleijn and van der Vaart (2006, 2012).

Define $V_{n,m}^n = \{P^n : P \in V_{n,m}\}$ and denote its corresponding convex hull by $\mathrm{co}(V_{n,m}^n)$ and the generic element in $\mathrm{co}(V_{n,m}^n)$ by P_n.

LEMMA 4.3 *Consider $n \geq 1$ and a sequence of model subsets $\{V_{n,m}\}_{m=1}^{N_n}$. Assume that $P_0^n(dP^n/dP_n^\Pi) < \infty$ for all $P \in V_{n,m}$, $1 \leq m \leq N_n$. Then there exists a test sequence $(\phi_{n,m})$ such that,*

$$P_0^n \phi_{n,m} + \sup_{P \in V_{n,m}} P_0^n \frac{dP^n}{dP_n^\Pi}(1 - \phi_{n,m}) \leq \sup_{P_n \in \mathrm{co}(V_{n,m}^n)} \inf_{0 \leq \alpha \leq 1} P_0^n \Big(\frac{dP_n}{dP_n^\Pi}\Big)^\alpha, \tag{4.12}$$

i.e. here the testing power embedded in some subset is appropriately bounded by means of the Hellinger transform.

PROOF OF LEMMA 4.3

This lemma can be completed by the method analogous to that used in Lemma 2.2 in Kleijn (2015) just replacing ϕ_n, V there with $\phi_{n,m}, V_{n,m}$ respectively. □

More specifically, the prior could be localized on some measurable set B_n with positive prior mass that mainly centers on P_0 in the same technical fashion as in Wong and Shen (1995); where the construction of the sieve could lead naturally to some good approximation to P_0. To this aim, we introduce the local prior predicative distribution proposed in Kleijn (2015). Given a prior Π and a measurable set B_n such that $\Pi(B_n) > 0$, the *local* prior predictive distributions Q_n^Π is defined as follows:

$$Q_n^\Pi(A) = \int Q^n(A)\, d\Pi(Q|B_n), \tag{4.13}$$

for each $n \geq 1$ and $A \in \sigma(X_1, \ldots, X_n)$. The following lemma makes use of the local prior predicative distribution to further establish some more specific conditions on model and prior.

LEMMA 4.4 *Fix $n \geq 1$, consider B_n and a collection of model subsets $\{V_{n,m}\}_{m=1}^{N_n}$ in $\mathscr{B}$. Assume that $\Pi(B_n) > 0$ and for all $P \in V_{n,m}$, $1 \leq m \leq N_n$,*

$$\sup_{Q \in B_n} P_0(dP/dQ) < \infty, \tag{4.14}$$

then there exists a test sequence $(\phi_{n,m})$ *such that,*

$$P_0^n \phi_{n,m} + \sup_{P \in V_{n,m}} P_0^n \frac{dP^n}{dP_n^\Pi}(1-\phi_{n,m}) \leq \inf_{0\leq\alpha\leq 1} \Pi(B_n)^{-\alpha} \int \Big[\sup_{P\in \mathrm{co}(V_{n,m})} P_0 \Big(\frac{dP}{dQ}\Big)^\alpha \Big]^n d\Pi(Q|B_n). \quad (4.15)$$

As a consequence, our main result, theorem 4.1, follows by invoking theorem 4.2 together with lemmas 4.3 and 4.4.

4.3 Sufficiency of GGV priors

In this Section the room of sufficiency of GGV priors will be exploited relative to condition (4.6) in theorem 4.1. We apply theorem 4.2 to encompass GGV priors under more stringent model conditions.

LEMMA 4.5 *Consider two model subsets* B, V *such that* $P_0 \in B$*. Suppose that* $\sup\limits_{Q\in B} \sup\limits_{P\in V} P_0(dP/dQ)$ *is finite. Given* $\varepsilon > 0$*, then,*

$$\sup_{Q\in B} -P_0 \log \frac{dQ}{dP_0} - \inf_{P\in V} \Big\{ -P_0 \log \frac{dP}{dP_0} \Big\} < -\tilde{M}\varepsilon^2, \quad (4.16)$$

if and only if,

$$\inf_{0\leq\alpha\leq 1} \sup_{Q\in B} \sup_{P\in V} P_0 \Big(\frac{dP}{dQ}\Big)^\alpha < e^{-\tilde{M}\varepsilon^2}, \quad (4.17)$$

where $\tilde{M}$ *is a positive constant. That is,* B *and* V *are strictly separated in Kullback-Leibler divergence with a slightly small difference iff we could uniformly control the Hellinger transform of two probability measures.*

It is clear that the above lemma does not require any restrictions related to the second moment of log-likelihood ratio $P_0\Big(\log \dfrac{dP}{dQ}\Big)^2$. Hence condition (4.16) offers us more room to explore the rate of posterior contraction without any prior mass put on the ball involving the higher moments of the Kullback-Leibler discrepancy. So we expect that condition (4.1) about GGV priors is sufficient but maybe not necessary to determine the rate of posterior convergence.

The following theorem precisely provides such an assertion in the preceding display without requiring more of the prior on neighborhood concerning the second moment of the Kullback-Leibler divergence.

THEOREM 4.6 *Fix $n \geq 1$, there exists a Kullback-Leibler neighbourhood B_n of P_0 in which $\sup_{Q \in B_n} P_0(dP/dQ)$ is finite for all $P \in \mathscr{P}$. Let Π satisfy (4.1) for a positive sequence (ε_n) such that $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to 0$ as $n \to 0$ and also d be either the Hellinger distance or total variation metric on $\mathscr{P}$. Assume that for $M > 0$, $V_n = \{P \in \mathscr{P} : d(P, P_0) > M\varepsilon_n\}$ is covered by a number of Hellinger balls $V_{n,1}, \ldots, V_{n,N_n}$ with the same radii ε_n, where $N_n \leq e^{Ln\varepsilon_n^2}$ for some positive constant L. Then we have that, for a sufficiently large $M > 0$,*

$$\Pi\big(\, P \in \mathscr{P} : d(P, P_0) > M\varepsilon_n \,\big|\, X_1, \ldots, X_n \,\big) \longrightarrow 0 \text{ in } P_0^n\text{-probability}. \tag{4.18}$$

So we have adapted GGV priors in our framework to obtain the rate of posterior contraction under some additional model assumptions. This flexible formulation enables us to accommodate a broad range of prior choices.

4.3.1 Hellinger prior

Generally speaking, under quite general assumptions on the model, theorem 4.6 examines the posterior contraction rate if there exists one Kullback-Leibler neighborhood with a large amount of prior mass. Since each Kullback-Leibler neighborhood is contained in a Hellinger ball, in this Subsection we try to show the assertion in theorem 4.6 also holds under a weak assumption about the prior charging Hellinger ball and more specific model conditions.

THEOREM 4.7 *Suppose that there exist a positive sequence $\varepsilon_n \downarrow 0$ and a positive constant $\tilde{C}$ such that for each $n \geq 1$,*

$$N(\varepsilon_n, \mathscr{P}, d_H) \leq e^{\tilde{C}n\varepsilon_n^2}. \tag{4.19}$$

Given a prior Π and for each $n \geq 1$, assume that there exist a Hellinger ball B'_n with a ε'_n radius and a *positive constants L' and C' such that for all $Q \in B_n'$, $P \in \mathscr{P}$,*

$$\left\| \frac{dP}{dQ} \right\|_{2,Q} \leq L', \tag{4.20}$$

$$\Pi(B_n') \geq e^{-C'n\varepsilon_n'^2}. \tag{4.21}$$

If $X_1, X_2, \dots, X_n$ constitute an i.i.d P_0-sample, where $P_0 \in \mathscr{P}$, then we get

$$\Pi\big(P \in \mathscr{P} : d_H(P, P_0) > M\varepsilon_n \bigm| X_1, \dots, X_n\big) \longrightarrow 0 \text{ in } P_0^n\text{-probability}, \tag{4.22}$$

for a sufficiently large $M > 0$.

The finiteness of the covering number relative to the Hellinger distance d_H in theorem 4.7 explicitly requires that $(\mathscr{P}, d_H)$ is a totally bounded space. In this case we could construct a so-called net prior under bracketing entropy conditions. Let (ε_n) be a positive monotone decreasing sequence with its limit be zero. Fixed $n \geq 1$, denote $N_{[]}(\varepsilon_n, \mathscr{P}, d_H)$ by N_n. Consider a totally bounded space $(\mathscr{P}, d_H)$ dominated by a σ-finite measure μ, then there exist a finite collection of functions $\{l_1, u_1, l_2, u_2 \dots, l_{N_n}, u_{N_n}\}$ such that for every $P \in \mathscr{P}$, $l_j < p < u_j$ and $d_H(l_j, u_j) < \varepsilon_n$ for some $j \in \{1, 2, \dots, N_n\}$. Now a prior Π_n is characterized as a uniform measure put on a kind of normalization sets $\mathscr{P}_n := \{u_1/\int u_1\, d\mu, \dots, u_{N_n}/\int u_{N_n}\, d\mu\}$. Therefore, the infinite mixture $\Pi = \sum_{n=1}^{\infty} \lambda_n \Pi_n$ is regarded as the net prior on $\cup_{i=1}^{\infty} \mathscr{P}_i$ where (λ_m) is a positive sequence such that $\sum_{m=1}^{\infty} \lambda_m = 1$. Here the support of this discrete prior includes all the upper brackets.

Given this prior Π constructed as above, the following variation of theorem 4.7 shows that the corresponding posterior contracts at a rate around the true distribution with respective to the Hellinger distance under some mild model conditions.

THEOREM 4.8 *Let net prior Π be described as above and consider a positive monotone decreasing sequence (ε_n) and a weighted sequence (λ_n) described above, such that for each $n \geq 1$,*

$$N_{[]}(\varepsilon_n, \mathscr{P}, d_H) \leq n\varepsilon_n^2, \tag{4.23}$$

and as $n \to \infty$,

$$\varepsilon_n \to 0,$$

$$n\varepsilon_n^2 / \log n \to \infty, \tag{4.24}$$

$$\log \lambda_n^{-1} = O(\log n). \tag{4.25}$$

In addition, if the model condition (4.20) in theorem 4.7 holds, then the claim (4.22) follows.

In the following three examples, we show that the corresponding posterior convergence rate under some prior agrees with the optimal rate of frequentist estimators. In particular, we look at Bayesian estimation of probability densities satisfying some regularity conditions, monotone nonincreasing densities and distribution functions in interval censoring case 2.

EXAMPLE 4.9 Given some $\gamma > 0$, let $\mathscr{P}$ be a class of probability measures on $[0,1]$ with associated probability densities p satisfying for any $x \in [0,1]$,

$$C_L\, g(x) \leq p(x) \leq C_U\, g(x) \text{ and } \|\sqrt{p}\|_\gamma \leq 1, \tag{4.26}$$

where C_L, C_U are positive constants, $g(x)$ is a known density function with support on $[0,1]$ and $\|\cdot\|_\gamma$ denotes the Hölder norm defined in Chapter 2.7 of van der Vaart and Wellner (1996).

Taking $\varepsilon_n = n^{-\gamma/(2\gamma+1)}$, by Corollary 2.7 in van der Vaart and Wellner (1996), the ε_n-bracket entropy numbers relative to the norm $\|\cdot\|_\gamma$ for every $1 \leq \gamma \leq \infty$ are bounded above by $n\varepsilon_n^2$. So we could construct a net prior Π in this case. Evidently, we see that the conditions in theorem 4.8 hold, giving rise to a $n^{-\gamma/(2\gamma+1)}$-rate of posterior contraction. What's more, it coincides with the optimal rate of the similar estimators in the frequentist context. Thus by theorem 4.8, the corresponding posterior from this net prior built on the finite set of brackets converges to the true distribution at the optimal rate.

EXAMPLE 4.10 Suppose we consider the problem of estimating a density function that is nonincreasing on $(0,\infty)$. This theme was introduced in Khazaei and Balabdaoui-Mohr (2010) and Salomond

(2013), which established posterior consistency and determined the rate of convergence under the Dirichlet prior and the finite mixture prior respectively.

Given a density function g supported on $(0, \infty)$ and a sufficiently small positive constant C'_L, let us consider the model $\mathscr{P}$ defined by,

$$\mathscr{P} := \left\{ p : \, p \downarrow, \, \int_0^1 p \, d\mu = 1, \, C'_L g(x) \leq p \leq C'_U \, g(x) \text{ for } x \in (0, \infty) \right\},$$

where C'_U is a positive constant. That is to say, the model $\mathscr{P}$ consists of all nonincreasing density functions supported on $(0, \infty)$, which are dominated by a known probability density function. According to Theorem 2.7.5 in van der Vaart and Wellner (1996), for any $\varepsilon > 0$, one could show,

$$\log N_{[]}(\varepsilon, \tilde{\mathscr{P}}, L^r(\mu)) \leq \tilde{C}'/\varepsilon, \tag{4.27}$$

for each $1 \leq r \leq \infty$, where $\tilde{\mathscr{P}} := \{\sqrt{p} : \, p \in \mathscr{P}\}$ and $\tilde{C}'$ is a positive constant that depends only on r.

Put $\epsilon = \varepsilon_n = n^{-1/3}$, then the monotonicity of $\sqrt{p}$ implies that ε-bracket entropy number $\log N_{[]}(\varepsilon_n, \mathscr{P}, d_H) \leq \tilde{C}' n \varepsilon_n^2$. Then a net prior could be designed in this statistical model. By virtue of the assumption that any nonincreasing density in $\mathscr{P}$ is dominated by a known density function, the model condition (4.20) in theorem 4.7 is fulfilled. Notice that $\varepsilon_n = n^{-1/3}$ is the optimal rate of the frequentist estimator, therefore in view of theorem 4.8 this net prior constructed from the finite approximation sets by means of brackets attains the optimal posterior contraction rate.

EXAMPLE 4.11 This example concerns the nonparametric Bayesian analysis in the field of survival analysis, with a particular focus on the estimation of the life distribution functions of the censored data from a Bayesian viewpoint, which often arises in a variety of contexts, such as medical research, actuarial sciences and reliability theory.

Suppose a representative sample about the ages of people who are possible to get lung cancer is observed during some period (say, 3 years). Two lung cancer tests are administered to each person in this sample within this specified period. Let X be the age at lung cancer with associated distribution being F, and a pair of observation times $(T_1,\, T_2)$ are the ages at which the person is implemented by these two lung cancer tests with a joint distribution G and its associated density g, and also X is independent with $(T_1,\, T_2)$. In this context, we observe a n i.i.d. sample $\{(X_i, T_{1i}, T_{2i}, \Delta_{1i}, \Delta_{2i}, \Delta_{3i})\}_{i=1}^n$, where $\Delta_{1i} = 1\{X_i \leq T_{1i}\}$, $\Delta_{2i} = 1\{T_{1i} < X_i \leq T_{2i}\}$ and $\Delta_{3i} = 1\{X_i > T_{2i}\}$, $i = 1, 2, \dots, n$. For example, if one realization for the observation of person j is $(x_j, t_{1j}, t_{2j}, \delta_{1j}, \delta_{2j}, \delta_{3j})$, where $\delta_{1j} = \delta_{2j} = 0$ and $\delta_{3j} = 1$, then we know this person j has not got lung cancer at up to age t_{2j} and $x_j > t_{2j}$. The probability density can be computed by considering three cases $\Delta_1 = 1$, $\Delta_2 = 1$ and $\Delta_3 = 1$ separately. Then the density at a realization $(t_1, t_2, \delta_1, \delta_2, \delta_3)$ is given by,

$$p_F(t_1, t_2, \delta_1, \delta_2, \delta_3) = F(t_1)^{\delta_1}(F(t_2) - F(t_1))^{\delta_2}(1 - F(t_2))^{\delta_3} g(t_1, t_2). \tag{4.28}$$

Moreover, some routine calculations show the likelihood at n i.i.d. realizations $\{(x_i, t_{1i}, t_{2i}, \delta_{1i}, \delta_{2i}, \delta_{3i})\}_{i=1}^n$ is:

$$\Pi_{i=1}^n (F(t_{1i}))^{\delta_{1i}}(F(t_{2i}) - F(t_{1i}))^{\delta_{2i}}(1 - F(t_{2i}))^{\delta_{3i}} g(t_{1i}, t_{2i}). \tag{4.29}$$

Note that the likelihood is factorized into the conditional likelihood of $(\Delta_1, \Delta_2, \Delta_3)$ given a pair of observation time (T_1, T_2) and the marginal likelihood g of this joint time (T_1, T_2), then the marginal density g vanishes in the expression of the posterior distribution of the life distribution F. Hence it is not necessary to specify a prior on the joint distribution G and in this case G could be regarded as an known distribution distribution.

Assume that F supports on a compact interval, say, $[0, 1]$ and $F = \Psi(W)$, where this link function $\Psi : \mathbb{R} \to (0, 1)$ is the normal or logistic distribution function and W is a Gaussian process chosen later. Given $\tilde{\alpha} > 0$, assume also that there is a true distribution F_0 for which $F_0 = \Psi(w_0)$, where $w_0 \in C^{\tilde{\alpha}}[0, 1]$. Let the support of G be a closed interval strictly contained in $(0, 1)$. Denote,

$$\mathscr{F} := \{F : \; F \in C^{\tilde{\alpha}}[0, 1], \; \|F\|_{\tilde{\alpha}} \leq 1 \text{ and } F \text{ is a distribution function on [0,1]}\}.$$

As prior on F, we consider the law Π of the transformed process $W^{\tilde{\alpha}}$ via $F = \Psi(W^{\tilde{\alpha}})$, where $W^{\tilde{\alpha}}$ is the modified Riemann-Liouville process given by,

$$W_t^{\tilde{\alpha}} = \sum_{j=0}^{\lfloor\tilde{\alpha}\rfloor+1} Z_j t^j + R_t^{\tilde{\alpha}}, \quad t \in [0,1], \tag{4.30}$$

where $R_t^{\tilde{\alpha}}$ is the Riemann-Liouville process with Hurst parameter $\tilde{\alpha}$ defined by $R_t^{\tilde{\alpha}} = \int_0^t (t - s)^{\lfloor\tilde{\alpha}\rfloor-1/2}\, dW_s,\ t \geq 0$, $\lfloor\tilde{\alpha}\rfloor$ is the largest integer that is strictly smaller than $\tilde{\alpha}$ and $Z_1, Z_2, \ldots, Z_{\lfloor\tilde{\alpha}\rfloor+1}$ are i.i.d standard normal random variables which are independent of $R_t^{\tilde{\alpha}}$. As shown in Theorem 4.3 in van der Vaart and van Zanten (2008), the support of this process $W^{\tilde{\alpha}}$ is $C[0,1]$ and the concentration function is smaller than $\varepsilon^{-1/\tilde{\alpha}}$ for each $\varepsilon > 0$ small enough. Hence according to Theorem 2.1 in van der Vaart and van Zanten (2008), it turns out that,

$$Pr\left(\|W^{\tilde{\alpha}} - w_0\|_\infty < 2\varepsilon_n\right) \geq e^{-n\varepsilon_n^2},$$

where $\varepsilon_n = n^{-\tilde{\alpha}/(2\tilde{\alpha}+1)}$. Using the elementary inequality $(\sqrt{a} - \sqrt{b})^2 \leq (a-b)^2$ for $a \geq 0,\ b \geq 0$, it follows that,

$$\begin{aligned} d_H^2(p_F, p_{F_0}) &= \int\int \left|F^{1/2}(t_1) - F_0^{1/2}(t_1)\right|^2 dG(t_1,t_2) + \int\int \left|(1-F(t_1))^{1/2} - (1-F_0(t_1))^{1/2}\right|^2 dG(t_1,t_2) \\ &\quad + \int\int \left|(F(t_1) - F(t_2))^{1/2} - (F_0(t_1) - F_0(t_2))^{1/2}\right|^2 dG(t_1,t_2) \\ &\leq 6 \sup_{t\in[0,1]} |F(t) - F_0(t)|^2 = 6\|F - F_0\|_\infty^2 \\ &\leq 6\|W^{\tilde{\alpha}} - w_0\|_\infty^2. \end{aligned}$$

where the last step follows the fact that the logistic function is differentiable with uniformly bounded derivative. Taking the square root of both sides of the inequality above, we get,

$$d_H(p_F, p_{F_0}) \leq \sqrt{6}\|F - F_0\|_\infty \leq \sqrt{6}\|W^{\tilde{\alpha}} - w_0\|_\infty. \tag{4.31}$$

So the Hellinger distance between P_F and P_{F_0} could be bounded above by the uniform norm of $W^{\tilde{\alpha}} - w_0$. Hence for $\varepsilon_n = n^{-\tilde{\alpha}/(2\tilde{\alpha}+1)}$ one could get,

$$N(\varepsilon_n, \{p_F : F \in \mathscr{F}\}, d_H) \leq N(\varepsilon_n/6, \mathscr{F}, \|\cdot\|_\infty) \leq e^{C'' n\varepsilon_n^2/36},$$
$$\Pi\{p_F : d_H(p_F, p_{F_0}) \leq \varepsilon_n\} \geq \Pi\{F : \|W^\alpha - w_0\|_\infty \leq \varepsilon_n/\sqrt{6}\} \geq e^{-n\varepsilon_n^2/24},$$

where C'' is positive constant. In addition, the model condition (4.20) in theorem 4.7 is automatically satisfied since any distribution function in $\mathscr{F}$ is bounded between 0 and 1. Therefore, by theorem 4.8, the prior based on this transformed process $W^{\tilde{\alpha}}$ in the interval censoring case 2 yields the optimal rate $n^{-\tilde{\alpha}/(2\tilde{\alpha}+1)}$.

4.4 Posterior contraction on separable models

The assumption that the existence of the finite number of model subsets to cover some measurable set in theorems 4.1 and 4.2 is admittedly restrictive. Unbounded parameter spaces in the Euclidean space, for example, do not coincide with this requirement. In this Section, we consider two alternatives to circumvent those problems due entirely to the finiteness of the order of the cover. The introduction of sieves that to a significant extent entails the model complexity to approximate the model can be found in Subsection 4.4.1. In Subsection 4.4.2, the number of subcovers will be allowed to be infinite accountable and the posterior convergence rate with respect to Hellinger distance is achieved for a wide range of priors that meet a summability condition developed in Walker et al. (2007). Meanwhile, we present a versatile version of Theorem 1 in Walker et al. (2007).

4.4.1 Finite covers

In a polish space, a sieve could be naturally regarded as a useful device to support the prior constructed in this space in a large part. A rate can be achieved by mild modifications on the assumptions in theorem 4.1 with the aid of sieves.

THEOREM 4.12 *Suppose that there exist a positive sequence* (ε_n) *with* $\varepsilon_n \to 0$ *and* $n\varepsilon_n^2 \to \infty$ *as* $n \to \infty$, *positive constants* K, L *and sets* $\mathscr{P}_n \subset \mathscr{P}$ *such that for sufficiently large* $n \geq 1$ *and* $M > 0$,

(i.) $V_n \cap \mathscr{P}_n = \{P \in \mathscr{P} : d(P, P_0) > M\varepsilon_n\} \cap \mathscr{P}_n$ *is covered by a finite number of measurable subsets* $V_{n,1}, V_{n,2}, \ldots, V_{n,N_n}$, *where* $N_n \leq e^{\frac{1}{2}Ln\varepsilon_n^2}$;

(ii.) for each $1 \leq m \leq N_n$, *there is a test* $\phi_{n,m}$ *such that,*

$$P_0^n \phi_{n,m} + \sup_{P \in V_{n,m}} P_0^n \frac{dP^n}{dP_n^\Pi}(1 - \phi_{n,m}) \leq e^{-Ln\varepsilon_n^2}; \tag{4.32}$$

(iii.) the prior mass on the complement of the set $\mathscr{P}_n$ *such that,*

$$\Pi(\mathscr{P} \backslash \mathscr{P}_n) \leq e^{-Kn\varepsilon_n^2}; \tag{4.33}$$

(iv.) there is a model subset B_n *with* $\Pi(B_n) \geq e^{-\frac{K}{2}n\varepsilon_n^2}$ *such that,*

$$\sup_{P \in V_n \cap \mathscr{P}_n^c} \sup_{Q \in B_n} P_0\Big(\frac{dP}{dQ}\Big) < e^{\frac{K}{4}\varepsilon_n^2}. \tag{4.34}$$

If $X_1, X_2, \ldots, X_n$ *constitute an i.i.d* P_0*-sample, where* $P_0 \in \mathscr{P}$, *then we get,*

$$\Pi\big(P \in \mathscr{P} : d(P, P_0) > M\varepsilon_n \bigm| X_1, \ldots, X_n\big) \longrightarrow 0 \text{ in } P_0^n\text{-probability}, \tag{4.35}$$

for a sufficiently large $M > 0$.

Conditions *(i.)* and *(ii.)* are analogous to conditions (4.8) and (4.9) of theorem 4.1 except that the covering sets are deeply embedded in the sieves $\mathscr{P}_n$. So that lemmas 4.3 and 4.4 that illustrate the existence and power of test sequences also hold within a sequence of sieves. Condition *(iii.)* that states the negligible prior mass outside the sieve is commonly used in the literature. Besides, condition *(iv.)* requires a large amount of prior mass on some model set and the Hellinger transform of the true distribution P_0 and another probability measures defined on the complement of sieves and this model subset uniformly vanish at an exponential rate. The latter could be typically viewed as a model condition.

Because of lemmas 4.3 and 4.4, we could formulate one alternative instead of condition *(ii.)* in the preceding theorem. More precisely, the following corollary illustrates this point.

COROLLARY 4.13 *Let the conditions (i.), (iii.) and (iv.) listed in theorem 4.12 hold as well. Assume that there exists some positive constant K for every $1 \leq m \leq N_n$, we have,*

$$\inf_{0 \leq \alpha \leq 1} \sup_{P \in \mathrm{co}(V_{n,m})} \sup_{Q \in B_n} P_0 \Big(\frac{dP}{dQ}\Big)^{\alpha} < e^{-K\varepsilon_n^2}, \tag{4.36}$$

and for all $P \in V_{n,m}$, $1 \leq m \leq N_n$,

$$\sup_{Q \in B_n} P_0 \Big(\frac{dP}{dQ}\Big) < \infty. \tag{4.37}$$

Then the claim (4.35) follows.

4.4.2 Infinite countable covers

We employed a sequence of sieves to alleviate the situations lack of total bounded requirement on the models by constructing a finite number of covering sets on a series of approximating models in the previous Subsection. Another way to address this question is to construct an infinite accountable number of covers. Following the approach in Walker et al. (2007), we center on the separable models with respect to some metric in which there exists countable model subsets to cover the model $\mathscr{P}$. More particularly, as was the case in Walker et al. (2007), we treat the rates theorem in Hellinger distance and a brief account for the definition of the Hellinger separable model can be found in Kleijn (2015).

In order to utilize the summability condition explored in Walker et al. (2007) that essentially imposes an upper bound for the prior mass, we formulate the equivalent version of theorem 4.4 in Kleijn (2015) that applies model subsets that display n-dependence.

THEOREM 4.14 *For a given prior Π, let $P_0^n \ll P_n^{\Pi}$ for each $n \geq 1$. Assume that $V_n = \{P \in \mathscr{P} :$*

$d(P, P_0) > M\varepsilon_n\}$ is covered by an infinite countable number of model subsets $V_{n,1}, V_{n,2}, \ldots$, where $M > 0$ and (ε_n) is a positive sequence such that $\varepsilon_n \to 0$, $n\varepsilon_n^2 \to \infty$ and also there exist a sequence of measurable subsets $(B_{n,i})_{i\geq 1}$ and a positive constant K' such that,

$$\Pi(B_{n,i}) \geq e^{-K'n\varepsilon_n^2}, \tag{4.38}$$

and,

$$\sup_{Q\in B_{n,i}} P_0\Big(\frac{dP}{dQ}\Big) < \infty, \tag{4.39}$$

for all $P \in V_{n,i}, i \geq 1$. Then we have,

$$P_0^n \Pi(V_n|X_1, \ldots, X_n) \leq \sum_{i=1}^{\infty} \inf_{0\leq\alpha\leq 1} \frac{\Pi(V_{n,i})^\alpha}{\Pi(B_{n,i})^\alpha} \left[\sup_{P\in \mathrm{co}(V_{n,i})} \sup_{Q\in B_{n,i}} P_0\Big(\frac{dP}{dQ}\Big)^\alpha\right]^n. \tag{4.40}$$

We will apply theorem 4.14 to present a generalization of theorem 1 in Walker et al. (2007) in the following two corollaries. The first corollary requires the prior to meet the summability requirement described in Walker et al. (2007) and also to admit a sufficient amount of probability mass on some model set that is beyond the scope of the Kullback-Leibler neighbourhood. Furthermore, we impose some model condition like that of (4.6).

COROLLARY 4.15 *Let $B_{n,i} = B_n$ for all $i \geq 1, n \geq 1$ and also the conditions stated in theorem 4.14 hold. Additionally, suppose,*

$$\sup_{P\in \mathrm{co}(V_{n,i})} \sup_{Q\in B_n} P_0\Big(\frac{dP}{dQ}\Big)^{1/2} < e^{-K''\varepsilon_n^2}, \tag{4.41}$$

for some positive constant $K'' > K'$. If there exist some positive constant $\tilde{K}$ and a positive sequence (ε_n) such that $\varepsilon_n \to 0$, $n\varepsilon_n^2 \to \infty$ and the prior satisfies,

$$e^{-\tilde{K}n\varepsilon_n^2} \sum_{i=1}^{\infty} \Pi(V_{n,i})^{1/2} \to 0 \quad \text{as} \quad n \to \infty. \tag{4.42}$$

Then for i.i.d P_0-distributed $X_1, X_2, \ldots,$ and a sufficiently large $M > 0$, we have,

$$\Pi\big(P \in \mathscr{P} : d(P, P_0) > M\varepsilon_n \bigm| X_1, \ldots, X_n\big) \longrightarrow 0 \textit{ in } P_0^n\textit{-probability}. \tag{4.43}$$

The second one allows the prior to satisfy a slightly stronger summability condition stated in Walker et al. (2007) with a large amount of probability mass on the Kullback-Leibler neighbourhood instead.

COROLLARY 4.16 *Let a prior Π and $\mathscr{P}$ be given and assume that $\mathscr{P}$ is separable with respect to the Hellinger metric d_H. For each $n \geq 1$, there exists a Kullback-Leibler neighborhood B_n of P_0 such that $\sup_{Q \in B_n} P_0\left(\frac{dP}{dQ}\right) \leq \infty$ for each $P \in \mathscr{P}$ and also $\Pi(B_n) \geq e^{-\tilde{K}_1 n\varepsilon_n^2}$ for some positive constant $\tilde{K}_1$ and a positive sequence (ε_n) with $\varepsilon_n \to 0$, $n\varepsilon_n^2 \to \infty$. Assume also there exists a positive constant $\tilde{K}_2$ such that for each $\beta \in [0, 1]$,*

$$e^{-\tilde{K}_2 n\varepsilon_n^2} \sum_{i=1}^{\infty} \Pi(V_{n,i})^{\beta} \to 0 \quad as \quad n \to \infty. \tag{4.44}$$

Then the claim (4.43) follows with the Hellinger metric d_H instead of d there.

4.5 Marginal posterior contraction

Semiparametric is by now viewed as a special and vibrant research area in statistics due to its novelty as well as the genuine scientific utility and intriguing theoretical complexity of models arisen in its field. There has been well established for the asymptotic theory of the frequentist semiparametric approaches, such as notions of optimality, asymptotically pivotal statistics, to name a few; See Bickel et al. (1993) for more details. However, from a Bayesian perspective, unfortunately few attempts have been developed for the asymptotic theory of Bayesian semiparametric methods beyond Bickel and Kleijn (2012), Kleijn and Knapik (2012) and Kleijn (2015), and more specifically, a Bernstein-von Mises theorem for LAN and LAE model under reasonably mild conditions is elegantly investigated if the posterior distribution of the parameter of interest is consistent. A further result along the line of this direction could be the establishment of posterior contraction rate involving this finite-dimensional parameter of interest. In this Section, we will demonstrate the posterior contraction rate under some mild model conditions and reasonable assumptions put on the prior.

To establish the basic framework, consider the model $\mathscr{P} := \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ where Θ is an open subset of $\mathbb{R}^k$ equipped with some metric g and H is an infinite-dimensional nuisance parameter space. We assume that $\mathscr{P}$ is dominated by some σ-finite measure defined on the sample space $(\mathcal{X}, \mathscr{X})$ with associated probability densities $p_{\theta,\eta}$. The prior on $\mathscr{P}$ is induced by a probability measure Π on $\Theta \times H$. The posterior distribution is said to contract at a rate ε_n if,

$$\Pi\big(\{P_{\theta,\eta} \in \mathscr{P} :\, g(\theta,\theta_0) > M_n\varepsilon_n,\, \eta \in H\} \;\big|\; X_1,\ldots,X_n\big) \longrightarrow 0 \text{ in } P_0^n\text{-probability}, \tag{4.45}$$

for every $(\theta_0, \eta_0) \in \Theta \times H$ and all $M_n \to \infty$, as $n \to \infty$.

Additionally, the identifiability of the parameter θ is defined by means of an identity mapping $\theta : \mathscr{P} \to \Theta$ with $\theta(P_{\theta,\eta}) = \theta$ for all $(\theta,\eta) \in \Theta \times H$. Hence the metric g in Θ could be characterized by the following pseudo-metric $d : \mathscr{P} \times \mathscr{P} \to [0,\infty)$,

$$d\big(P_{\theta,\eta}, P_{\theta',\eta'}\big) = g(\theta,\theta'), \tag{4.46}$$

for all $\theta, \theta' \in \Theta$ and $\eta, \eta' \in H$.

4.5.1 Density support boundary estimation

The properties of the boundary domain problem, such as consistency of parameter of interest, have been explored from the frequentist perspective, see for instance the introduction of this subject in Ibragimov and Hasminskii (1981), and more recently in a Bayesian viewpoint, such as posterior consistency discussed in Kleijn (2015) and the Bernstein-von Mises property exhibited in Kleijn and Knapik (2012). However, less is known about the rate of posterior contraction to our best of knowledge. In this Subsection, a preliminary attempt is conducted to establish this result under certain prior conditions for the nuisance space and the parameter space of interest, as well as some additional model assumptions.

Before moving on to major results, a basic specification of this model is introduced below. Our model $\mathscr{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ is described by means of Lebesgue densities given as follows:

$$p_{\theta,\eta}(x) = \frac{1}{\theta_2 - \theta_1}\, \eta\Big(\frac{x - \theta_1}{\theta_2 - \theta_1}\Big)\, 1_{\{\theta_1 \le x \le \theta_2\}},$$

where $(\theta_1, \theta_2) \in \Theta$ and $\eta \in H$ which is defined as a collection of Lebesgue probability densities with their support in accordance with $[0, 1]$. Moreover, there exists a continuous, monotone nondecreasing $f : (0, \infty) \to (0, \infty)$ such that,

$$\inf_{\eta \in H} \min\Big\{\int_0^{\epsilon} \eta \, d\mu, \int_{1-\epsilon}^{1} \eta \, d\mu\Big\} \ge f(\epsilon), \quad (0 < \epsilon < 1). \tag{4.47}$$

Condition (4.47) is of paramount importance to determine the posterior contraction rate in this scenario. This requirement on the nuisance space H indicates that a positive probability mass shall be put on any neighborhood of parameters θ_1, θ_2, especially around its boundary. In this context, the Bayesian estimation of the parameters θ_1, θ_2 is of primary interest and we need to specify some priors on Θ and H to obtain the relevant asymptotic results for θ_1 and θ_2. The following theorem states the rate of posterior convergence of the parameters θ_1, θ_2 with respect to the Euclidean norm $\|\cdot\|_2$ in $\mathbb{R}^2$.

THEOREM 4.17 *Given $\beta > 0$, $\sigma > 0$, let $\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 : 0 < \theta_2 - \theta_1 < \sigma\}$ and $f(x) = x^{\beta}$. Suppose there exist some positive constants $\tilde{L}$ and T, such that, for any $n \ge 1$,*

$$\Pi(B_n) \equiv \Pi\Big\{Q \in \mathscr{P} : \Big\|\frac{dP_0}{dQ} - 1\Big\|_{s,Q} < \delta_n^2\Big\} \ge e^{-\tilde{L}n\delta_n^2}, \tag{4.48}$$

$$\sup_{P \in \mathscr{P}} \sup_{Q \in B_n} \Big\|\frac{dP}{dQ}\Big\|_{r,Q} \le T, \tag{4.49}$$

where $\delta_n = \big[\frac{1}{2Tn\sigma^{\beta}}\big]^{1/2}$, $1/r + 1/s = 1$ and $\|\cdot\|_{s,Q}$ stands for the $L_s(Q)$-norm. If $X_1, X_2, \dots$ form an *i.i.d.-P_0 sample, where $P_0 = P_{\theta_0,\eta_0} \in \mathscr{P}$, then,*

$$\Pi\big(\|\theta - \theta_0\|_2 > M_n\varepsilon_n \bigm| X_1, \ldots, X_n \big) \longrightarrow 0 \text{ in } P_0^n\text{-probability}, \tag{4.50}$$

for every $M_n \to \infty$ as $n \to \infty$.

LEMMA 4.18 *Let $w : [0,1] \to \mathbb{R}$ be a continuous function and a weighted logistic density function on $[0,1]$ is defined as,*

$$p_w(t) = \frac{e^{w(t)} g(t)}{\int_0^1 e^{w(s)} g(s)\, ds},$$

where $g : [0,1] \to (0,\infty)$ be a known probability function and denote its associated distribution function by G. Then for any continuous functions $v, w : [0,1] \to \mathbb{R}$ we have the following,

$$d_H(p_v, p_w) \le \|v - w\|_\infty \times e^{\|v-w\|_\infty/2}.$$

Here we consider one example to illustrate the use of theorem 4.17. We particularly construct the priors on the infinite-dimensional space H and finite-dimensional space Θ of interest that meet (4.48) and the model condition (4.49) demonstrated in theorem 4.17 also holds.

EXAMPLE 4.19 Let a sequence of i.i.d sample $X_1, X_2, \ldots$ distributed from some unknown distribution function $P_0 = P_{\theta_0,\eta_0}$ with a Lebesgue density p_0 supported on an interval $[\theta_{0,1}, \theta_{0,2}]$ such that $0 < \theta_{0,2} - \theta_{0,1} < \delta$ for a known $\delta > 0$.

Given $M > 0$, define,

$$C_M := \left\{ h \in C[0,1] : e^{-M} \le h \le e^M \right\},$$

and,

$$H := \left\{ \eta(x) = \frac{g(x)\, h(x)}{\int_0^1 g(y)\, h(y)\, dy} : h \in C_M,\ x \in [0,1] \right\},$$

where $C[0,1]$ denotes the class of continuous functions on $[0,1]$ and $g : [0,1] \to (0,\infty)$ is a known density function. Note that any element in H is the Esscher transform of the function in C_M.

Take $h(x) = e^{Z(x)}$ for all $x \in [0,1]$, now a prior on nuisance space H is induced by considering a stochastic process prior $Z(x)$. Let U be a uniform random variable on $[-M,M]$ and $W = \{W(x) : x \in [0,1]\}$ be a Brownian motion independent of U. The stochastic process $Z(x)$ is defined as the process $U + W(x)$ conditioning on $U + W(x) \in [-M,M]$ for all $x \in [0,1]$. Hence the process Z has full support with respect to the uniform norm in C_M. Assume that a prior distribution Π_Θ on Θ admits a strictly positive and continuous Lebesgue density on its support. Of course, condition (4.47) is fulfilled in this case with f defined by,

$$f(\epsilon) = e^{-2M} \min\Big\{\int_0^\epsilon g(x)\,dx,\ \int_{1-\epsilon}^1 g(x)\,dx\Big\}, \quad (0 < \epsilon < 1).$$

Denote $P_0 = P_{\theta_0,\eta_0}$ and $Q = P_{\theta,\eta}$ with $\theta_0 = (\theta_{01},\theta_{02})$ and $\theta = (\theta_1,\theta_2)$, and put,

$$B_n := \{\theta \in \Theta, \eta \in H : \|P_0 - Q\|_{TV} < \delta_n\}.$$

Now, the corresponding densities can be explicitly written in the following forms:

$$\begin{aligned}
\eta(x) &= \frac{e^{U+W(x)}g(x)}{\int_0^1 e^{U+W(y)}g(y)\,dy} = \frac{e^{W(x)}g(x)}{\int_0^1 e^{W(y)}g(y)\,dy}, \qquad (4.51)\\
p_{\theta_{01},\theta_{02},\eta_0}(x) &= \frac{1}{\theta_{02}-\theta_{01}}\,\eta_0\Big(\frac{x-\theta_{01}}{\theta_{02}-\theta_{01}}\Big)\,1_{\{\theta_{01}\le x\le\theta_{02}\}},\\
p_{\theta_1,\theta_{02},\eta_0}(x) &= \frac{1}{\theta_{02}-\theta_1}\,\eta_0\Big(\frac{x-\theta_1}{\theta_{02}-\theta_1}\Big)\,1_{\{\theta_1\le x\le\theta_{02}\}},\\
p_{\theta_1,\theta_2,\eta_0}(x) &= \frac{1}{\theta_2-\theta_1}\,\eta_0\Big(\frac{x-\theta_1}{\theta_2-\theta_1}\Big)\,1_{\{\theta_1\le x\le\theta_2\}},\\
p_{\theta_1,\theta_2,\eta}(x) &= \frac{1}{\theta_2-\theta_1}\,\eta\Big(\frac{x-\theta_1}{\theta_2-\theta_1}\Big)\,1_{\{\theta_1\le x\le\theta_2\}}.
\end{aligned}$$

Consider $s = 1$, then,

$$\left\| \frac{dP_0}{dQ} - 1 \right\|_{s,Q} = \|P_0 - Q\|_{TV} = \int |p_0 - q| \, d\mu.$$

Notice that applying triangle inequality for the total variation distance between $p_{\theta_{01},\theta_{02},\eta_0}$ and $p_{\theta_1,\theta_2,\eta}$ yields,

$$\|p_{\theta_{01},\theta_{02},\eta_0} - p_{\theta_1,\theta_2,\eta}\|_1 \leq \|p_{\theta_{01},\theta_{02},\eta_0} - p_{\theta_1,\theta_{02},\eta_0}\|_1 + \|p_{\theta_1,\theta_{02},\eta_0} - p_{\theta_1,\theta_2,\eta_0}\|_1 + \|p_{\theta_1,\theta_2,\eta_0} - p_{\theta_1,\theta_2,\eta}\|_1,$$

that means the difference of these two densities in terms of total-variation metric is split into three parts that we will discuss in details in the sequel separately.

For the first part, let $\theta_1 < \theta_{01} < \theta_{02}$ and $\kappa = \frac{\theta_{01}-\theta_1}{\theta_{02}-\theta_{01}}$ together with $\kappa \leq \|\theta - \theta_0\| \leq \tilde{\delta}/2$, where $\tilde{\delta}$ is small enough. Note that,

$$\begin{aligned}
&\|p_{\theta_{01},\theta_{02},\eta_0} - p_{\theta_1,\theta_{02},\eta_0}\|_1 \\
&= \int_{\theta_{01}}^{\theta_{02}} \left| \frac{1}{\theta_{02}-\theta_{01}} \eta_0\Big(\frac{x-\theta_{01}}{\theta_{02}-\theta_{01}}\Big) - \frac{1}{\theta_{02}-\theta_1} \eta_0\Big(\frac{x-\theta_1}{\theta_{02}-\theta_{01}}\Big) \right| dx + \int_{\theta_1}^{\theta_{01}} \frac{1}{\theta_{02}-\theta_1} \eta_0\Big(\frac{x-\theta_1}{\theta_{02}-\theta_1}\Big) dx \\
&= \int_0^1 \Big| \eta_0(x) - \frac{1}{1+\kappa} \eta_0(x+\kappa) \Big| \, dx + \int_0^{\frac{\kappa}{1+\kappa}} \eta_0(x) \, dx \\
&= I + II.
\end{aligned}$$

Direct calculation for (I) and using the property of uniform continuous of $\eta_0(x)$ on $[0, 1]$ show that,

$$\begin{aligned}
I &= \int_0^1 \Big| \eta_0(x) - \frac{1}{1+\kappa} \eta_0(x+\kappa) \Big| \, dx \\
&\leq \frac{\kappa}{1+\kappa} \int_0^1 \eta_0(x) \, dx + \frac{1}{1+\kappa} \int_0^1 \Big| \eta_0(x+\kappa) - \eta_0(x) \Big| \, dx \\
&\leq \frac{2\kappa}{1+\kappa} \leq 2\|\theta - \theta_0\|.
\end{aligned}$$

Similarly one could find,

$$\begin{aligned} II &= \int_0^{\frac{\kappa}{1+\kappa}} \eta_0(x)\,dx \\ &\leq e^{2M}\frac{\kappa}{1+\kappa} \leq e^{2M}\|\theta-\theta_0\|. \end{aligned}$$

For the second part, let $\theta_1 < \theta_2 < \theta_{02},\ \tau = \dfrac{\theta_{02}-\theta_2}{\theta_2-\theta_1}$ and $\tau \leq \|\theta-\theta_0\| \leq \tilde{\delta}/2$, where $\tilde{\delta}$ is sufficiently small. Observe that,

$$\begin{aligned} &\|p_{\theta_1,\theta_{02},\eta_0} - p_{\theta_1,\theta_2,\eta_0}\|_1 \\ &= \int_{\theta_1}^{\theta_2}\left|\frac{1}{\theta_{02}-\theta_1}\eta_0\Big(\frac{x-\theta_1}{\theta_{02}-\theta_1}\Big) - \frac{1}{\theta_2-\theta_1}\eta_0\Big(\frac{x-\theta_1}{\theta_2-\theta_1}\Big)\right|\,dx + \int_{\theta_2}^{\theta_{02}}\frac{1}{\theta_{02}-\theta_1}\eta_0\Big(\frac{x-\theta_1}{\theta_{02}-\theta_1}\Big)\,dx \\ &= III + IV. \end{aligned}$$

A straightforward calculation for (III) and using the fact that $\eta_0(x)$ is uniform continuous on $[0,1]$ again yield that,

$$\begin{aligned} III &= \int_{\theta_1}^{\theta_2}\left|\frac{1}{\theta_{02}-\theta_1}\eta_0\Big(\frac{x-\theta_1}{\theta_{02}-\theta_1}\Big) - \frac{1}{\theta_2-\theta_1}\eta_0\Big(\frac{x-\theta_1}{\theta_2-\theta_1}\Big)\right|\,dx \\ &= \frac{1}{1+\tau}\int_0^1\left|\Big(\eta_0\Big(\frac{x}{1+\tau}\Big) - \eta_0(x)\Big) - \tau\eta_0(x)\right|\,dx \\ &\leq \frac{1}{1+\tau}\int_0^1\left|\Big(\eta_0\Big(\frac{x}{1+\tau}\Big) - \eta_0(x)\Big)\right|\,dx + \frac{\tau}{1+\tau}\int_0^1 \eta_0(x)\,dx \\ &\leq \frac{1}{1+\tau}\int_0^1 \tau\,dx + \frac{\tau}{1+\tau} \\ &= \frac{2\tau}{1+\tau} \leq 2\|\theta-\theta_0\|. \end{aligned}$$

Analogously simple algebra for (IV), one could get,

$$\begin{aligned}
IV &= \int_{\theta_2}^{\theta_{02}} \frac{1}{\theta_{02}-\theta_1}\,\eta_0\Big(\frac{x-\theta_1}{\theta_{02}-\theta_1}\Big)\,dx\\
&= \int_{\frac{\theta_2-\theta_1}{\theta_{02}-\theta_1}}^{1} \eta_0(x)\,dx = \int_{\frac{1}{\tau+1}}^{1} \eta_0(x)\,dx\\
&\le \frac{\tau}{1+\tau}\,e^{2M} \le e^{2M}\|\theta-\theta_0\|.
\end{aligned}$$

A direct computation for the third part guarantees that,

$$\begin{aligned}
\|p_{\theta_1,\theta_2,\eta_0}-p_{\theta_1,\theta_2,\eta}\|_1 &= \int_{\theta_1}^{\theta_2}\Big|\frac{1}{\theta_2-\theta_1}\,\eta_0\Big(\frac{x-\theta_1}{\theta_2-\theta_1}\Big) - \frac{1}{\theta_2-\theta_1}\,\eta\Big(\frac{x-\theta_1}{\theta_2-\theta_1}\Big)\Big|\,dx\\
&= \int_0^1 |\eta_0(x)-\eta(x)|\,dx\\
&= \|\eta_0-\eta\|_1\\
&\le 2h(\eta_0,\eta)\\
&\le 2\|W-w_0\|_\infty \times e^{\|W-w_0\|_\infty/2},
\end{aligned}$$

where the last inequality is due to lemma 4.18 and the fact that $\eta_0(x)=e^{w_0(x)+U_0}g(x)$.

Hence, the arguments in the preceding display imply that,

$$\begin{aligned}
\|P_0-Q\|_{TV} &= \|P_{\theta_0,\eta_0}-P_{\theta,\eta}\|_{TV}\\
&= \|p_{\theta_{01},\theta_{02},\eta_0}-p_{\theta_1,\theta_2,\eta_0}\|_1\\
&\le (4+2e^{2M})\|\theta-\theta_0\| + 2\|W-w_0\|_\infty \times e^{\|W-w_0\|_\infty/2}.
\end{aligned}$$

Let $K_0=4+2e^{2M}$ and $0<\delta_n<4\sqrt{e}$. Thus if,

$$\|\theta-\theta_0\| \le \frac{\delta_n}{2K_0} \quad \text{and} \quad \|W-w_0\|_\infty \le \delta_n/(2\sqrt{e}),$$

then it follows that,

$$K_0\|\theta - \theta_0\| + 2\|W - w_0\|_\infty \times e^{\|W-w_0\|_\infty/2} \leq \delta_n/2 + \delta_n/2 = \delta_n.$$

Moreover, we could explore the lower bound of the prior on B_n, i.e. $\Pi(B_n)$, as follows,

$$\begin{aligned}
\Pi(B_n) &\geq \Pi\Big(\theta \in \Theta, \eta \in H : \, K_0\|\theta - \theta_0\| + L_0^{'}\|W - w_0\|_\infty \times e^{\|W-w_0\|_\infty/2} < \delta_n\Big)\\
&\geq \Pi\Big(\theta \in \Theta, \eta \in H : \, \|\theta - \theta_0\| < \frac{\delta_n}{2K_0}, \|W - w_0\|_\infty \times e^{\|W-w_0\|_\infty/2} < \frac{\delta_n}{4}\Big)\\
&= \Pi_\Theta\Big(\theta \in \Theta : \, \|\theta - \theta_0\| < \frac{\delta_n}{2K_0}\Big) \times \Pi_W\Big(W : \, \|W - w_0\|_\infty \times e^{\|W-w_0\|_\infty/2} < \frac{\delta_n}{4}\Big)\\
&\geq \Pi_\Theta\Big(\theta \in \Theta : \, \|\theta - \theta_0\| < \frac{\delta_n}{2K_0}\Big) \times \Pi_W\Big(W : \, \|W - w_0\|_\infty < \frac{\delta_n}{4\sqrt{e}}\Big)\\
&\geq \Pi_\Theta\Big(\theta \in \Theta : \, \|\theta - \theta_0\| < \frac{\delta_n}{2K_0}\Big) \times \exp\Big\{-\frac{n\delta_n^2}{16e}\Big\}\\
&\geq \exp\{-K_2 n\delta_n^2\}.
\end{aligned}$$

where $K_2 > 0$ a sufficiently large constant. Since the prior probability density of θ is strictly positive and continuous on its support, the first factor of the last second inequality is bounded below by a constant multiple of δ_n^2. Furthermore, this factor could be absorbed into the second factor of this same inequality. Finally, it can be easily shown that the model condition (4.49) in theorem 4.17 follows since every η in H satisfies $e^{-2M}g(x) \leq \eta(x) \leq e^{2M}g(x)$ for all $x \in [0, 1]$ and a known density function $g(x)$.

4.6 Discussion

As far as the property of the posterior contraction rate is concerned, one of the main conditions appears to involve the GGV priors that charge a specialized Kullback-Leibler ball with a sufficient amount of probability mass. Especially, taking into account the difficulty to construct and even appropriately control the priors on infinite-dimensional spaces, a lack of flexibility in GGV priors then limits its application to a wide range of statistical models. For instance, any prior involved in boundary support estimation does not belong to this class of GGV priors.

To conclude, this present study is preliminary research on the flexible criteria on the prior at the cost of the relevant constraints on the model. A major finding is that we have explored the possibility of examining the contraction properties of the posterior distribution under some flexible condition for priors as well as stringent assumptions imposed on the model, followed by a comparison with that described in Ghosal et al. (2000). This flexible selection criteria permits us to tail the prior to act in accordance with Kullback-Leibler property stated in (4.1).

More specifically, it can be seen that the proposed approach encompasses the GGV priors and the posterior under the prior built on bracketing approximation attains the optimal rate of convergence in the context of the Bayesian nonparametric estimation of monotone nonincreasing densities, some class of suitably chosen probability densities and the similar notion of optimality is also demonstrated in Bayesian analysis of the interval censoring case 2 using a type of transformed Gaussian process prior. Furthermore, it can be reasoned that semiparametric estimation of the boundary support problem in violation of Kullback-Leibler requirement falls in this framework.

This kind of criteria for prior selection is still very much in its earlier stage and merits further investigation. Much more also needs to be known about the translation of this freedom to select priors into accommodating more choices for the loss functions, such as the uniform norm addressed in Giné and Nickl (2011) and Rousseau (2013). This study should yield the optimal rate of posterior contraction with respect to this uniform distance for future research.

Chapter 5

Future Works

5.1 Bayesian nonparametric estimation of functions subjects to shape restrictions

Shape restrictions often occur in a wide range of domains, such as monotonicity property entailing in the proportional hazard function models and the imposition of the monotonicity and concavity constraints on the indirect utility function playing a fundamental role in the estimation of the demand and cost functions and a single minimum attained in the electricity consumption model that characterizes the relationship between electricity demand and temperature, among others.

Both fixed-knot and free-knot regression spline models in a Bayesian context have been developed for the nonparametric estimation of functions in the presence of shape constraints in Shively et al. (2009) and Shively et al. (2011). In a frequentist side Wang and Ghosh (2012) proposed a general framework on a basis of the Bernstein polynomials for the estimation of the shape restricted nonparametric regression functions. More particularly, the Bernstein polynomials based estimators maintained the shape constraints such as monotonicity, convexity or concavity and simultaneous monotonicity and convexity via a series of linear restrictions on the associated coefficients.

This nice shape-preserving property that the Bernstein Polynomial enjoys could be further exploited in the context of regression function estimation in a Bayesian paradigm. The natural way to impose shape constraints in a Bayesian context is through the prior distributions on the coefficients of the Bernstein polynomial. Here we adopt the coarsened Bernstein polynomial proposed by Kruijer and van der Vaart (2008) that segment the terms involved in Bernstein polynomial expansion of the objective function into lower dimensional groups and keep the coefficients within the groups be equal to each other. The modified version also retained the desired properties of the original Bernstein polynomial. A bread spectrum of nonparametric priors would be imposed on the coefficients that satisfy some linear constraints arisen from the shape restrictions on the primitive function of interest.

Most studies up to now in the case have focus on some aspects of asymptotic properties such as posterior consistency and posterior contraction rate. However, far too little attention has been paid to the Bernstein-von Mises phenomenon in this case. Further research might establish the Bernstein-von Mises property in the context of shape restricted function estimation that enables us to construct approximate credible sets for the objective function which would be made feasible by the popular Markov chain Monte Carlo algorithms.

5.2 Bayesian semiparametric inference of linear model with conditional moment constraints

A linear regression model with conditional moment restrictions has been broadly documented in the field of econometrics. A semiparametric Bayesian approach could be used to perform an analysis of this model in the presence of conditional moment restrictions. The conditional distribution of the error on predictors is modelled to be covariate dependent via a finite mixture of normal distributions with the associated mixing probabilities be dependent on predictors.

One possible extension is to establish the Bernstein-von Mises property of this model by relaxing the assumptions in Norets (2015) that the regression error terms conditional on the covariates subjects to the normal distribution associated with the variance varying with predictors, namely, heteroskedasticity variance. We consider a more general case to assign a prior on this conditional distribution directly on a basis of the approach proposed in Shen and Ghosal (2014) instead of an induced prior on the distribution of error terms via endowing a probability distribution on the variance in Norets (2015). Then this setup readily eschews the misspecification for the data generating process of the error terms. It would be interesting to demonstrate that the limiting distribution of the parameters of interest is asymptotically normal with its variance that attains the semiparametric efficiency bound.

5.3 Bayesian nonparametric regression in the presence of endogeneity

The problem of endogeneity frequently occurs in the regression analysis of observational data when the covariates are correlated with the error term. A number of sources for the endogeneity issue include omitted variables, simultaneity and measurement errors. The estimation of regression coefficients by standard regression methods in the presence of endogeneity could be biased and inefficient.

The past decade has seen the rapid development of addressing the potential of the resulting endogeneity bias in the nonparametric regression context in a classical perspective. See Blundell and Powell (2003) for an extensive survey of nonparametric instrumental variable model (thereafter called NPIV) and Chen and Pouzo (2013) for a recent development for this topic. However, there have been a few investigations on asymptotic properties of posterior distributions in these models in the context of Bayesian nonparametrics and much of the research up to now are limited in a quasi-Bayesian method. Liao and Jiang (2011) proposed a unified framework for a general conditional moment restriction model, of which NPIV model is a special case, to establish the posterior consistency property on a basis of a quasi-Bayesian approach, and then Kato (2013) explored a quasi-Bayesian analysis of the model, with emphasis on the theoretical properties of quasi-posterior distributions. For the practical side, Wiesenfarth et al. (2014) presented a flexible nonparametric approach for this model with one endogenous regressor and implemented it via a Dirichlet process mixture prior.

One possible direction would be to investigate the pure Bayesian analysis of this model by endowing the nonparametric priors on the functions of interest. The respective asymptotic properties could be established in the context of this NPIV estimation.

Appendix A

Appendix to Chapter 2

A.1 Proof of Theorems

A.1.1 Proof of Lemma 2.1

PROOF OF LEMMA 2.1

We use $D_i^1 g$ to denote the partial derivative of the m-dimensional function g relative to the argument x_i for $i = 1, 2, \ldots, m$, where the partial derivative operator of the multivariate function has been defined in (3.7) of Chapter 3. An application of Taylor expansion of $g(\mathbf{x})$ at point $\mathbf{x} = \mathbf{z}$ together with its integral expression in (2.8) yields that,

$$\begin{aligned} g(\mathbf{x}) - b(\mathbf{x}; k, G) &= k^m \mathbb{E}\left(\int_{J_1/k}^{(J_1+1)/k} \cdots \int_{J_m/k}^{(J_m+1)/k} (g(\mathbf{x}) - g(\mathbf{z}))\, \mathrm{d}\mathbf{z} \right) \\ &= k^m \mathbb{E}\left(\int_{J_1/k}^{(J_1+1)/k} \cdots \int_{J_m/k}^{(J_m+1)/k} \left(\sum_{i=1}^{m} D_i^1 g|_{\mathbf{x}=\mathbf{z}} (x_i - z_i) + \frac{1}{2} (\mathbf{x} - \mathbf{z})^T H(\tilde{\mathbf{z}}) (\mathbf{x} - \mathbf{z}) \right) \mathrm{d}\mathbf{z} \right), \end{aligned}$$

where $\tilde{\mathbf{z}} := \mathbf{z} + \theta_{\mathbf{z}}(\mathbf{x} - \mathbf{z})$ for some $\theta_{\mathbf{z}} \in (0, 1)$.

Let $|trace\,(H(\tilde{\mathbf{z}}))| \leq M_1$ for some $M_1 > 0$ since the determinate of the Hessian matrix is bounded on $[0,1]^m$. Then one could obtain,

$$\begin{aligned}\left|\frac{1}{2}(\mathbf{x}-\mathbf{z})^T H(\tilde{\mathbf{z}})(\mathbf{x}-\mathbf{z})\right| &= \left|trace\left(\frac{1}{2}(\mathbf{x}-\mathbf{z})^T H(\tilde{\mathbf{z}})(\mathbf{x}-\mathbf{z})\right)\right| \\ &= \left|trace\left(\frac{1}{2}(\mathbf{x}-\mathbf{z})^T (\mathbf{x}-\mathbf{z}) H(\tilde{\mathbf{z}})\right)\right| \\ &= \left|\frac{1}{2}(\mathbf{x}-\mathbf{z})^T (\mathbf{x}-\mathbf{z})\, trace\left(H(\tilde{\mathbf{z}})\right)\right| \\ &\le \frac{M_1}{2}\sum_{i=1}^m (x_i - z_i)^2.\end{aligned}$$

Also the density $g(\mathbf{x})$ admits second partial derivative on $[0,1]^m$, then there exists a positive constant M_2 such that for $i = 1, 2, \ldots, m$ and $\mathbf{z} \in [0,1]^m$,

$$\left|D_i^1 g|_{\mathbf{x}=\mathbf{z}}\right| \le M_2.$$

Let $M_3 = \max(M_1/2, M_2)$, then,

$$\begin{aligned}&\left|\sum_{i=1}^m D_i^1 g|_{\mathbf{x}=\mathbf{z}}(x_i - z_i) + \frac{1}{2}(\mathbf{x}-\mathbf{z})^T H(\tilde{\mathbf{z}})(\mathbf{x}-\mathbf{z})\right| \\ &\le \sum_{i=1}^m \left[M_2|x_i - z_i| + \frac{M_1}{2}(x_i - z_i)^2\right] \\ &\le M_3 \sum_{i=1}^m \left[|x_i - z_i| + (x_i - z_i)^2\right].\end{aligned}$$

Hence,

$$\begin{aligned}
|g(\mathbf{x}) - b(\mathbf{x}; k, G)| &= \left| k^m \mathbb{E} \left(\int_{J_1/k}^{(J_1+1)/k} \dots \int_{J_m/k}^{(J_m+1)/k} (g(\mathbf{x}) - g(\mathbf{z}))\, \mathrm{d}\mathbf{z} \right) \right| \\
&\le M_3 k^m \mathbb{E} \left(\int_{J_1/k}^{(J_1+1)/k} \dots \int_{J_m/k}^{(J_m+1)/k} \sum_{i=1}^{m} \left[|x_i - z_i| + (x_i - z_i)^2 \right] \mathrm{d}\mathbf{z} \right) \\
&= M_3 k \mathbb{E} \left(\int_{J_1/k}^{(J_1+1)/k} \left[|x_1 - z_1| + (x_1 - z_1)^2 \right] \mathrm{d}z_1 \right) \\
&= O(1/k),
\end{aligned}$$

where the last step follows the similar arguments as (2.1) in Ghosal (2001). Thus we have proved this lemma. □

A.1.2 Proof of Theorem 2.2

PROOFS OF THEOREM 2.2

Define $K(f_0, f) = \int \log \frac{f_0}{f} dP_0$, $V(f_0, f) = \int \left(\log \frac{f_0}{f}\right)^2 dP_0$ and $N(\epsilon, f_0) = \{f : K(f_0, f) \le \epsilon^2, V(f_0, f) \le \epsilon^2\}$.

THEOREM A.1 (Ghosal (2001)) *Let Π_n be a sequence of priors on $\mathcal{F}$. Suppose that for positive sequences $\bar{\epsilon}_n, \tilde{\epsilon}_n \to 0$ with $n \min(\bar{\epsilon}_n, \tilde{\epsilon}_n)^2 \to \infty$, constants $c_1, c_2, c_3, c_4 > 0$ and sets $\mathcal{F}_n \subset \mathcal{F}$, we have,*

$$\begin{aligned}
\log D(\bar{\epsilon}_n, \mathcal{F}_n, d) &\le c_1 n \bar{\epsilon}_n^2, & \text{(A.1)} \\
\Pi_n(\mathcal{F} \setminus \mathcal{F}_n) &\le c_3 \exp(-(c_2 + 4) n \tilde{\epsilon}_n^2), & \text{(A.2)} \\
\Pi_n(N(\tilde{\epsilon}_n, f_0)) &\ge c_4 \exp(-c_2 n \tilde{\epsilon}_n^2). & \text{(A.3)}
\end{aligned}$$

Then for $\epsilon_n = \max(\tilde{\epsilon}_n, \bar{\epsilon}_n)$ and a sufficiently large $M > 0$, we have

$$\Pi_n(f : d_H(f, f_0) > M\epsilon_n | X_1, ..., X_n) \to 0 \quad \textit{in } P_0^n \textit{ probability.}$$

For $k \ge 1$, let $f_k(x_1, x_2, \dots, x_m) = b(x_1, x_2, \dots, x_m; k, F_0)$, where F_0 is the cumulative distribution function of f_0. Note that f_k is uniformly bounded away from 0 for all large k by (2.7).

Now let $\boldsymbol{\omega}_k^{m0}$ be the set of weights associated with $b(x_1, ..., x_m; k, F_0)$ and let $\boldsymbol{\omega}_k^{m}$ be the weights associated with $b(x_1, ..., x_m; k, F)$ in (2.9). Then,

$$\begin{aligned} \mid b(x_1, x_2, \ldots, x_m; k, F) - b(x_1, x_2, \ldots, x_m; k, F_0) \mid &\leq k^m \max_{1 \leq i_1, \ldots, i_m \leq k} |w_{i_1 i_2 \ldots i_m; k} - w^0_{i_1 i_2 \ldots i_m; k}| \\ &\leq k^m \|\boldsymbol{\omega}_k^m - \boldsymbol{\omega}_k^{m0}\|_1. \qquad \text{(A.4)} \end{aligned}$$

Now, if $\|\boldsymbol{\omega}_k^m - \boldsymbol{\omega}_k^{m0}\|_1 \leq \epsilon^{m+1}$ and $d_1 \varepsilon^{-1} \leq k \leq d_2 \varepsilon^{-1}$ for positive constants d_1 and d_2, then $\sup\limits_{0 < x_1, x_2, \ldots, x_m \leq 1} |f_0(x_1, x_2, \ldots, x_m) - b(x_1, x_2, \ldots, x_m; k, F)| \leq D_1 \varepsilon$ for some positive constant D_1 and $b(x_1, \ldots, x_m; k, F)$ is bounded away from 0 for sufficiently small ϵ. Now, for some positive constant D_2, $d_H(f_0, b(\cdot, \cdots, \cdot; k, F)) \leq D_2 \epsilon$ and so (8.6) of Ghosal et al. (2000) implies that $b(\cdot, \cdots, \cdot; k, F) \in N(C_1 \epsilon, f_0)$ for some positive constant C_1. Hence,

$$N(C_1 \varepsilon, f_0) \supset \{b(x_1, \ldots, x_m; k, F) : \|\boldsymbol{\omega}_k^m - \boldsymbol{\omega}_k^{m0}\|_1 \leq \varepsilon^{m+1}\}.$$

Now let k_n be such that,

$$b_1 \left(\frac{n}{\log n} \right)^{1/(m+2)} \leq k_n \leq b_2 \left(\frac{n}{\log n} \right)^{1/(m+2)},$$

for positive constants b_1 and b_2 and $\tilde{\varepsilon}_n = k_n^{-1}$. Then Lemma A.1 of the Appendix in Ghosal (2001) implies that there exist positive constants C_3, C_4, D and d such that,

$$\begin{aligned} \Pi(N(C_1 \tilde{\epsilon}_n, f_0)) &\geq p(k_n) C_2 \exp(-C_3 k_n^m \log(1/\tilde{\epsilon}_n)) \\ &\geq B_1 \exp(-\beta_1 (1/\tilde{\varepsilon}_n)^m) \times C_2 \exp(-C_3 (1/\tilde{\varepsilon}_n)^m \log(1/\tilde{\epsilon}_n)) \\ &\geq D \exp(-d (1/\tilde{\epsilon}_n)^m \log(1/\tilde{\varepsilon}_n)). \end{aligned}$$

Hence $\tilde{\epsilon}_n = n^{-1/(m+2)}(\log n)^{1/(m+2)}$ satisfies condition (A.3). Let s_n be an integer such that,

$$L_1(1/\tilde{\varepsilon}_n)^m \log(1/\tilde{\varepsilon}_n) \leq s_n \leq L_2(1/\tilde{\epsilon}_n)^m \log(1/\tilde{\epsilon}_n),$$

for positive constants L_1 and L_2. Now note that,

$$L_1^{'}\, n^{m/(m+2)}(\log n)^{2/(m+2)} \leq s_n \leq L_2^{'}\, n^{m/(m+2)}(\log n)^{2/(m+2)},$$

where we may choose $L_1^{'} = \frac{L_1}{2m+4}$ and $L_2^{'} = \frac{L_2}{m+2}$. Let $\mathcal{F}_n := \bigcup_{r=1}^{\sqrt[m]{s_n}} \mathscr{B}_r^m$. Note that for positive constants B_3 and L,

$$\Pi(\mathcal{F}_n^c) \leq \sum_{r=\sqrt[m]{s_n}+1}^{\infty} \rho(r) \leq \sum_{r=\sqrt[m]{s_n}+1}^{\infty} B_2 e^{-\beta_2 r^m} \leq B_3 e^{-\beta_2 s_n} \leq B_3 \exp(-L(1/\tilde{\epsilon}_n)^m \log(1/\tilde{\epsilon}_n)).$$

and L can be made as large as required by choosing L_1 large enough. As $(1/\tilde{\epsilon}_n)^m \log(1/\tilde{\varepsilon}_n)$ and $n\tilde{\varepsilon}_n^2$ have the same order, the condition (A.2) holds. Observe also that by virtue of (A.4), the covering number $D(\epsilon, \mathscr{B}_r^m, d)$ is bounded from above by the corresponding covering number of the unit l_1-simplex in $\mathbb{R}^{r^m}$ for $r = 1, 2, \ldots$, that is,

$$D(\epsilon, \mathscr{B}_r^m, \|\cdot\|_1) \leq D(\epsilon/r^m, \Delta_r^m, \|\cdot\|_1).$$

Moreover, in view of Lemma A.4 in Ghosal and van der Vaart (2001), one could get,

$$D(\epsilon/r^m, \Delta_r^m, \|\cdot\|_1) \leq \left(\frac{5\, r^m}{\epsilon}\right)^{r^m-1}.$$

Hence,

$$\begin{aligned}
D(\epsilon, \mathcal{F}_n, \|\cdot\|_1) &\leq \sum_{r=1}^{\sqrt[m]{s_n}} D(\epsilon, \mathscr{B}_r^m, \|\cdot\|_1) \\
&\leq \sum_{r=1}^{\sqrt[m]{s_n}} \left(\frac{5\, r^m}{\epsilon}\right)^{r^m - 1} \\
&\leq \sqrt[m]{s_n} \left(\frac{5\, s_n}{\epsilon}\right)^{s_n - 1}.
\end{aligned}$$

Then it follows that for some positive constants C and L_3,

$$\begin{aligned}
\log D(\epsilon, \mathcal{F}_n, \|\cdot\|_1) &\leq \frac{1}{m} \log s_n + s_n \log \frac{5\, s_n}{\epsilon} \\
&\leq C s_n \log \frac{5\, s_n}{\epsilon} \\
&\leq L_3\, n^{m/(m+2)} (\log n)^{2/(m+2)} \log \frac{5\, s_n}{\epsilon}.
\end{aligned}$$

Since the topology generated by the Hellinger distance and L_1-norm in $\mathscr{P}$ is equivalent, then the choice $\epsilon = \bar{\epsilon}_n = n^{-1/(m+2)} (\log n)^{(m+4)/(2m+4)}$ satisfies condition (A.1) and the posterior converges at the rate $n^{-1/(m+2)} (\log n)^{(m+4)/(2m+4)}$. Hence the proof of this theorem is complete. □

Appendix B

Appendix to Chapter 3

B.1 Useful Lemmas for Chapter 3

To prove the main theorems in Section 3.4, we need the following supplementary lemmas. For brevity of notations, we use the generic positive constant C throughout this Appendix.

LEMMA B.1 *If $x > 0$, then the following inequality holds.*

$$1 - \sqrt{\frac{2x}{x^2+1}} \leq \log x^2 - 1 + \frac{1}{x^2}. \tag{B.1}$$

PROOF OF LEMMA B.1

Let us introduce a new function $f(x)$ as follows,

$$h(x) = \log x^2 + \frac{1}{x^2} + \sqrt{\frac{2x}{x^2+1}}. \tag{B.2}$$

The claim holds if $h(x) \geq 2$ for all $x > 0$. Note that the first derivative of $h(x)$ could be written as,

$$h'(x) = 2(x^2-1)\left(\frac{1}{x^3} - \frac{1}{2(x^2+1)\sqrt{2x(x^2+1)}}\right).$$

Noting also that,

$$\frac{2(x^2+1)\sqrt{2x(x^2+1)}}{x^3} = 2\sqrt{2}\sqrt{x\left(1+\frac{1}{x^2}\right)}\left(1+\frac{1}{x^2}\right) \geq 2\sqrt{2}\sqrt{x\frac{1}{2x}} \times 1 = 2 > 1.$$

Hence $h'(x) \geq 0$ if $x \geq 1$ and $h'(x) < 0$ otherwise. That is to say, $h(x)$ attains the minimum at $x = 1$. Using the fact that $h(1) = 2$ we then obtain $h(x) \geq 2$ for all $x > 0$. So the proof of this lemma is complete. □

The following lemma states that the order of the Hellinger distance between (β_1, σ_1) and (β_2, σ_2) is controlled by the Euclidean distance of finite-dimensional parametric parts β_1 and β_2 as well as the uniform norm of the difference on infinite-dimensional parts σ_1 and σ_2.

LEMMA B.2 *Let $\lambda_{max}(E(X_iX_i'))$ be bounded by some positive constant m_2 for $i = 1, 2, \ldots, n$, then we have,*

$$\begin{aligned} d_H^2(\eta_1,\eta_2) &= 2 - 2\int_{\mathscr{X}} \exp\left\{-\frac{((\beta_1-\beta_2)^T\boldsymbol{x})^2}{4(\sigma_1^2(\boldsymbol{x})+\sigma_2^2(\boldsymbol{x}))}\right\}\sqrt{\frac{2\sigma_1(\boldsymbol{x})\sigma_2(\boldsymbol{x})}{\sigma_1^2(\boldsymbol{x})+\sigma_2^2(\boldsymbol{x})}}\,dG_0(\boldsymbol{x}) \\ &\leq \frac{m_2}{4\underline{\sigma}^2}\|\beta_1-\beta_2\|_2^2 + \frac{1}{4}z\left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2}\right)\frac{\overline{\sigma}^2}{\underline{\sigma}^4}\sup_{\boldsymbol{x}\in\mathscr{X}}|\sigma_1(\boldsymbol{x})-\sigma_2(\boldsymbol{x})|^2. \end{aligned} \tag{B.3}$$

PROOF OF LEMMA B.2

An application of the elementary inequality $1 - ab \leq 1 - a + 1 - b$ for $a \leq 1$ and $b \leq 1$ yields,

$$\begin{aligned} d_H^2(\eta_1,\eta_2) &= 2 - 2\int_{\mathscr{X}} \exp\left\{-\frac{((\beta_1-\beta_2)^T\boldsymbol{x})^2}{4(\sigma_1^2(\boldsymbol{x})+\sigma_2^2(\boldsymbol{x}))}\right\}\sqrt{\frac{2\sigma_1(\boldsymbol{x})\sigma_2(\boldsymbol{x})}{\sigma_1^2(\boldsymbol{x})+\sigma_2^2(\boldsymbol{x})}}\,dG_0(\boldsymbol{x}) \\ &\leq \int_{\mathscr{X}} 2\left(1-\exp\left\{-\frac{((\beta_1-\beta_2)^T\boldsymbol{x})^2}{4(\sigma_1^2(\boldsymbol{x})+\sigma_2^2(\boldsymbol{x}))}\right\}\right) + 2\left(1-\sqrt{\frac{2\sigma_1(\boldsymbol{x})\sigma_2(\boldsymbol{x})}{\sigma_1^2(\boldsymbol{x})+\sigma_2^2(\boldsymbol{x})}}\right)dG_0(\boldsymbol{x}) \\ &\leq \int_{\mathscr{X}} \left\{\frac{((\beta_1-\beta_2)^T\boldsymbol{x})^2}{2(\sigma_1^2(\boldsymbol{x})+\sigma_2^2(\boldsymbol{x}))} + \log\left(\frac{\sigma_1^2(\boldsymbol{x})}{\sigma_2^2(\boldsymbol{x})}\right) - 1 + \frac{\sigma_2^2(\boldsymbol{x})}{\sigma_1^2(\boldsymbol{x})}\right\}dG_0(\boldsymbol{x}) \\ &\leq \frac{1}{4\underline{\sigma}^2}\lambda_{max}(E(X_iX_i'))\|\beta_1-\beta_2\|_2^2 + \frac{1}{4}z\left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2}\right)\frac{\overline{\sigma}^2}{\underline{\sigma}^4}\sup_{\boldsymbol{x}\in\mathscr{X}}|\sigma_1(\boldsymbol{x})-\sigma_2(\boldsymbol{x})|^2, \end{aligned}$$

where the penultimate inequality follows from the elementary inequality $1-e^{-x} \leq x$ for $x \geq 0$ and

lemma B.1. Thus the assertion follows by the assumption $\lambda_{max}(E(X_iX_i')) \leq m_2$ for $i = 1, 2, \ldots, n$.
□

The following lemma states that we could bound the first and second moments of log-likelihood ratio from above.

LEMMA B.3 *Let $\lambda_{max}(E(X_iX_i')) \leq m_2$, where $m_2 > 0$, then the following inequalities hold.*

$$K(\eta, \eta_0) \leq m_3 \left(\sup_{\boldsymbol{x} \in \mathscr{X}} |\sigma(\boldsymbol{x}) - \sigma_0(\boldsymbol{x})|^2 + \|\beta - \beta_0\|_2^2 \right), \tag{B.4}$$

$$V(\eta, \eta_0) \leq m_4 \left(\sup_{\boldsymbol{x} \in \mathscr{X}} |\sigma(\boldsymbol{x}) - \sigma_0(\boldsymbol{x})|^2 + \|\beta - \beta_0\|_2^2 \right). \tag{B.5}$$

PROOF OF LEMMA B.3

A straightforward computation for $K(\eta, \eta_0)$ shows that,

$$\begin{aligned}
K(\eta, \eta_0) &= \int_{\mathscr{X}} \int_{\mathscr{Y}} f_{\boldsymbol{x}\eta_0}(y) \log \frac{f_{\boldsymbol{x}\eta_0}(y)}{f_{\boldsymbol{x}\eta}(y)} \, dy \, dG_0(\boldsymbol{x}) \\
&= \int_{\mathscr{X}} \int_{\mathscr{Y}} f_{\boldsymbol{x}\eta_0}(y) \frac{1}{2} \left\{ \log \frac{\sigma^2(\boldsymbol{x})}{\sigma_0^2(\boldsymbol{x})} - \frac{(y - \beta_0'\boldsymbol{x})^2}{\sigma_0^2(\boldsymbol{x})} + \frac{(y - \beta'\boldsymbol{x})^2}{\sigma^2(\boldsymbol{x})} \right\} dy \, dG_0(\boldsymbol{x}) \\
&= \int_{\mathscr{X}} \frac{1}{2} \left\{ \log \frac{\sigma^2(\boldsymbol{x})}{\sigma_0^2(\boldsymbol{x})} - 1 \right\} dG_0(\boldsymbol{x}) + \int_a^b \int_{\mathscr{Y}} f_{\boldsymbol{x}\eta_0}(y) \left\{ \frac{1}{2\sigma^2(\boldsymbol{x})} (y - \beta_0'\boldsymbol{x} + \beta_0'\boldsymbol{x} - \beta'\boldsymbol{x})^2 \right\} dy \, dG_0(\boldsymbol{x}) \\
&= \int_{\mathscr{X}} \frac{1}{2} \left\{ \log \frac{\sigma^2(\boldsymbol{x})}{\sigma_0^2(\boldsymbol{x})} - 1 + \frac{\sigma_0^2(\boldsymbol{x})}{\sigma^2(\boldsymbol{x})} + \frac{1}{\sigma^2(\boldsymbol{x})} (\beta_0 - \beta)'\boldsymbol{x}\boldsymbol{x}'(\beta_0 - \beta) \right\} dG_0(\boldsymbol{x}) \\
&\leq \; 2z \left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2} \right) \frac{\overline{\sigma}^2}{\underline{\sigma}^4} \sup_{\boldsymbol{x} \in \mathscr{X}} |\sigma(\boldsymbol{x}) - \sigma_0(\boldsymbol{x})|^2 + \underline{\sigma}^{-2} \lambda_{max}(E(X_iX_i')) \|\beta - \beta_0\|_2^2,
\end{aligned}$$

where the final line follows from lemma B.5. Thus the assertion (B.4) follows by taking,

$$m_3 = \left\{ 2z \left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2} \right) \frac{\overline{\sigma}^2}{\underline{\sigma}^4}, \, \underline{\sigma}^{-2} \lambda_{max}(E(X_iX_i')) \right\}.$$

For $V(\eta, \eta_0)$, simple algebra delivers that,

$$\begin{aligned} V(\eta,\eta_0) &= \int_{\mathscr{X}} \int_{\mathscr{Y}} f_{\boldsymbol{x}\eta_0}(y) \left(\log \frac{f_{\boldsymbol{x}\eta_0}(y)}{f_{\boldsymbol{x}\eta}(y)} \right)^2 dy \, dG_0(\boldsymbol{x}) \\ &= \int_{\mathscr{X}} \left\{ \left(\frac{\sigma_0^2(\boldsymbol{x})}{\sigma^2(\boldsymbol{x})} - 1 \right)^2 + \frac{\sigma_0^4(\boldsymbol{x})}{\sigma^4(\boldsymbol{x})} (\beta_0 - \beta)' \boldsymbol{x}\boldsymbol{x}' (\beta_0 - \beta) \right\} dG_0(\boldsymbol{x}) \\ &\leq \underline{\sigma}^{-2} \int_{\mathscr{X}} (\sigma^2(\boldsymbol{x}) - \sigma_0^2(\boldsymbol{x}))^2 \, dG_0(\boldsymbol{x}) + \left(\frac{\overline{\sigma}}{\underline{\sigma}} \right)^4 \lambda_{max}(E(XX')) \|\beta - \beta_0\|_2^2 \\ &\leq \frac{4\overline{\sigma}^2}{\underline{\sigma}^2} \sup_{\boldsymbol{x} \in \mathscr{X}} |\sigma(\boldsymbol{x}) - \sigma_0(\boldsymbol{x})|^2 + \left(\frac{\overline{\sigma}}{\underline{\sigma}} \right)^4 \lambda_{max}(E(X_i X_i') \|\beta - \beta_0\|_2^2. \end{aligned}$$

Here we let,

$$m_4 = \left\{ \frac{4\overline{\sigma}^2}{\underline{\sigma}^2}, \left(\frac{\overline{\sigma}}{\underline{\sigma}} \right)^4 \lambda_{max}(E(X_i X_i')) \right\},$$

therefore the assertion (B.5) follows. □

An immediate consequence from lemma B.3 implies that the following result holds.

COROLLARY B.4 *Under the condition described in lemma B.3, we have,*

$$\max\left\{ K(\eta,\eta_0),\, V(\eta,\eta_0) \right\} \leq m_5 \left(\sup_{\boldsymbol{x} \in \mathscr{X}} |\sigma(\boldsymbol{x}) - \sigma_0(\boldsymbol{x})|^2 + \|\beta - \beta_0\|_2^2 \right), \tag{B.6}$$

for some positive constant m_5.

LEMMA B.5 *Let $z(t) = \dfrac{t - 1 - \log t}{(t-1)^2}$ be a positive decreasing function on $(0, \infty)$, then for any $t \in \left[\frac{\underline{\sigma}^2}{\overline{\sigma}^2}, \frac{\overline{\sigma}^2}{\underline{\sigma}^2} \right]$, the following inequality holds,*

$$\frac{4\underline{\sigma}^2}{\overline{\sigma}^4} z\left(\frac{\overline{\sigma}^2}{\underline{\sigma}^2} \right) \tilde{d}_2^2(\sigma, \sigma_0) \leq \int_a^b \left(\frac{\sigma_0^2(\boldsymbol{x})}{\sigma^2(\boldsymbol{x})} - 1 - \log \frac{\sigma_0^2(\boldsymbol{x})}{\sigma^2(\boldsymbol{x})} \right) dG_0(\boldsymbol{x}) \leq \frac{4\overline{\sigma}^2}{\underline{\sigma}^4} z\left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2} \right) \tilde{d}_2^2(\sigma, \sigma_0),$$

where $\tilde{d}_2^2(\sigma, \sigma_0) = \int_a^b (\sigma(\boldsymbol{x}) - \sigma_0(\boldsymbol{x}))^2 \, dG_0(\boldsymbol{x})$.

PROOF OF LEMMA B.5

Observe that,

$$(t-1)^2 \, z\left(\frac{\overline{\sigma}^2}{\underline{\sigma}^2} \right) \leq t - 1 - \log t \leq (t-1)^2 \, z\left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2} \right).$$

Let $t = \frac{\sigma_0^2(\boldsymbol{x})}{\sigma^2(\boldsymbol{x})}$ and notice that,

$$\frac{(\sigma^2(X)-\sigma_0^2(X))^2}{\sigma^4(X)}\, z\left(\frac{\overline{\sigma}^2}{\underline{\sigma}^2}\right) \leq \frac{\sigma_0^2(X)}{\sigma^2(X)} - 1 - \log\frac{\sigma_0^2(X)}{\sigma^2(X)} \leq \frac{(\sigma^2(X)-\sigma_0^2(X))^2}{\sigma^4(X)}\, z\left(\frac{\underline{\sigma}^2}{\overline{\sigma}^2}\right).$$

Therefore the claim follows by taking expectation with respect to the distribution function $G_0(\boldsymbol{x})$ on the inequality above. □

LEMMA B.6 *Given $0 < \alpha \leq q$, and for each function $f \in C^\alpha[(0,1)^d]$, there exist some $\theta \in \mathbb{R}^{J^d}$ and a positive constant C that depends solely on q such that,*

$$\|f - \theta^T\xi\|_\infty \leq CJ^{-\alpha}\|D^{(\alpha)}f\|_\infty.$$

Futhermore, if $\underline{\sigma} < f < \overline{\sigma}$, every element of θ could be chosen to be between $\underline{\sigma}$ and $\overline{\sigma}$.

PROOF OF LEMMA B.6

The first part is as same as that in Lemma 1 in Shen and Ghosal (2012). And the proof of the second part goes throughout the part (b) of Lemma 1 in Shen and Ghosal (2012) by choosing $\tilde{f} = f - \underline{\sigma}$ and $\tilde{g} = \overline{\sigma} - f$. □

The following two lemmas state that the approximation error of the transform stochastic process could be controlled by the corresponding primitive process with respective to the uniform norm.

LEMMA B.7

$$\sup_{x\in\mathscr{X}}\left|\tilde{\Psi}(W(\boldsymbol{x})) - \tilde{\Psi}(w_0(\boldsymbol{x}))\right| \leq C\sup_{\boldsymbol{x}\in\mathscr{X}}|W(\boldsymbol{x}) - w_0(\boldsymbol{x})|\,. \tag{B.7}$$

B.2 Proof of Theorems

B.2.1 Proof of Theorem 3.2

Here we provide the proof of all the results developed in Section 3.4.

PROOF OF THEOREM 3.2

We apply Theorem 4 of Ghosal and van der Vaart (2007a) to prove this theorem in a similar manner as Lin and Dunson (2014). In particular, let,

$$V_n = \{\sigma = \tilde{\Psi}(W) : W \in U_n\}, \tag{B.8}$$

where U_n is a the measurable subset described in Theorem B.14. Now we determine the upper bound on the entropy number on the sieve of the support of the product prior $\Pi_\eta = \Pi_\sigma \times \Pi_\beta$. Define,

$$\mathscr{F}_n = \left\{(\sigma, \beta) : \sigma \in V_n, \, \beta \in [\underline{\beta}, \overline{\beta}]^d \right\}. \tag{B.9}$$

Since $\tilde{\Psi}$ is a one-to-one map from $\mathbb{R}$ to $[\underline{\sigma}, \overline{\sigma}]$, then $V_n \subset U_n$. Hence the number of ε_n-balls needed to cover V_n is less than U_n in terms of the uniform distance. That is,

$$\log N(\varepsilon_n, V_n, \|\cdot\|_\infty) \leq \log N(\varepsilon_n, U_n, \|\cdot\|_\infty), \tag{B.10}$$

which is bounded by $Dn\varepsilon_n^2$ by (B.45). To bound from above the entropy number on $\mathscr{F}_n$, we consider the covering number of the one-dimensional set $\{\beta_1 : \beta_1 \in [\underline{\beta}, \overline{\beta}]\}$. Let $N = \left\{\left[\frac{\overline{\beta}-\underline{\beta}}{2\varepsilon_n}\right] + 1\right\}$, the interval $[\underline{\beta}, \overline{\beta}]$ could be partitioned into N sub-intervals with the equal length $\frac{\overline{\beta}-\underline{\beta}}{N}$. We denote all the middle points of these equidistant intervals by the set,

$$T = \left\{\underline{\beta} + i\,\frac{\overline{\beta} - \underline{\beta}}{2N} : i = 1, 3, \ldots, 2N-1\right\}.$$

Then every equidistant interval could be covered by one neighborhood of some point in T with radius $\bar{\varepsilon}_n$. Thus the covering number of the set $\{\beta : \beta \in [\underline{\beta}, \overline{\beta}]^d\}$ is,

$$N\left(\varepsilon_n d^{1/2}, \, [\underline{\beta}, \overline{\beta}]^d, \, \|\cdot\|_2\right) \leq N^d.$$

In view of (B.3), observe that if $\sup_{\boldsymbol{x}\in\mathscr{X}} |\sigma(\boldsymbol{x}) - \sigma_0(\boldsymbol{x})| \leq C\varepsilon_n$ and $\|\beta - \beta_0\|_2 \leq \varepsilon_n d^{1/2}$, then we have that,

$$\begin{aligned} d_H^2(\eta, \eta_0) &\lesssim \sup_{\boldsymbol{x}\in\mathscr{X}} |\sigma(\boldsymbol{x}) - \sigma_0(\boldsymbol{x})|^2 + \|\beta - \beta_0\|_2^2, \\ &\leq \varepsilon_n^2(C^2 + d). \end{aligned}$$

Therefore, the $\varepsilon_n(C^2+d)^{1/2}$-covering number of $\mathscr{F}_n$ is bounded by $e^{Dn\varepsilon_n^2} \times N^d$, that is,

$$\log N\left(\varepsilon_n(C^2+d)^{1/2},\, \mathscr{F}_n,\, d_H\right) \leq Dn\varepsilon_n^2 + d\log N.$$

Using the assumption $\log\left\{\left\lceil \frac{\overline{\beta}-\underline{\beta}}{2\varepsilon_n}\right\rceil + 1\right\} \leq n\varepsilon_n^2$ we obtain,

$$\log N\left((C^2+d)^{1/2}\varepsilon_n,\, \mathscr{F}_n,\, d_H\right) \leq (D+d)n\varepsilon_n^2.$$

We proceed to show that the prior Π_η assigns a large amount of probability mass on some specialized Kullback-Leibler ball of the true value η_0. Let,

$$B^*(\eta_0, \varepsilon_n) = \{\eta :\ K(\eta, \eta_0) < \varepsilon_n^2,\, V(\eta, \eta_0) < \varepsilon_n^2\}. \tag{B.11}$$

We need to bound from below $\Pi(B^*(\eta_0, \varepsilon_n))$. By corollary B.4, it follows that,

$$B^*(\eta_0, \varepsilon_n) \supset \left\{\eta = (\beta, \sigma) :\ \|\beta - \beta_0\|_2 \leq \frac{\tilde{D}\varepsilon_n}{2},\, \|\sigma - \sigma_0\|_\infty \leq \frac{\tilde{D}\varepsilon_n}{2}\right\}. \tag{B.12}$$

for some constant $\tilde{D}$. Therefore the prior mass on $B^*(\eta_0, \varepsilon_n)$ could be lower bounded by,

$$\Pi_\sigma\left(\|\sigma - \sigma_0\|_\infty \leq \frac{\tilde{D}\varepsilon_n}{2}\right) \times \Pi_\beta\left(\|\beta - \beta_0\|_2 \leq \frac{\tilde{D}\varepsilon_n}{2}\right).$$

Applying lemma B.7 gives rise to,

$$\Pi_\sigma\left(\|\sigma-\sigma_0\|_\infty \le \frac{\tilde{D}\varepsilon_n}{2}\right) \ge \Pi_W\left(\|W-w_0\|_\infty \le \frac{\tilde{D}\varepsilon_n}{2C}\right),$$

which is greater than $\exp\left\{-\frac{\tilde{D}^2 n\varepsilon_n^2}{16C^2}\right\}$. In view of the assumption on the prior of β, we have ,

$$\begin{aligned}\Pi(B^*(\eta_0,\varepsilon_n)) &\ge \Pi_\sigma\left(\|\sigma-\sigma_0\|_\infty \le \frac{\tilde{D}\varepsilon_n}{2}\right)\times \Pi_\beta\left(\|\beta-\beta_0\|_2 \le \frac{\tilde{D}\varepsilon_n}{2}\right),\\ &\ge \exp\left\{-\frac{\tilde{D}^2 n\varepsilon_n^2}{16C^2}\right\}\times\exp(-\bar{D}n\varepsilon_n^2),\\ &\ge \exp(-\tilde{D}_1 n\varepsilon_n^2),\end{aligned}$$

for some positive constants $\bar{D}$ and $\tilde{D}_1$.

It remains to show that prior on the complement of the sieve is negligible. In fact, since $\{\eta : \eta \notin \mathscr{F}_n\} \subset \{\sigma : \sigma \notin V_n\}$, it is easy to say, by (B.44),

$$\Pi_\eta\{\eta : \eta\notin\mathscr{F}_n\} \subset \Pi_\sigma\{\sigma : \sigma\notin V_n\} \subset \Pi_W\{W : W\notin U_n\} \le \exp\{-n\varepsilon_n^2\}. \tag{B.13}$$

So the claim follows since all the three key conditions listed in Theorem 4 of Ghosal and van der Vaart (2007a) are satisfied. □

In order to prove theorem 3.3, we first present a variant of main results stated in Shen and Ghosal (2012) in the following two technical lemmas.

LEMMA B.8 *Let,*

$$\tilde{V}_{J_n,M_n} = \{\sigma = \tilde{\Phi}(W^{J,\theta}) : W^{J,\theta} = \theta^T\xi,\ \theta\in\mathbb{R}^j,\ j\le J_n,\ \|\theta\|_\infty \le M_n\}, \tag{B.14}$$

$$\tilde{\mathcal{W}}_{J_n,M_n} = \{(\sigma,\beta) : \sigma\in\tilde{V}_{J_n,M_n},\ \beta\in[\underline{\beta},\overline{\beta}]^d\}, \tag{B.15}$$

$$d_2(\eta,\eta_0) = \left\{\int_0^1 [\sigma(\boldsymbol{x})-\sigma_0(\boldsymbol{x})]^2\, dG_0(\boldsymbol{x})\right\}^{1/2} + \|\beta-\beta_0\|_2. \tag{B.16}$$

Assume that the conditions listed in Theorem 1 *of Shen and Ghosal (2012) hold relative to uniform metric* $\|\cdot\|_\infty$*, then for some positive constants* $\tilde{a}_1$, $\tilde{a}_2$, $\tilde{b}$*, we have the following,*

$$\log D(\varepsilon_n, \tilde{\mathcal{W}}_{J_n,M_n}, d_2) \le n\varepsilon_n^2, \tag{B.17}$$

$$\Pi(W \notin \tilde{\mathcal{W}}_{J_n,M_n}) \le \tilde{a}_1 \exp\{-\tilde{b}n\bar{\varepsilon}_n^2\}, \tag{B.18}$$

$$-\log \Pi\{\eta = (\sigma,\beta) : \|\sigma - \sigma_0\|_\infty^2 + \|\beta - \beta_0\|_2^2 \le \bar{\varepsilon}_n^2\} \le \tilde{a}_2 n\bar{\varepsilon}_n^2. \tag{B.19}$$

PROOF OF LEMMA B.8

We omit the proof of assertions (B.17) and (B.18) since it is similar to the corresponding parts in the proof of theorem 3.2. We are in a position to show (B.19). Observe that,

$$\begin{aligned}
&\Pi\{\eta = (\sigma,\beta) : \|\sigma - \sigma_0\|_\infty^2 + \|\beta - \beta_0\|_2^2 \le \bar{\varepsilon}_n^2\},\\
&\quad \ge \Pi_\sigma\left(\|\sigma - \sigma_0\|_\infty \le \frac{\bar{\varepsilon}_n}{2}\right) \times \Pi_\beta\left(\|\beta - \beta_0\|_2 \le \frac{\bar{\varepsilon}_n}{2}\right),\\
&\quad \ge \Pi_w\left(\|w - w_0\|_\infty \le \frac{\bar{\varepsilon}_n}{2}\right) \times \exp(-cd\log(1/\bar{\varepsilon}_n)),\\
&\quad \ge \exp\left\{-a_2 n\bar{\varepsilon}_n^2\right\} \times \exp(-\tilde{b}_2 n\bar{\varepsilon}_n^2),\\
&\quad \ge \exp(-\tilde{a}_2 n\bar{\varepsilon}_n^2),
\end{aligned}$$

where $\tilde{a}_2 = a_2 + \tilde{b}_2$. The assertion (B.19) follows by taking logarithm transformation on both sides above. We thus complete the proof of this lemma. □

LEMMA B.9 *Suppose that the conditions except (B.40) listed in Theorem 2 of Shen and Ghosal (2012) hold for the case* $r = \infty$*, then the posterior distribution of* η *converges at rate* ε_n *with respective to the Hellinger distance.*

PROOF OF LEMMA B.9

Notice that $K(p_{f_0}, p_f)$ and $V(p_{f_0}, p_f)$ exhibited in Theorem 2 of Shen and Ghosal (2012) are essentially the same as $K(\eta_0, \eta)$ and $V(\eta_0, \eta)$ respectively described in (3.1) and (3.2). We employ the similar arguments in the proof of Theorem 2 in Shen and Ghosal (2012) to show this lemma. It suffices to show that the following conditions stated in Theorem 4 of Ghosal and van der Vaart (2007a) hold.

$$\log D(\varepsilon_n, \tilde{\mathcal{W}}_{J_n,M_n}, d_H) \leq b_1 n\varepsilon_n^2, \tag{B.20}$$

$$\Pi(W \notin \tilde{\mathcal{W}}_{J_n,M_n}) \leq b_3 \exp\{-(b_2+4)n\bar{\varepsilon}_n^2\}, \tag{B.21}$$

$$\Pi(B^*(\eta_0, \bar{\varepsilon}_n)) \geq b_4 \exp\{-b_2 n\bar{\varepsilon}_n^2\}, \tag{B.22}$$

for some positive constants $b_1,\ b_2,\ b_3,\ b_4$, where $\tilde{\mathcal{W}}_{J_n,M_n}$ is described in lemma B.8 and $B^*(\eta_0, \bar{\varepsilon}_n) = \{\eta :\ K(\eta,\eta_0) < \bar{\varepsilon}_n^2,\ V(\eta,\eta_0) < \bar{\varepsilon}_n^2\}$. It is easy to show (B.20) and (B.21) hold by the same arguments used in the proof of Theorem 2 in Shen and Ghosal (2012). Now it remains to check (B.22). In fact, observe that by corollary B.4,

$$B^*(\eta_0, \bar{\varepsilon}_n) \supset \{\eta = (\sigma, \beta) :\ \|\sigma - \sigma_0\|_\infty^2 + \|\beta - \beta_0\|_2^2 \leq \bar{\varepsilon}_n^2\}.$$

It follows that by (B.19) in lemma B.8,

$$\begin{aligned}\Pi(B^*(\eta_0, \bar{\varepsilon}_n)) &\geq \Pi\{\eta = (\sigma, \beta) :\ \|\sigma - \sigma_0\|_\infty^2 + \|\beta - \beta_0\|_2^2 \leq \bar{\varepsilon}_n^2\} \\ &\geq \exp(-\tilde{a}_2 n\bar{\varepsilon}_n^2).\end{aligned}$$

Then the proof of this lemma is complete. □

B.2.2 Proof of Theorem 3.3

PROOF OF THEOREM 3.3

In order to obtain the rate ε_n like this, we only need to apply lemma B.9 with the appropriate choice of $\bar{J}_n,\ J_n,\ M_n,\ \bar{\varepsilon}_n$. It is easy to say that (B.38) and (B.39) described in Theorem 2 of Shen and Ghosal (2012) in terms of tensor-product spline basis. An application of corollary B.4 yields that,

$$\max(K(\eta_0,\eta),\ V(\eta_0,\eta)) \lesssim (\|\sigma - \sigma_0\|_\infty^2 + \|\beta - \beta_0\|_2^2).$$

Meanwhile, lemma 3.1 implies the approximation error $e(J) \approx J^{-\alpha}$. We proceed to determine the rate ε_n as follows. Firstly, it follows that $\bar{J}_n^{-\alpha} \leq \bar{\varepsilon}_n$ and $\bar{J}_n \log n \leq n\bar{\varepsilon}_n^2$ by (B.39). Hence we can choose $M_n = n^{1/t_3}$, $\bar{J}_n = (n/\log n)^{1/(2\alpha+d)}$ and $\bar{\varepsilon}_n = (n/\log n)^{-\alpha/(2\alpha+d)}$. Observe that $n\bar{\varepsilon}_n^2 \lesssim J_n \log^{t_1} n$ by (B.37), we can also choose $J_n = n^{1/(2\alpha+d)}(\log n)^{2\alpha/(2\alpha+d)-t_2}$. Noting also that $J_n \log n \lesssim n\varepsilon_n^2$ by (B.38), so that we get the rate ε_n as $n^{-\alpha/(2\alpha+d)}(\log n)^{\alpha/(2\alpha+d)-(t_2-1)/2}$. Then the proof of this theorem is complete. □

B.3 Auxiliary Theorems for Chapter 3

For easy reference, we collect some complementary results in the literature in aid of the proof of the theorems in this present article.

THEOREM B.10 (Ghosal and van de Vaart (2007)) *Let $P_\theta^{(n)}$ be product measures and d_n be defined as follows:*

$$d_n(\theta, \theta') = \frac{1}{n}\int (\sqrt{p_{\theta,i}} - \sqrt{p_{\theta',i}})^2 \, d\mu_i. \tag{B.23}$$

Suppose that for a sequence $\varepsilon_n \to 0$ such that $n\varepsilon_n^2$ is bounded away from zero, some $k > 1$, all sufficiently large j and sets $\Theta_n \subset \Theta$, the following conditions hold:

$$\sup_{\varepsilon > \varepsilon_n} \log N(\varepsilon/36, \{\theta \in \Theta_n : d_n(\theta, \theta_0) < \varepsilon\}, d_n) \leq n\varepsilon_n^2, \tag{B.24}$$

$$\frac{\Pi_n(\Theta \backslash \Theta_n)}{\Pi_n(B_n^*(\theta_0, \varepsilon_n; k))} = o(e^{-2n\varepsilon_n^2}), \tag{B.25}$$

$$\frac{\Pi_n(\theta \in \Theta_n : \; j\varepsilon_n \leq d_n(\theta, \theta_0) \leq 2j\varepsilon_n)}{\Pi_n(B_n^*(\theta_0, \varepsilon_n; k))} \leq e^{n\varepsilon_n^2 j^2/4}. \tag{B.26}$$

Then $P_\theta^{(n)}\Pi_n(\theta : d_n(\theta, \theta_0) \geq M_n\varepsilon_n | X^{(n)}) \to 0$ for every $M_n \to \infty$.

LEMMA B.11 (Shen and Ghosal (2012)) *For any $1 \leq p \leq \infty$, we have,*

$$\|\theta_1^T\xi - \theta_2^T\xi\|_r \leq \sum_{j=1}^{J} |\theta_{1j} - \theta_{2j}| \max_{1 \leq j \leq J} \|\xi_j\|_p \leq \sqrt{J}\|\theta_1 - \theta_2\|_2 \, C_{p,J}, \tag{B.27}$$

where,

$$C_{p,J} \equiv \max_{1\leq j\leq J} \|\xi_j\|_p \asymp \begin{cases} 1 & p=2 \\ \sqrt{J} & p=\infty \end{cases}$$

THEOREM B.12 (Shen and Ghosal (2012)) *Let $\varepsilon_n \geq \bar{\varepsilon}_n$ be two sequence of positive numbers satisfying $\varepsilon_n \to 0$ and $n\bar{\varepsilon}_n^2 \to \infty$ as $n \to \infty$. For a function w_0, suppose that there exist sequences of positive numbers J_n, $\bar{J}_n$ and M_n, a strictly decreasing, nonnegative function $e(\cdot)$ and a $\theta_{0,j} \in \mathbb{R}^j$ for any $j \in \mathbb{N}$, such that the following conditions hold for some positive constants a_1, a_1', a_2:*

$$\|\theta_{0,j}\| \leq H,\ d_2(w_0, \theta_{0,j}^T\xi) \leq e(j), \tag{B.28}$$

$$J_n\{\log J_n + \log a(J_n) + \log M_n + \log(1/\varepsilon_n)\} \leq n\varepsilon_n^2, \tag{B.29}$$

$$e(\bar{J}_n) \leq \bar{\varepsilon}_n,\ \log\{1/B(\bar{J}_n)\} + c_2\bar{J}_n \log(2b(\bar{J}_n)/\bar{\varepsilon}_n) \leq a_2 n\bar{\varepsilon}_n^2, \tag{B.30}$$

$$A(J_n) \leq a_1 \exp\{-(a_2+4)n\bar{\varepsilon}_n^2\},\ J_n \exp\{-CM_n^{t_3}\} \leq a_1' \exp\{-(a_2+4)n\bar{\varepsilon}_n^2\}. \tag{B.31}$$

Let $\mathcal{W} = \{w = \theta^T\xi : \theta \in \mathbb{R}^j, j \leq J_n, \|\theta\|_\infty \leq M_n\}$. Then the following assertions hold:

$$\log D(\varepsilon_n, \mathcal{W}_{J_n,M_n}, d_2) \leq n\varepsilon_n^2, \tag{B.32}$$

$$\Pi(W \notin \mathcal{W}_{J_n,M_n}) \leq (a_1 + a_1')\exp\{-(a_2+4)n\bar{\varepsilon}_n^2\}, \tag{B.33}$$

$$-\log \Pi\{w = \theta^T\xi : d_2(w_0, w) \leq \bar{\varepsilon}_n\} \leq a_2 n\bar{\varepsilon}_n^2. \tag{B.34}$$

THEOREM B.13 (Shen and Ghosal (2012)) *Suppose that we have independent observations X_i following some distributions with densities $p_{i,w} : i = 1, \dots, n$ respectively. Let $w_0 \in C^\alpha(\Omega_0)$ be the true value of w. let r be either 2 or ∞. Let $\varepsilon_n \geq \bar{\varepsilon}_n$ be two sequences of positive numbers satisfying $\varepsilon_n \to 0$ and $n\bar{\varepsilon}_n^2 \to \infty$ as $n \to \infty$. Assume that there exist a $\theta_0 \in \mathbb{R}^J$, $\|\theta_0\| \leq H$ and some positive constants C_1, C_2 satisfying,*

$$\|w_0 - \theta_0^T\xi\|_r \leq C_1 J^\alpha (\log J)^s,\ s \geq 0, \tag{B.35}$$

$$\|\theta_1^T\xi - \theta_2^T\xi\|_r \leq C_2 J^{K_0}\|\theta_1 - \theta_2\|_2,\ K_0 \geq 0,\ \textit{for any}\ \theta_1, \theta_2 \in \mathbb{R}^J. \tag{B.36}$$

Assume that the prior on J and θ satisfy some conditions $(A2)$ and $(A3)$ in their paper. Let J_n, $\bar{J}_n \geq 2$ and M_n be sequences of positive numbers such that the following hold:

$$J_n \log^{t_1} J_n \geq 6n\bar{\varepsilon}_n^2,\ \log J_n + 6n\varepsilon_n^2 \leq c_1 M_n^{t_3}, \tag{B.37}$$

$$J_n\{(K_0+1)\log J_n + \log M_n + \log(1/\varepsilon_n) + \log n\} \leq n\varepsilon_n^2, \tag{B.38}$$

$$\bar{J}_n^{-\alpha}(\log \bar{J}_n)^s \leq \bar{\varepsilon}_n,\ \bar{J}_n\{\log^{t_2} \bar{J}_n + c_2 K_0 \log(\bar{J}_n) + c_2 \log(1/\bar{\varepsilon}_n)\} \leq 2n\bar{\varepsilon}_n^2, \tag{B.39}$$

$$\rho_n(w_1, w_2) \lesssim n^{C_3}\|w_1 - w_2\|_r \textit{ for any } w_1,\ w_2 \in \mathcal{W}_{J_n,M_n} \textit{ and some constant } C_3 > 0, \tag{B.40}$$

$$\max_{1\leq i\leq n}\{\,K(p_{i,w_0}, p_{i,w}),\ V(p_{i,w_0}, p_{i,w})\,\} \lesssim \|w_1 - w_2\|_r, \tag{B.41}$$

provided $\|w_1 - w_2\|_r$ is sufficiently small. Then the posterior of w converges around w_0 at the rate ε_n with respect to ρ_n.

THEOREM B.14 (de Jonge and van Zanten (2012)) *Suppose that for every $m \geq 1$,*

$$C_1 \exp(-D_1 m^d \log^t m) \leq P(M = m) \leq C_2 \exp(-D_2 m^d \log^t m), \tag{B.42}$$

for some constants $C_1, C_2, D_1, D_2, t \geq 0$. If $w_0 \in C^r([0,1]^d)$ for some integer $r \leq q$, then there exist a constant $C > 0$, a constant $D > 0$ and measurable subsets U_n of $C([0,1]^d)$ such that,

$$P\left(\|W - w_0\|_\infty \leq 2\varepsilon_n\right) \geq \exp\left(-n\varepsilon_n^2\right), \tag{B.43}$$

$$P(W \notin U_n) \leq \exp\left(-Cn\varepsilon_n^2\right), \tag{B.44}$$

$$\log N(2\bar{\varepsilon}_n, U_n, \|\cdot\|_\infty) \leq Dn\bar{\varepsilon}_n^2, \tag{B.45}$$

are satisfied for sufficiently large n, and for ε_n and $\bar{\varepsilon}_n$ given by,

$$\varepsilon_n = c(n/\log^{1\vee t} n)^{-\frac{r}{d+2r}} \qquad \bar{\varepsilon}_n = n^{-\frac{r}{d+2r}}(\log n)^{\frac{(1\vee t)r}{d+2r}+(\frac{1-t}{2})+}, \tag{B.46}$$

for $c > 0$ a large enough constant.

Appendix C

Appendix to Chapter 4

C.1 Useful complementary Lemmas for Chapter 4

In this Appendix C.1 we will state some auxiliary lemmas to complement the main results in the context.

Let $P, Q \in \mathscr{P}$ and consider a prior Π and a measurable set $B \in \mathscr{B}$ such that $\Pi(B) > 0$, recall that the *prior predictive distribution* P_n^Π and the *local prior predictive distribution* Q_n^Π are two distributions on n-sample space $(\mathcal{X}^n, \mathscr{X}^n)$ respectively, defined by,

$$P_n^\Pi(A) := \int_{\mathscr{P}} P^n(A)\, d\Pi(P), \tag{C.1}$$

$$Q_n^\Pi(A) := \int_{\mathscr{P}} Q^n(A)\, d\Pi(Q|B), \tag{C.2}$$

for all $n \geq 1$ and $A \in \sigma(X_1, \ldots, X_n)$. The next lemma describes the relationship between these two distributions.

LEMMA C.1 *For each $A \in \sigma(X_1, \ldots, X_n)$ and each $n \geq 1$, it holds,*

$$P_n^\Pi(A) \geq \Pi(B)\, Q_n^\Pi(A). \tag{C.3}$$

That is, the prior predictive distribution is greater than the product of the local prior predictive distribution and prior mass on a given measurable set in the model.

PROOF OF LEMMA C.1

Observe that for all $n \geq 1$ and $A \in \sigma(X_1, \ldots, X_n)$,

$$Q_n^{\Pi}(A) = \int_{\mathscr{P}} Q^n(A)\, d\Pi(Q|B) = \int \frac{Q^n(A)}{\Pi(B)}\, d\Pi(Q \cap B).$$

Define $\nu_B(B') := \Pi(B' \cap B)$ for any $B' \in \mathscr{B}$, then by Theorem 16.11 and Example 16.9 in Billingsley (1995), we arrive at,

$$\nu_B \ll \Pi, \qquad \frac{d\nu_B}{d\Pi} = 1_B, \qquad \int_D f\, d\nu_B = \int_D f 1_B\, d\Pi = \int_{D\cap B} f\, d\Pi,$$

for any nonnegative measurable function f on $(\mathscr{P}, \mathscr{B})$ and all $D \in \mathscr{B}$.

Hence it follows that,

$$Q_n^{\Pi}(A) = \int_{\mathscr{P}} \frac{Q^n(A)}{\Pi(B)}\, d\nu_B(Q) = \int_{\mathscr{P}} \frac{Q^n(A)}{\Pi(B)} 1_B\, d\Pi(Q).$$

Now note that $\mathscr{P} \cap B \subset \mathscr{P}$, we conclude that,

$$Q_n^{\Pi}(A) \leq \frac{1}{\Pi(B)} \int_{\mathscr{P}} Q^n(A)\, d\Pi(Q) = \frac{P_n^{\Pi}(A)}{\Pi(B)}.$$

□

The following lemma deals directly with the comparison between a collection of n-fold probability measures in some measurable set with the Cartesian product of these n-copies of this same set.

LEMMA C.2 *Let* $V^n := \{P^n :\ P \in V\} = \{P \times P \times \ldots \times P :\ P \in V\}$, *then we have,*

$$V^n \subset V \times V \times \ldots \times V, \tag{C.4}$$

$$\mathrm{co}(V^n) \subset \mathrm{co}(V \times V \times \ldots \times V). \tag{C.5}$$

PROOF OF LEMMA C.2

It is trivial that (C.4) holds. It remains to show (C.5). According to the definition of the convex hull, $\mathrm{co}(V^n)$ is the smallest convex set that contain V^n. Since $\mathrm{co}(V \times V \times \ldots \times V)$ is a convex set that contains $V \times V \times \ldots \times V$, then we infer that (C.5) holds by (C.4). □

For sake of clarity, we denote the element in $\mathrm{co}(V^n)$ by P_n and P^n is the generic element included in V^n.

LEMMA C.3 *Let $Q^n \ll P_0^n$ for any $Q \in \mathscr{P}$ and $\alpha \in [0,1]$, then,*

$$\frac{dQ_n^\Pi}{dP_n} = \int_{\mathscr{P}} \frac{dQ^n}{dP_n}\, d\Pi(Q|B), \tag{C.6}$$

$$\left(\frac{dQ_n^\Pi}{dP_n}\right)^{-\alpha} \leq \int_{\mathscr{P}} \left(\frac{dQ^n}{dP_n}\right)^{-\alpha} d\Pi(Q|B). \tag{C.7}$$

PROOF OF LEMMA C.3

Since $Q^n \ll P_0^n$, the Radon-Nikodym theorem tells us that there is a Radon-Nikodym derivative $\frac{dQ^n}{dP_0^n}$ for which,

$$Q_n^\Pi(A) = \int_A \frac{dQ^n}{dP_0^n}\, dP_0^n,$$

holds for all $A \in \sigma(X_1, \ldots, X_n)$. Note that in view of the Fubini's theorem, for any $A \in \sigma(X_1, \ldots, X_n)$, one could find,

$$\begin{aligned} Q_n^\Pi(A) &= \int_{\mathscr{P}} Q^n(A)\, d\Pi(Q|B) \\ &= \int_{\mathscr{P}} \int_A \frac{dQ^n}{dP_0^n}\, dP_0^n\, d\Pi(Q|B) \\ &= \int_A \left[\int_{\mathscr{P}} \frac{dQ^n}{dP_0^n}\, d\Pi(Q|B)\right] dP_0^n. \end{aligned}$$

Then an application of the Radon-Nikodym theorem again gives,

$$\frac{dQ_n^\Pi}{dP_0^n} = \int_{\mathscr{P}} \frac{dQ^n}{dP_0^n}\, d\Pi(Q|B).$$

Note also that $P_n \ll P_0^n$ and a straightforward computation shows that,

$$\begin{aligned}\frac{dQ_n^\Pi}{dP_n} &= \frac{\int_{\mathscr{P}} dQ^n/dP_0^n \, d\Pi(Q|B)}{dP_n/dP_0^n} \\ &= \int_{\mathscr{P}} \frac{dQ^n/dP_0^n}{dP_n/dP_0^n} \, d\Pi(Q|B) \\ &= \int_{\mathscr{P}} \frac{dQ^n}{dP_n} \, d\Pi(Q|B).\end{aligned}$$

Hence, (C.6) is as desired.

Now we turn to show(C.7). Define $f : (0,\infty) \to \mathbb{R}$ by $f(x) = x^{-\alpha}$. Notice that f is convex on $(0,\infty)$ and applying Jensen's inequality, one could get,

$$f\left(\frac{dQ_n^\Pi}{dP_n}\right) = f\left(\int_{\mathscr{P}} \frac{dQ^n}{dP_n} \, d\Pi(Q|B)\right) \le \int_{\mathscr{P}} f\left(\frac{dQ^n}{dP_n}\right) d\Pi(Q|B).$$

Then (C.7) follows by substituting f with its associated explicit function form. Therefore the proof of this lemma is complete. □

The following corollary is an immediate consequence of lemma C.3, which provides us a further characterization of the Hellinger transform conditioning on some measurable set with positive prior mass.

COROLLARY C.4

$$P_0^n\Big(\frac{dP_n}{dP_n^\Pi}\Big)^\alpha \le \Pi(B)^{-\alpha}\, P_0^n\Big(\frac{dP_n}{dQ_n^\Pi}\Big)^\alpha \le \Pi(B)^{-\alpha}\, P_0^n \int_{\mathscr{P}} \Big(\frac{dP_n}{dQ^n}\Big)^\alpha d\Pi(Q|B). \tag{C.8}$$

The following lemma states that Hellinger transform taken on covex hulls of V^n admits a factorization on convex hull of V.

LEMMA C.5

$$\sup_{P_n \in \mathrm{co}(V^n)} P_0^n\Big(\frac{dP_n}{dQ^n}\Big)^\alpha \le \Big[\sup_{P \in \mathrm{co}(V)} P_0\Big(\frac{dP}{dQ}\Big)^\alpha\Big]^n. \tag{C.9}$$

PROOF OF LEMMA C.5

First of all, set,

$$d\,\nu_{P,Q,i} := \frac{dP_0}{dQ}\,dP_i, \quad P_i \in V, \quad i = 1, 2, \ldots, J,$$
$$\mathscr{Q} := \left\{ \nu_{P,Q} := \int \frac{dP_0}{dQ}\,dP : P \in V \right\}.$$

Suppose that $P_n \in \operatorname{co}(V^n)$ be given, then there exist $\{P_1^n, P_2^n, \ldots, P_J^n\} \subset V^n$ and $\lambda_1, \lambda_2, \ldots, \lambda_J \geq 0$, $\sum_{i=1}^{J} \lambda_i = 1$ such that,

$$P_n = \sum_{i=1}^{J} \lambda_i P_i^n.$$

Next, observe that the left side of (C.9) could be written as,

$$\sup \left\{ \int \Big(\sum_{i=1}^{J} \lambda_i \frac{dP_i^n}{dQ^n} \Big)^{\alpha} dP_0^n : P_i^n \in V^n, i = 1, 2, \ldots, J \right\},$$

which is weakly less than,

$$\sup \left\{ \int \sum_{i=1}^{J} \lambda_i \Big(\frac{dP_i^n}{dQ^n} \Big)^{\alpha} dP_0^n : P_i^n \in V^n, i = 1, 2, \ldots, J \right\} \leq \sum_{i=1}^{J} \lambda_i \sup_{P_i^n \in V^n} \left\{ \int \Big(\frac{dP_i^n}{dQ^n} \Big)^{\alpha} dP_0^n \right\}.$$

And the right side of the inequality above is equal to,

$$\sum_{i=1}^{J} \lambda_i \sup_{P_i \in V} \left\{ \int \Big[\Big(\frac{dP_0}{dQ} dP_i \Big)^{n} \Big]^{\alpha} \big(dP_0^n\big)^{1-\alpha} \right\} = \sum_{i=1}^{J} \lambda_i \sup_{\nu_{P,Q,i} \in \mathscr{Q}} \left\{ \int \Big(d\,\nu_{P,Q,i}^n \Big)^{\alpha} \big(dP_0^n\big)^{1-\alpha} \right\}.$$

Moreover, the right side of the equality above is less than,

$$\begin{aligned}
&\sum_{i=1}^{J} \lambda_i \sup \left\{ \int \Big(d\,\nu_{P,Q,i}^n \Big)^{\alpha} \big(dP_0^n\big)^{1-\alpha} : \nu_{P,Q,i}^n \in \operatorname{co}(\mathscr{Q} \times \mathscr{Q} \ldots \times \mathscr{Q}) \right\} \\
&\leq \sum_{i=1}^{J} \lambda_i \, \rho_\alpha\big(\mathscr{Q} \times \mathscr{Q} \ldots \times \mathscr{Q}, P_0^n\big) \\
&\leq \sum_{i=1}^{J} \lambda_i \left[\rho_\alpha(\mathscr{Q}, P_0) \right]^n = \left[\rho_\alpha(\mathscr{Q}, P_0) \right]^n,
\end{aligned}$$

where the second inequality follows from Lemma 6.2 in Kleijn and van der Vaart (2006) and,

$$\rho_\alpha(\mathscr{Q}, P_0) := \sup_{\nu_{P,Q}\in \mathrm{co}(\mathscr{Q})} P_0 \left(\frac{d\,\nu_{P,Q}}{dQ}\right)^\alpha .$$

Finally, simple algebra shows that,

$$\rho_\alpha(\mathscr{Q}, P_0) = \sup_{P\in \mathrm{co}(V)} P_0 \left(\frac{dP}{dQ}\right)^\alpha .$$

This concludes the proof of this lemma. □

If we replace V with a convex set $V_{n,m}$, then we easily draw an analogous claim of lemma C.5 by taking, $\mathrm{co}(V_{n,m}) = V_{n,m}$.

COROLLARY C.6

$$\sup_{P_n\in \mathrm{co}(V^n_{n,m})} P_0^n \left(\frac{dP_n}{dQ^n}\right)^\alpha \leq \Big[\sup_{P\in \mathrm{co}(V_{n,m})} P_0 \Big(\frac{dP}{dQ}\Big)^\alpha\Big]^n = \Big[\sup_{P\in V_{n,m}} P_0 \Big(\frac{dP}{dQ}\Big)^\alpha\Big]^n . \tag{C.10}$$

C.2 Proofs of Theorems

C.2.1 Proofs for Section 2

PROOF OF THEOREM 4.2

For each $n \geq 1$, let us first introduce a new test function as follows,

$$\phi_n := \max_{1\leq m\leq N_n} \phi_{n,m},$$

Then according to assumption (4.9), it follows that,

$$P_0^n \phi_n \leq \sum_{m=1}^{N_n} P_0^n \phi_{n,m} \leq e^{(L-KM^2)n\varepsilon_n^2},$$

$$\sup_{P\in V_n} P_0^n \frac{dP^n}{dP_n^\Pi}(1-\phi_n) \leq \min_{1\leq m\leq N_n} \sup_{P\in V_{n,m}} P_0^n \frac{dP^n}{dP_n^\Pi}(1-\phi_{n,m}) \leq e^{-KM^2 n\varepsilon_n^2}.$$

Next, observe that,

$$P_0^n \Pi(V_n|X_1,\dots,X_n) \leq P_0^n \Pi(V_n|X_1,\dots,X_n)(1-\phi_n) + P_0^n \phi_n,$$

In addition, an application of the Fubini's theorem yields,

$$\begin{aligned} P_0^n \Pi(V_n|X_1,\dots,X_n)(1-\phi_n) &= P_0^n \int_{V_n} \frac{dP^n}{dP_n^\Pi}(1-\phi_n)\, d\Pi(P) \\ &= \int_{V_n} P_0^n \frac{dP^n}{dP_n^\Pi}(1-\phi_n)\, d\Pi(P) \\ &\leq \sup_{P\in V_n} P_0^n \frac{dP^n}{dP_n^\Pi}(1-\phi_n) \\ &\leq e^{-KM^2 n\varepsilon_n^2}. \end{aligned}$$

Hence, for a sufficiently large $M > 0$, one could find,

$$P_0^n \Pi(V_n|X_1,\dots,X_n) \leq e^{(L-KM^2)n\varepsilon_n^2} + e^{-KM^2 n\varepsilon_n^2} \longrightarrow 0 \text{ as } n \to \infty.$$

We have thus proved this theorem. □

PROOF OF THEOREM 4.3

This lemma can be completed by the method analogous to that used in Lemma 2.2 in Kleijn (2015) just replacing ϕ_n, V there with $\phi_{n,m}, V_{n,m}$ respectively. □

PROOF OF LEMMA 4.4

Suppose $\alpha \in [0,1]$, observe that by lemma C.1,

$$P_n^\Pi(A) \geq \Pi(B_n)\, Q_n^\Pi(A),$$

for each $n \geq 1$ and all $A \in \sigma(X_1,\dots,X_n)$. Using the fact that the function $x \mapsto x^{-\alpha}$ is convex on

$(0, \infty)$, we get that,

$$P_0^n \Big(\frac{dP_n}{dP_n^\Pi} \Big)^\alpha \leq \Pi(B_n)^{-\alpha} \, P_0^n \Big(\frac{dP_n}{dQ_n^\Pi} \Big)^\alpha \leq \Pi(B_n)^{-\alpha} \, P_0^n \int \Big(\frac{dP_n}{dQ^n} \Big)^\alpha \, d\Pi(Q|B_n). \tag{C.11}$$

Applying Fubini's theorem together with the factorization of the Hellinger transform over the convex hulls of products by Lemma 6.2 in Kleijn and van der Vaart (2006) gives that,

$$\begin{aligned} \sup_{P_n \in \mathrm{co}(V_{n,m}^n)} \inf_{0 \leq \alpha \leq 1} \Pi(B_n)^{-\alpha} \int P_0^n \Big(\frac{dP_n}{dQ^n} \Big)^\alpha d\Pi(Q|B_n) \\ \leq \inf_{0 \leq \alpha \leq 1} \Pi(B_n)^{-\alpha} \int \sup_{P_n \in \mathrm{co}(V_{n,m}^n)} P_0^n \Big(\frac{dP_n}{dQ^n} \Big)^\alpha d\Pi(Q|B_n) \\ \leq \inf_{0 \leq \alpha \leq 1} \Pi(B_n)^{-\alpha} \int \Big[\sup_{P \in \mathrm{co}(V_{n,m})} P_0 \Big(\frac{dP}{dQ} \Big)^\alpha \Big]^n d\Pi(Q|B_n). \end{aligned}$$

Here the last inequality follows lemma C.5. Combining (4.14) and (C.11) with $\alpha = 1$ shows that $P_0^n(dP^n/dP_n^\Pi) < \infty$ for all $P \in V_{n,m}$, $1 \leq m \leq N_n$. According to (4.12), we have thus proved the lemma. □

C.2.2 Proofs for Section 3

PROOF OF LEMMA 4.5

We firstly assume (4.16) holds. A Taylor expansion of $P_0\big(\frac{dP}{dQ}\big)^\alpha$ at $\alpha = 0$ yields that for all $\alpha \in (0, 1)$,

$$\sup_{Q \in B} \sup_{P \in V} P_0 \Big(\frac{dP}{dQ} \Big)^\alpha \leq 1 + \alpha \sup_{\alpha' \in (0,\alpha]} \sup_{Q \in B} \sup_{P \in V} P_0 \Big(\frac{dP}{dQ} \Big)^{\alpha'} \log \frac{dP}{dQ}. \tag{C.12}$$

Define,

$$z(\alpha') = \sup_{Q \in B} \sup_{P \in V} P_0 \Big(\frac{dP}{dQ} \Big)^{\alpha'} \log \frac{dP}{dQ}. \tag{C.13}$$

It is easy to see that $z(\alpha')$ is convex on $[0, 1]$ which hence implies that it is also continuous on $(0, 1)$ and upper-semicontinuous at 0. Note that based on (4.16) and Lemma A.1 in the Appendix of

Kleijn (2015), we see that,

$$\begin{aligned}\lim_{\alpha'\to 0} z(\alpha') &= \sup_{Q\in B}\sup_{P\in V} P_0\, 1_{\{p>0\}} \log\frac{dP}{dQ} \leq \sup_{Q\in B}\sup_{P\in V} P_0 \log\frac{dP}{dQ}\\ &= \sup_{Q\in B} -P_0\log\frac{dQ}{dP_0} - \inf_{P\in V}\Big\{-P_0\log\frac{dP}{dP_0}\Big\} < -\tilde{M}\varepsilon^2.\end{aligned}$$

Let $\tilde{\varepsilon} = \frac{1}{2}(-\tilde{M}\varepsilon^2 - \lim_{\alpha'\to 0} z(\alpha'))$. As $z(\alpha')$ is upper-semicontinuous at 0, for $\tilde{\varepsilon} > 0$, there exists $0 < \alpha_0 < 1$ such that,

$$z(\alpha') \leq \lim_{\alpha\to 0+} z(\alpha) + \tilde{\varepsilon} = \frac{1}{2}(-\tilde{M}\varepsilon^2 + \lim_{\alpha'\to 0} z(\alpha')) < -\tilde{M}\varepsilon^2, \tag{C.14}$$

for all $\alpha' \in (0, \alpha_0]$. Additionally, noting also that by (C.12) and (C.14),

$$\begin{aligned}\inf_{0\leq\alpha\leq 1}\sup_{Q\in B}\sup_{P\in V} P_0\Big(\frac{dP}{dQ}\Big)^{\alpha} &\leq 1 + \inf_{0<\alpha\leq 1}\sup_{\alpha'\in(0,\alpha]} z(\alpha') \leq 1 + \inf_{0<\alpha\leq\alpha_0}\sup_{\alpha'\in(0,\alpha]} z(\alpha')\\ &\leq 1 + \inf_{0<\alpha\leq\alpha_0}(-\tilde{M}\varepsilon^2) = 1 - \tilde{M}\varepsilon^2 \leq e^{-\tilde{M}\varepsilon^2}.\end{aligned}$$

Hence (4.17) follows. Conversely, we prove the other side by contradiction. Assume that the assertion (4.16) does not hold, then applying Jensen's inequality for the exponential function e^x yields,

$$\begin{aligned}\sup_{Q\in B}\sup_{P\in V} P_0\Big(\frac{dP}{dQ}\Big)^{\alpha} &= \sup_{Q\in B}\sup_{P\in V} P_0\Big(e^{\alpha\log\frac{dP}{dQ}}\Big) \geq \sup_{Q\in B}\sup_{P\in V} e^{\alpha\, P_0\log\frac{dP}{dQ}}\\ &= \exp\left\{\alpha \sup_{Q\in B}\sup_{P\in V} P_0\log\frac{dP}{dQ}\right\}\\ &= \exp\left\{\alpha\sup_{Q\in B} -P_0\log\frac{dQ}{dP_0} - \alpha\inf_{P\in V}\Big\{-P_0\log\frac{dP}{dP_0}\Big\}\right\} \geq e^{-\alpha\tilde{M}\varepsilon^2}.\end{aligned}$$

Taking infinitum with $0 \leq \alpha \leq 1$ on both sides above, one could obtain,

$$\inf_{0\leq\alpha\leq 1}\sup_{Q\in B}\sup_{P\in V} P_0\Big(\log\frac{dP}{dQ}\Big) \geq e^{-\tilde{M}\varepsilon^2},$$

which is contrary to (4.17). Therefore this completes the proof of this lemma. □

Proof of theorem 4.6

First of all, let us define the Kullback-Leibler B_n as follows:

$$B_n := \left\{ P \in \mathscr{P} \,:\, -P_0 \log \frac{dP}{dP_0} < \varepsilon_n^2 \right\}.$$

Note that V_n is covered by some finite collection of Hellinger balls $\{V_{n,i}\}_{i=1}^{N_n}$ of radii $\frac{M\varepsilon_n}{2}$. Thus for every $P \in V_{n,i}$, $1 \le i \le N_n$, we obtain,

$$-P_0 \log \frac{dP}{dP_0} \ge d^2(P, P_0) > \frac{M^2 \varepsilon_n^2}{4}.$$

Consequently,

$$\sup_{Q \in B_n} -P_0 \log \frac{dQ}{dP_0} - \inf_{P \in V_{n,m}} \left\{ -P_0 \log \frac{dP}{dP_0} \right\} < (-M^2/4 + 1)\varepsilon_n^2,$$

for all $1 \le m \le N_n$, hence (4.16) is fulfilled by taking $\tilde{M} = M^2/4 - 1$ for $M > 2\sqrt{2}$. Next, notice also that Hellinger ball is convex and in order to get the exponential bound for the type I and type II error probability, it is necessary to just upper bound the quantity below for each $1 \le m \le N_n$,

$$\inf_{0 \le \alpha \le 1} \sup_{Q \in B_n} \sup_{P \in V_{n,m}} \Pi(B_n)^{-\alpha/n} P_0 \left(\frac{dP}{dQ} \right)^{\alpha}.$$

Thus if the GGV priors assumption (4.1) holds, then,

$$\Pi(B_n) \ge \Pi\Big(P \in \mathscr{P} : -P_0 \log \frac{dP}{dP_0} < \epsilon_n^2,\, P_0 \Big(\log \frac{dP}{dP_0} \Big)^2 < \epsilon_n^2 \Big) \ge e^{-n\epsilon_n^2},$$

so that $\Pi(B_n)^{-\alpha/n} < e^{\alpha\epsilon_n^2} \leq e^{\epsilon_n^2}$ since $\alpha \in [0,1]$. What's more, lemma 4.5 guarantees that,

$$\begin{aligned}&\inf_{0\leq\alpha\leq 1} \sup_{Q\in B_n} \sup_{P\in V_{n,m}} \Pi(B_n)^{-\alpha/n} P_0 \left(\frac{dP}{dQ}\right)^{\alpha} \\ &\leq e^{\epsilon_n^2} \inf_{0\leq\alpha\leq 1} \sup_{Q\in B_n} \sup_{P\in V_{n,m}} P_0 \left(\frac{dP}{dQ}\right)^{\alpha} \\ &\leq e^{(-M^2/4+2)\varepsilon_n^2}.\end{aligned}$$

An application of lemma 4.4 with $K = \frac{M^2-8}{4M^2}$ implies that (4.9) in theorem 4.2 holds. Thus we have completed the proof of this theorem. □

PROOF OF THEOREM 4.7

Let us first consider the following two sets,

$$\begin{aligned}V_n :=& \{P \in \mathscr{P} : \, d_H(P, P_0) > M\varepsilon_n\}, \\ B_n :=& \{Q \in \mathscr{P} : \, d_H(Q, P_0) < \frac{\varepsilon_n^2}{2L'} \wedge \varepsilon_n'\}.\end{aligned}$$

Note that $V_n \subset \mathscr{P}$ and $N(\varepsilon_n, \mathscr{P}, d_H) \leq e^{\tilde{C}n\varepsilon_n^2}$, then there exist a finite collection of probability measures $\{P_1, P_2, \ldots, P_{N_n}\} \subset V_n$ such that,

$$V_n \subset \bigcup_{m=1}^{N_n} V_{n,m},$$

where,

$$\begin{aligned}&N_n \leq e^{\tilde{C}n\varepsilon_n^2}, \\ &V_{n,m} := \{P \in \mathscr{P} : \, d_H(P, P_m) < \varepsilon_n\}, \; m = 1, 2, \ldots, N_n.\end{aligned}$$

By the triangle inequality, one could see that for every $P \in V_{n,m}$, $m = 1, 2, \ldots, N_n$,

$$d_H(P, P_0) \geq d_H(P_0, P_m) - d_H(P, P_m) > (M-1)\varepsilon_n.$$

Moreover, observe that for $Q \in B_n$, $P \in V_{n,m}$,

$$\begin{aligned}
P_0\left(\frac{p}{q}\right)^{1/2} &= \int p^{1/2}p_0^{1/2}\,d\mu + \int\left(\sqrt{\frac{p_0}{q}}-1\right)\left(\frac{p}{q}\right)^{1/2}\left(\frac{p_0}{q}\right)^{1/2} dQ \\
&\leq 1-\frac{1}{2}d_H^2(P_0,P)+d_H(P_0,Q)\left\|\frac{p_0}{q}\right\|_{2,Q}\left\|\frac{p}{q}\right\|_{2,Q} \\
&\leq 1-\frac{1}{2}\varepsilon_n^2(M-1)^2+L'^2\left(\frac{\varepsilon_n^2}{2L'}\wedge\varepsilon_n'\right) \\
&\leq \exp\left\{-\frac{1}{2}\varepsilon_n^2(M^2-2M+1-L'/2)\right\} \\
&\leq \exp\left\{-\frac{1}{2}\varepsilon_n^2(M+1)\right\}.
\end{aligned}$$

where M is chosen sufficiently large and the first inequality above follows from the Cauchy-Schwarz inequality. Using the convex property of the Hellinger ball, one could obtain for each $1 \leq m \leq N_n$ and a sufficiently large $M > 0$,

$$\begin{aligned}
\Pi(B_n)^{-1/2n}\sup_{Q\in B_n}\sup_{P\in \text{co}(V_n)} P_0\left(\frac{dP}{dQ}\right)^{1/2} &\leq \exp\left\{\frac{C'}{2}\left(\frac{\varepsilon_n^2}{2L'}\wedge\varepsilon_n'\right)^2\right\}\exp\left\{-\frac{1}{2}\varepsilon_n^2(M+1)\right\} \\
&\leq \exp\left\{-\frac{1}{2}\varepsilon_n^2 M-\frac{1}{2}\varepsilon_n^2+\frac{C'\varepsilon_n^4}{8L'2}\right\} \\
&\leq \exp\left\{-\frac{1}{2}M\varepsilon_n^2\right\}.
\end{aligned}$$

The last inequality follows from the fact that $-\frac{1}{2}\varepsilon_n^2+\frac{C'\varepsilon_n^4}{8L'2}<0$ when ε_n is small enough as $n \to 0$. Notice that by the Cauchy-Schwarz inequality again, for each $Q \in B_n$, $P \in V_n$,

$$P_0\left(\frac{dP}{dQ}\right) \leq \left\|\frac{p_0}{q}\right\|_{2,Q}\left\|\frac{p}{q}\right\|_{2,Q} \leq L'^2 < \infty.$$

Therefore, we conclude that the claim (4.22) is as desired according to theorem 4.1 and lemma 4.4. □

PROOF OF THEOREM 4.8

This proof can be proved in a similar way as shown in Theorem 3.1 in Ghosal et al. (2000). Write

$\| \cdot \|_{2,\mu}$ for the norm in $L^2(\mu)$. According to theorem 4.7, it suffices to show the following prior condition holds.

$$\Pi(B_n) = \Pi\{Q \in \mathscr{P} : d_H(Q, P_0) < \varepsilon_n\} \geq e^{-C' n\varepsilon_n^2}.$$

where C' is a positive constant. Observe that given $\varepsilon_n > 0$, the bracketing number of $\mathscr{P}$ is finite, then there exists a upper bracket $u_{n'}$ in $\mathscr{P}_n$ such that $p_0 < u_{n'}$ and $d_H(p_0, u_{n'}) < \varepsilon_n$. Then taking the squared root of the integral $\int u_{n'} d\,\mu$, one could get,

$$1 = \left(\int p_0 \, d\mu\right)^{1/2} \leq \left(\int u_{n'} \, d\mu\right)^{1/2} = \|\sqrt{u_{n'}}\|_{2,\mu} \leq d_H(p_0, u_{n'}) + \|\sqrt{p_0}\|_{2,\mu} \leq \varepsilon_n + 1.$$

Let $\varepsilon_n < 1$, note also that by the triangle inequality,

$$\begin{aligned} d_H\left(p_0, u_{n'}/\int u_{n'} \, d\mu\right) &\leq d_H(p_0, u_{n'}) + d_H\left(u_{n'}, u_{n'}/\int u_{n'} \, d\mu\right) \\ &\leq \varepsilon_n + (\|\sqrt{u_{n'}}\|_{2,\mu} - 1)^2 \\ &\leq \varepsilon_n + \varepsilon_n^2 < 2\varepsilon_n. \end{aligned}$$

So that the Hellinger ball $\{Q \in \mathscr{P} : d_H(Q, P_0) < 2\varepsilon_n\}$ contains at least one point $u_{n'}/\int u_{n'} \, d\mu$ in $\mathscr{P}_n$ that is defined as above. Hence,

$$\Pi(\{Q \in \mathscr{P} : d_H(Q, P_0) < 2\varepsilon_n\}) \geq \frac{\lambda_n}{N_n} \geq \exp(-n\varepsilon_n^2 - O(\log n)) \geq \exp(-2n\varepsilon_n^2).$$

Therefore taking change of variables on two sides of the inequality above, one could get,

$$\Pi(B_n) = \Pi(\{Q \in \mathscr{P} : d_H(Q, P_0) < \varepsilon_n\}) \geq \exp(-n\varepsilon_n^2/2).$$

We thus proved this theorem by setting $C' = 1/2$. □

C.2.3 Proofs for Section 4

Proof of theorem 4.12

For each $n \geq 1$, let us first introduce a new test in the following,

$$\phi_n = \max_{1 \leq i \leq N_n} \phi_{n,i},$$

then we can bound the posterior distribution $\Pi(V_n | X_1, \ldots, X_n)$ by,

$$\phi_n + \Pi(V_n | X_1, \ldots, X_n)(1 - \phi_n). \tag{C.15}$$

The expected value under the n-th posterior P_0^n of the first term above vanishes by (4.32), i.e.

$$P_0^n \phi_n = P_0^n \max_{1 \leq i \leq N_n} \phi_{n,i} \leq \sum_{i=1}^{N_n} P_0^n \phi_{n,i} \leq N_n e^{-Ln\varepsilon_n^2} \leq e^{-\frac{1}{2}Ln\varepsilon_n^2},$$

which goes to zero as $n \to \infty$.

Next, we split the second term of (C.15) in parts on $\mathscr{P}_n$ and its complement as follows,

$$\Pi(V_n \cap \mathscr{P}_n | X_1, \ldots, X_n)(1 - \phi_n) + \Pi(V_n \cap \mathscr{P}_n^c | X_1, \ldots, X_n)(1 - \phi_n). \tag{C.16}$$

By the construction of the test ϕ_n, the first term in (C.16) under P_0^n could be written as,

$$\begin{aligned} P_0^n \Pi(V_n \cap \mathscr{P}_n | X_1, \ldots, X_n)(1 - \phi_n) &\leq \sum_{i=1}^{N_n} P_0^n \Pi(V_{n,i} \cap \mathscr{P}_n | X_1, \ldots, X_n)(1 - \phi_{n,i}) \\ &\leq \sum_{i=1}^{N_n} \sup_{P \in V_{n,i}} P_0^n \frac{dP^n}{dP_n^\Pi}(1 - \phi_{n,i}) \\ &\leq N_n e^{-Ln\varepsilon_n^2} \leq e^{-\frac{1}{2}Ln\varepsilon_n^2}. \end{aligned}$$

Finally, applying (C.11) with $\alpha = 1$ enables us to write out the second term in (C.16),

$$\begin{aligned}
P_0^n \Pi(V_n \cap \mathscr{P}_n^c | X_1, \ldots, X_n)(1-\phi_n) &\leq P_0^n \Pi(V_n \cap \mathscr{P}_n^c | X_1, \ldots, X_n) \\
&= \int_{V_n \cap \mathscr{P}_n^c} P_0^n \frac{dP^n}{dP_n^{\Pi}}\, d\Pi(P) \\
&\leq \frac{1}{\Pi(B_n)} \int_{V_n \cap \mathscr{P}_n^c} P_0^n \int_{B_n} \frac{dP^n}{dQ^n}\, d(Q|B_n)\, d\Pi(P) \\
&\leq \frac{\Pi(V_n \cap \mathscr{P}_n^c)}{\Pi(B_n)} \sup_{P \in V_n \cap \mathscr{P}_n^c} \sup_{Q \in B_n} P_0^n \Big(\frac{dP^n}{dQ^n}\Big) \\
&\leq \frac{\Pi(\mathscr{P}_n^c)}{\Pi(B_n)} \left[\sup_{P \in V_n \cap \mathscr{P}_n^c} \sup_{Q \in B_n} P_0 \Big(\frac{dP}{dQ}\Big)\right]^n \\
&\leq \frac{\Pi(\mathscr{P}_n^c)}{\Pi(B_n)} e^{Kn\varepsilon_n^2/4},
\end{aligned}$$

which is bounded by $e^{-Kn\varepsilon_n^2/4}$ by conditions *(iii.)* and *(iv.)*. Since $n\varepsilon_n^2 \to \infty$ as $n \to \infty$, so that all the terms tend to zero. □

PROOF OF THEOREM 4.14

The proof of Theorem 4.4 in Kleijn (2015) goes through with a substitution of $V_n, V_{n,i}$ and $B_{n,i}$ for V, V_i and B_i respectively. □

PROOF OF COROLLARY 4.15

An application of theorem 4.14 with $\alpha = 1/2$ and choosing $B_{n,i} = B_n$ for each $i \geq 1, n \geq 1$ implies that,

$$\begin{aligned}
P_0^n \Pi(V_n | X_1, \ldots, X_n \leq \Pi(B_n)^{-1/2} \sum_{i=1}^{\infty} \Pi(V_{n,i})^{1/2} &\left[\sup_{P \in \mathrm{co}(V_{n,i})} \sup_{Q \in B_n} P_0 \Big(\frac{dP}{dQ}\Big)^{1/2}\right]^n \\
&\leq e^{\frac{1}{2}K' n\varepsilon_n^2} e^{-K'' n\varepsilon_n^2} \sum_{i=1}^{\infty} \Pi(V_{n,i})^{1/2},
\end{aligned}$$

which tends to zero by applying (4.42) with $\tilde{K} = K'' - \frac{1}{2}K'$. □

PROOF OF COROLLARY 4.16

Since $\mathscr{P}$ is Hellinger separable, then there exist an infinite countable number of model subsets $\{V_{n,m}\}_{m \geq 1}$ that cover V_n, where $V_n = \{P \in \mathscr{P} : d_H(P, P_0) > M\varepsilon_n\}$, $M > 0$ and $V_{n,m} = \{P \in \mathscr{P} :$

$d_H(P_m, P) < \varepsilon_n\}$, $P_m \in V_n$ for $m = 1, 2, \ldots$. Then applying the triangle inequality to give that,

$$\inf_{m \geq 1} \inf_{P \in \mathrm{co}(V_{n,m})} d_H(P, P_0) > (M-1)\varepsilon_n.$$

Note that for each $P \in \mathrm{co}(V_{n,m}), m = 1, 2, \ldots$,

$$-P_0 \log \frac{dP}{dP_0} \geq d_H^2(P, P_0) > (M-1)^2 \varepsilon_n^2.$$

Let's choose $B_n = \{P \in \mathscr{P} : -P_0 \log \frac{dP}{dP_0} < \varepsilon_n^2\}$, then one can obtain for $m = 1, 2, \ldots$,

$$\sup_{Q \in B_n} \left\{ -P_0 \log \frac{dQ}{dP_0} \right\} - \inf_{P \in \mathrm{co}(V_{n,m})} \left\{ -P_0 \log \frac{dP}{dP_0} \right\} < -(M^2 - 2M)\varepsilon_n^2.$$

Noting also that (4.17) in lemma 4.5 holds with $\mathrm{co}(V_{n,m})$ instead of $V_{n,m}$ there as well, that is , for $m = 1, 2, \ldots$,

$$\inf_{0 \leq \beta \leq 1} \sup_{Q \in B_n} \sup_{P \in \mathrm{co}(V_{n,m})} \left\{ P_0 \left(\frac{dP}{dQ} \right)^\beta \right\} < e^{-(M^2 - 2M - 1)\varepsilon_n^2}.$$

Hence, invoking theorem 4.14 to give rise to,

$$P_0^n \Pi(V_n | X_1, \ldots, X_n) \leq e^{\tilde{K}_1 n \varepsilon_n^2} e^{-(M^2 - 2M - 1)n\varepsilon_n^2} \inf_{0 \leq \beta \leq 1} \sum_{i=1}^{\infty} \Pi(V_{n,i})^\beta.$$

Choosing $M > 0$ sufficiently large could guarantee that $\tilde{K}_2 = M^2 - 2M - 1 - \tilde{K}_1 > 0$, so that the *r.h.s* of the preceding display goes to zero by (4.44). □

C.2.4 Proofs for Section 5

PROOF OF THEOREM 4.17

Let $\varepsilon_n = n^{-1/\beta}$ and define,

$$\begin{aligned} g(\theta,\theta') &:= \max\{|\theta_1 - \theta'_1|,\, |\theta_2 - \theta'_2|\}, \\ V_n &:= \{P_{\theta,\eta} \in \mathscr{P} :\, g(\theta,\theta') > M_n\varepsilon_n\}, \end{aligned}$$

for each $\theta := (\theta_1, \theta_2) \in \Theta$ and $\theta' := (\theta'_1, \theta'_2) \in \Theta$. Notice that B_n is contained in any Hellinger ball around P_0 and then $\Pi(B_n) > 0$ by (4.48), so that $P_0^n \ll P_n^\Pi$ holds for all $n \geq 1$ by Lemma 2.1 in Kleijn (2015).

Applying Lemma A.1 in Kleijn (2015) with $P \in V_n$ and $Q \in B_n$ and Hölder's inequality yields,

$$\begin{aligned} \Pi(B_n)^{-\alpha/n} P_0\Bigl(\frac{dP}{dQ}\Bigr)^{\alpha}\Bigg|_{\alpha\downarrow 0} &= P_0(p > 0), \\ \Pi(B_n)^{-\alpha/n} P_0\Bigl(\frac{dP}{dQ}\Bigr)^{\alpha}\Bigg|_{\alpha\uparrow 1} &= \Pi(B_n)^{-1/n}\left(\int \frac{dP_0}{dQ}\, 1_{\{p_0>0,p>0,q>0\}}\, dP\right) \\ &\leq \Pi(B_n)^{-1/n}\left(P(p_0 > 0) + \int \Bigl|\frac{dP_0}{dQ} - 1\Bigr| 1_{\{q>0\}}\, dP\right) \\ &\leq \Pi(B_n)^{-1/n}\left(P(p_0 > 0) + \Bigl\|\frac{dP_0}{dQ} - 1\Bigr\|_{s,Q}\Bigl\|\frac{dP}{dQ}\Bigr\|_{r,Q}\right) \\ &< \Pi(B_n)^{-1/n}\left(P(p_0 > 0) + \tfrac{1}{2}\bigl(\tfrac{\varepsilon_n}{\sigma}\bigr)^{\beta}\right). \end{aligned}$$

Hence it is easy to see that,

$$\inf_{0\leq\alpha\leq 1}\sup_{Q\in B_n}\sup_{P\in \mathrm{co}(V_n)} \Pi(B_n)^{-\alpha}\left[P_0\left(\frac{dP}{dQ}\right)^{\alpha}\right]^n,$$

is bounded above by,

$$\min\Bigl\{\sup_{Q\in B_n}\sup_{P\in \mathrm{co}(V_n)} \bigl[P_0(p > 0)\bigr]^n,\ \sup_{Q\in B_n}\sup_{P\in \mathrm{co}(V_n)} \Pi(B_n)^{-1}\Bigl[P(p_0 > 0) + \tfrac{1}{2}\bigl(\tfrac{\varepsilon_n}{\sigma}\bigr)^{\beta}\Bigr]^n\Bigr\}. \tag{C.17}$$

That is to say that we just concentrate in two cases to characterize the bound of the Hellinger transform above. Next, we just consider whether each element in (C.17) has the desired exponential bound. Let $P_0 = P_{(\theta_{0,1},\theta_{0,2}),\eta_0}$ and $P = P_{(\theta_1,\theta_2),\eta}$. It is possible to construct the covers of V_n in the following:

$$\begin{aligned}
V_{n,+,1} &= \{P_{\theta,\eta} : \theta_1 \geq \theta_{0,1} + M_n\varepsilon_n,\, \eta \in H\},\\
V_{n,-,1} &= \{P_{\theta,\eta} : \theta_1 \leq \theta_{0,1} - M_n\varepsilon_n,\, \eta \in H\},\\
V_{n,+,2} &= \{P_{\theta,\eta} : \theta_2 \geq \theta_{0,2} + M_n\varepsilon_n,\, \eta \in H\},\\
V_{n,-,2} &= \{P_{\theta,\eta} : \theta_2 \leq \theta_{0,2} - M_n\varepsilon_n,\, \eta \in H\}.
\end{aligned}$$

Now we explore the upper bound of (C.17) in these four cases above respectively.

(i) Let $P \in \mathrm{co}(V_{n,+,1})$, here we consider the lower bound of $P_0(p = 0)$,

$$\begin{aligned}
P_0(p=0) &\geq \int_{\theta_{0,1}}^{\theta_{0,1}+M_n\varepsilon_n} p_0(x)\,dx = \int_{\theta_{0,1}}^{\theta_{0,1}+M_n\varepsilon_n} \frac{1}{\theta_{0,2}-\theta_{0,1}}\eta_0\Big(\frac{x-\theta_{0,1}}{\theta_{0,2}-\theta_{0,1}}\Big)dx\\
&= \int_0^{M_n\varepsilon_n/(\theta_{0,2}-\theta_{0,1})} \eta_0(z)\,dz \geq \int_0^{\frac{M_n\varepsilon_n}{\sigma}} \eta_0(z)\,dz \geq \big(\tfrac{M_n\varepsilon_n}{\sigma}\big)^\beta,
\end{aligned}$$

the last inequality is derived from (4.47). Then we get,

$$(C.17) \leq \sup_{Q\in B_n}\ \sup_{P\in\mathrm{co}(V_{n,+,1})} \big[P_0(p>0)\big]^n \leq \big[1-\big(\tfrac{M_n\varepsilon_n}{\sigma}\big)^\beta\big]^n \leq e^{-n\left(\frac{M_n\varepsilon_n}{\sigma}\right)^\beta}.$$

(ii) Let $P \in \mathrm{co}(V_{n,-,1})$, assume that there exist some positive integer I and $\lambda_1, \lambda_2, \ldots, \lambda_I \geq 0$ such that $\sum_{i=1}^I \lambda_i = 1$. Write $P_i = P_{\theta_i,\eta_i}$ with $\theta_{i,1} \leq \theta_{0,1} - M_n\varepsilon_n$ as well as $\eta_i \in H$, for all $1 \leq i \leq I$. Here we take into account the lower bound of $P(p_0 = 0)$. Using (4.47), it is not difficult to see that,

$$\begin{aligned} P(p_0 = 0) &= \sum_{i=1}^{I} \lambda_i\, P_i(p_0 = 0) \geq \sum_{i=1}^{I} \lambda_i \int_{\theta_{i,1}}^{\theta_{i,1}+M_n\varepsilon_n} p_i(x)\, dx \\ &= \sum_{i=1}^{I} \lambda_i \int_{\theta_{i,1}}^{\theta_{i,1}+M_n\varepsilon_n} \frac{1}{\theta_{i,2}-\theta_{i,1}} \eta_i\Big(\frac{x-\theta_{i,1}}{\theta_{i,2}-\theta_{i,1}}\Big) dx \\ &= \sum_{i=1}^{I} \lambda_i \int_{0}^{M_n\varepsilon_n/(\theta_{i,2}-\theta_{i,1})} \eta_i(z)\, dz \geq \sum_{i=1}^{I} \lambda_i \int_{0}^{\frac{M_n\varepsilon_n}{\sigma}} \eta_i(z)\, dz \geq \big(\tfrac{M_n\varepsilon_n}{\sigma}\big)^{\beta}. \end{aligned}$$

Then we have,

$$\begin{aligned} (C.17) &\leq \sup_{Q\in B_n} \sup_{P\in \mathrm{co}(V_{n,-,1})} \Pi(B_n)^{-1} \Big[P(p_0 > 0) + \tfrac{1}{2}\big(\tfrac{\varepsilon_n}{\sigma}\big)^{\beta}\Big]^n \\ &\leq e^{\tilde{L}n\delta_n^2}\Big[1 - \big(\tfrac{M_n\varepsilon_n}{\sigma}\big)^{\beta} + \tfrac{1}{2}\big(\tfrac{\varepsilon_n}{\sigma}\big)^{\beta}\Big]^n \\ &\leq e^{\tilde{L}n\left(\frac{\varepsilon_n}{\sigma}\right)^{\beta}/2T} e^{-n\left(\frac{M_n\varepsilon_n}{\sigma}\right)^{\beta} + \frac{n}{2}\left(\frac{\varepsilon_n}{\sigma}\right)^{\beta}} \\ &\leq e^{-Cn\left(\frac{M_n\varepsilon_n}{\sigma}\right)^{\beta}}. \end{aligned}$$

The last inequality holds since $(\tilde{L}/2T + 1/2)\big(\frac{\varepsilon_n}{\sigma}\big)^{\beta} - \big(\frac{M_n\varepsilon_n}{\sigma}\big)^{\beta} < -C\big(\frac{M_n\varepsilon_n}{\sigma}\big)^{\beta}$ for some positive constant C when M_n is chosen sufficiently large as $n \to \infty$.

(iii) Let $P \in \mathrm{co}(V_{n,+,2})$, assume again that there exist some positive integer I and $\lambda_1, \lambda_2, \ldots, \lambda_I \geq 0$ such that $\sum_{i=1}^{I} \lambda_i = 1$. Write $P_i = P_{\theta_i,\eta_i}$ with $\theta_{i,2} \geq \theta_{0,2} + M_n\varepsilon_n$ and $\eta_i \in H$, for all $1 \leq i \leq I$. Observe that by (4.47),

$$\begin{aligned} P(p_0 = 0) &= \sum_{i=1}^{I} \lambda_i\, P_i(p_0 = 0) \geq \sum_{i=1}^{I} \lambda_i \int_{\theta_{i,2}-M_n\varepsilon_n}^{\theta_{i,2}} p_i(x)\, dx \\ &= \sum_{i=1}^{I} \lambda_i \int_{\theta_{i,2}-M_n\varepsilon_n}^{\theta_{i,2}} \frac{1}{\theta_{i,2}-\theta_{i,1}} \eta_i\Big(\frac{x-\theta_{i,1}}{\theta_{i,2}-\theta_{i,1}}\Big) dx \\ &= \sum_{i=1}^{I} \lambda_i \int_{1-M_n\varepsilon_n/(\theta_{i,2}-\theta_{i,1})}^{1} \eta_i(z)\, dz \geq \sum_{i=1}^{I} \lambda_i \int_{1-M_n\varepsilon_n/\sigma}^{1} \eta_i(z)\, dz \geq \big(\tfrac{M_n\varepsilon_n}{\sigma}\big)^{\beta}. \end{aligned}$$

Then we obtain,

$$
\begin{aligned}
(C.17) &\leq \sup_{Q\in B_n} \sup_{P\in \mathrm{co}(V_{n,+,2})} \Pi(B_n)^{-1}\left[P(p_0>0)+\tfrac{1}{2}\big(\tfrac{\varepsilon_n}{\sigma}\big)^{\beta}\right]^n\\
&\leq e^{\tilde{L}n\delta_n^2}\left[1-\big(\tfrac{M_n\varepsilon_n}{\sigma}\big)^{\beta}+\tfrac{1}{2}\big(\tfrac{\varepsilon_n}{\sigma}\big)^{\beta}\right]^n\\
&\leq e^{\tilde{L}n\left(\frac{\varepsilon_n}{\sigma}\right)^{\beta}/2T}e^{-n\left(\frac{M_n\varepsilon_n}{\sigma}\right)^{\beta}+\frac{n}{2}\left(\frac{\varepsilon_n}{\sigma}\right)^{\beta}}\\
&\leq e^{-Cn\left(\frac{M_n\varepsilon_n}{\sigma}\right)^{\beta}}.
\end{aligned}
$$

The last inequality follows since $(\tilde{L}/2T+1/2)\big(\frac{\varepsilon_n}{\sigma}\big)^{\beta}-\big(\frac{M_n\varepsilon_n}{\sigma}\big)^{\beta}<-C\big(\frac{M_n\varepsilon_n}{\sigma}\big)^{\beta}$ for some positive constant C when M_n is chosen sufficiently large as $n\to\infty$.

(iv) Let $P\in \mathrm{co}(V_{n,-,2})$, then $\theta_{0,2}\leq\theta_{0,2}-M_n\varepsilon_n$. Note that, using (4.47), we see that,

$$
\begin{aligned}
P_0(p=0) &\geq \int_{\theta_{0,2}-M_n\varepsilon_n}^{\theta_{0,2}} p_0(x)\,dx=\int_{\theta_{0,2}-M_n\varepsilon_n}^{\theta_{0,2}}\frac{1}{\theta_{0,2}-\theta_{0,1}}\eta_0\Big(\frac{x-\theta_{0,1}}{\theta_{0,2}-\theta_{0,1}}\Big)dx\\
&=\int_{1-M_n\varepsilon_n/(\theta_{0,2}-\theta_{0,1})}^{1}\eta_0(z)\,dz\\
&\geq\int_{1-M_n\varepsilon_n/\sigma}^{1}\eta_0(z)\,dz\\
&\geq\big(\tfrac{M_n\varepsilon_n}{\sigma}\big)^{\beta}.
\end{aligned}
$$

Hence we get,

$$
(C.17)\leq\sup_{Q\in B_n}\sup_{P\in\mathrm{co}(V_{n,-,2})}\big[P_0(p>0)\big]^n\leq\big[1-\big(\tfrac{M_n\varepsilon_n}{\sigma}\big)^{\beta}\big]^n\leq e^{-n\left(\frac{M_n\varepsilon_n}{\sigma}\right)^{\beta}}.
$$

So that it follows that for $\tilde{C}=\min(1,C)$,

$$
\inf_{0\leq\alpha\leq1}\sup_{Q\in B_n}\sup_{P\in\mathrm{co}(V_{n,\cdot})}\Pi(B_n)^{-\alpha}\left[P_0\Big(\frac{dP}{dQ}\Big)^{\alpha}\right]^n\leq e^{-\tilde{C}n\left(\frac{M_n\varepsilon_n}{\sigma}\right)^{\beta}},
$$

for $V_{n,\cdot}$ equal to $V_{n,+,1}$, $V_{n,-,1}$, $V_{n,+,2}$ and $V_{n,-,2}$. The assumption $\varepsilon_n=n^{-1/\beta}$ suffices to assure that $n\big(\frac{M_n\varepsilon_n}{\sigma}\big)^{\beta}\to\infty$ and $e^{-\tilde{C}n\left(\frac{M_n\varepsilon_n}{\sigma}\right)^{\beta}}\to0$ as $n\to\infty$.

Therefore a combination of theorem 4.2, lemma 4.3 and lemma 4.4 implies that,

$$\Pi\big(g(\theta,\theta_0) > M_n\varepsilon_n \;\big|\; X_1,\ldots,X_n \big) \longrightarrow 0 \text{ in } P_0^n\text{-probability.}$$

Since all the norms defined on the finite dimension space are equivalent, then (4.50) follows. Thus the proof of this theorem is complete. □

PROOF OF LEMMA 4.18

Let us consider the following norm,

$$\|f\|_{2,G} := \left\{ \int_0^1 f^2(t)g(t)\,dt \right\}^{1/2} = \left\{ \int_0^1 f^2(t)\,dG(t) \right\}^{1/2}.$$

Note that $p_v(t)$ and $p_w(t)$ can be rewritten as follows:

$$p_v(t) = \frac{e^{v(t)}g(t)}{\|e^{v/2}\|_{2,G}} \quad \text{and} \quad p_w(t) = \frac{e^{w(t)}g(t)}{\|e^{w/2}\|_{2,G}}.$$

By the same arguments in the proof of lemma 3.1 in van der Vaart and van Zanten (2008), we see that,

$$\begin{aligned} d_H(p_v,p_w) &= \left\| \frac{e^{v/2}}{\|e^{v/2}\|_{2,G}} - \frac{e^{w/2}}{\|e^{w/2}\|_{2,G}} \right\|_{2,G} \\ &= \left\| \frac{e^{v/2}-e^{w/2}}{\|e^{v/2}\|_{2,G}} + e^{v/2}\Big(\frac{1}{\|e^{v/2}\|_{2,G}} - \frac{1}{\|e^{w/2}\|_{2,G}} \Big) \right\|_{2,G} \\ &\le \frac{2\|e^{v/2}-e^{w/2}\|_{2,G}}{\|e^{w/2}\|_{2,G}}. \end{aligned}$$

Since $|e^{v/2}-e^{w/2}| = e^{w/2}|e^{v/2-w/2}-1| < e^{w/2}e^{|v-w|/2}|v-w|/2$, then the squared Hellinger distance is further bounded from above by,

$$\frac{\int_0^1 e^{w(t)}e^{|v(t)-w(t)|}|v(t)-w(t)|^2\,dG(t)}{\int_0^1 e^{w(t)}\,dG(t)} \le \|v-w\|_\infty^2 \times e^{\|v-w\|_\infty}.$$

Then this concludes the proof. □

Bibliography

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174.

Barron, A., Schervish, M. J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561.

Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *The Annals of Statistics*, 31(2):536–559.

Belitser, E. and Serra, P. (2014). Adaptive priors based on splines with random knots. *Bayesian Analysis*, 9(4):859–882.

Bernstein, S. (1912). On the best approximation of continuous functions by polynomials of a given degree. *Communications of the Kharkov Mathematical Society, Series XIII*, 2:49–194.

Bickel, P. (1982). On adaptive estimation. *The Annals of Statistics*, 10(3):647–671.

Bickel, P. and Kleijn, B. (2012). The semiparametric Bernstein-von Mises theorem. *The Annals of Statistics*, 40(1):206–237.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore, Maryland.

Billingsley, P. (1995). *Probability and Measure*. John Wiley & Sons, New York.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65(2):181–237.

Birgé, L. (1984). Sur un théorème de minimax et son application aux tests. *Probability and Mathematical Statistics*, 3(2):259–282.

Blundell, R. and Powell, J. L. (2003). Endogeneity in nonparametric and semiparametric regression models. *Econometric Society Monographs*, 36:312–357.

Chen, X. and Pouzo, D. (2013). Sieve quasi likelihood ratio inference on semi/nonparametric conditional moment models.

Chib, S. and Greenberg, E. (2013). On conditional variance estimation in nonparametric regression. *Statistics and Computing*, 23(2):261–270.

Crainiceanu, C. M., Ruppert, D., Carroll, R. J., Joshi, A., and Goodner, B. (2007). Spatially adaptive bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, 16(2).

de Boor, C. (2001). *A Practical Guide to Splines*. Springer, New York.

de Jonge, R. and van Zanten, J. H. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *The Annals of Statistics*, 38(6):3300–3320.

de Jonge, R. and van Zanten, J. H. (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electronic Journal of Statistics*, 6:1984–2001.

Efroimovich, S. Y. (1986). Nonparametric estimation of a density of unknown smoothness. *Theory of Probability & Its Applications*, 30(3):557–568.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403.

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *The Annals of Statistics*, 29:1264–1280.

Ghosal, S. (2010). The Dirichlet process, related priors and posterior asymptotics. In Hjort, N., Holmes, C., and Walker, S., editors, *Bayesian Nonparametrics*, pages 35–79. Cambridge University Press, New York.

Ghosal, S., Ghosh, J., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158.

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5):1233–1263.

Ghosal, S. and van der Vaart, A. W. (2007a). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223.

Ghosal, S. and van der Vaart, A. W. (2007b). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723.

Ghosh, J. and Ramamoorthi, R. (2003a). *Bayesian Nonparametrics*. Springer, Berlin.

Ghosh, J. K. and Ramamoorthi, R. (2003b). *Bayesian nonparametrics*, volume 1. New York, NY: Springer New York.

Giné, E. and Nickl, R. (2011). Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *The Annals of Statistics*, 39(6):2883–2911.

Goldberg, P. W., Williams, C. K., and Bishop, C. M. (1997). Regression with input-dependent noise: A Gaussian process treatment. *Advances in Neural Information Processing Systems*, 10:493–499.

Hjort, N., Holmes, C., Müller, P., and Walker, S. (2010). *Bayesian Nonparametrics*. Cambridge University Press, New York.

Huang, T.-M. (2004). Convergence rates for posterior distributions and adaptive estimation. *The Annals of Statistics*, 32(4):1556–1593.

Ibragimov, I. A. and Hasminskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*, volume 2. Springer-Verlag, New York.

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of statistical software*, 40(5):1–30.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and computing*, 21(1):93–105.

Kato, K. (2013). Quasi-Bayesian analysis of nonparametric instrumental variables models. *The Annals of Statistics*, 41(5):2359–2390.

Khazaei, Soleiman, R. J. and Balabdaoui-Mohr, F. (2010). Bayesian Nonparametric Inference of decreasing densities. In *42èmes Journées de Statistique*, Marseille, France.

Kleijn, B. (2003). *Bayesian asymptotics under misspecification*. PhD thesis, VU University Amsterdam.

Kleijn, B. (2015). Criteria for posterior consistency. *Submitted for publication*.

Kleijn, B. and Knapik, B. (2012). Semiparametric posterior limits under local asymptotic exponentiality. *arXiv preprint arXiv:1210.6204*.

Kleijn, B. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34(2):837–877.

Kleijn, B. and van der Vaart, A. W. (2012). The Bernstein-von Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Kruijer, W., Rousseau, J., and van der Vaart, A. W. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257.

Kruijer, W. and van der Vaart, A. W. (2008). Posterior convergence rates for Dirichlet mixtures of beta densities. *Journal of Statistical Planning and Inference*, 138(7):1981–1992.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53.

Le Cam, L. (1975). On local and global properties in the theory of asymptotic normality of experiments. *Stochastic Processes and Related Topics*, 1:13–54.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

Lepskii, O. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466.

Lepskii, O. (1992). Asymptotically minimax adaptive estimation. I: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697.

Lepskii, O. (1993). Asymptotically minimax adaptive estimation. II. schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications*, 37(3):433–448.

Liao, Y. and Jiang, W. (2011). Posterior consistency of nonparametric conditional moment restricted models. *The Annals of Statistics*, 39(6):3003–3031.

Lin, L. and Dunson, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika*, 101(2):303–317.

Lorentz, G. (1986). *Bernstein Polynomials*. Chelsea, New York.

Norets, A. (2015). Bayesian regression with nonparametric heteroskedasticity. *Journal of Econometrics*, 185(2):409–419.

Norets, A. and Pati, D. (2014). Adaptive Bayesian estimation of conditional densities. *arXiv preprint arXiv:1408.5355*.

Olkin, I. and Liu, R. (2003). A bivariate beta distribution. *Statistics & Probability Letters*, 62(4):407–412.

Papaspiliopoulos, O. and Roberts, G. (2008). Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, 95(1):169–186.

Pelenis, J. (2014). Bayesian regression with heteroscedastic error density and parametric mean function. *Journal of Econometrics*, 178(3):624–638.

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics*, 27(1):105–126.

Petrone, S. (1999b). Random Bernstein polynomials. *Scandinavian Journal of Statistics*, 26(3):373–393.

Petrone, S. and Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(1):79–100.

Phillips, G. (2003). *Interpolation and Approximation by Polynomials*. Springer, New York.

Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *The Annals of Statistics*, 38(1):146–180.

Rousseau, J. (2013). Some recent advances in the asymptotic properties of Bayesian nonparametric approaches. *Invited talk at BNP9 conference*.

Salomond, J.-B. (2013). Concentration rate and consistency of the posterior under monotonicity constraints. *arXiv preprint arXiv:1301.1898*.

Sancetta, A. and Satchell, S. (2004). The Bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric Theory*, 20(03):535–562.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press, New York.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26.

Scricciolo, C. (2006). Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *The Annals of Statistics*, 34(6):2897–2920.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Shen, W. and Ghosal, S. (2012). MCMC-free adaptive Bayesian procedures using random series prior. *arXiv preprint arXiv:1204.4238*.

Shen, W. and Ghosal, S. (2014). Adaptive Bayesian density regression for high-dimensional data. *arXiv preprint arXiv:1403.2695*.

Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714.

Shively, T. S., Sager, T. W., and Walker, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):159–175.

Shively, T. S., Walker, S. G., and Damien, P. (2011). Nonparametric function estimation subject to monotonicity, convexity and other shape constraints. *Journal of Econometrics*, 161(2):166–181.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, 22(1):118–171.

Szabo, B. (2014). *Adaptation and confidence in nonparametric Bayes*. PhD thesis, Eindhoven University of Technology.

Szabo, B., van der Vaart, A. W., and van Zanten, J. H. (2013). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *arXiv preprint arXiv:1310.4489*.

Tenbusch, A. (1994). Two-dimensional Bernstein polynomial density estimators. *Metrika*, 41:233–253.

Trippa, L., Bulla, P., and Petrone, S. (2011). Extended Bernstein prior via reinforced urn processes. 63(3):481–496.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435.

van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, 37(5):2655–2675.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Process*. Springer, New York.

Vitale, R. (1975). A Bernstein polynomial approach to density function estimation. *Statistical Inference and Related Topics*, 2:87–99.

Walker, S. (2004). New approaches to Bayesian consistency. *The Annals of Statistics*, 32(5):2028–2043.

Walker, S. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation*, 36(1):45–54.

Walker, S. and Hjort, N. L. (2001). On Bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821.

Walker, S., Lijoi, A., and Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *The Annals of Statistics*, 35(2):738–746.

Wang, J. and Ghosh, S. (2012). Shape restricted nonparametric regression with bernstein polynomials. *Computational Statistics & Data Analysis*, 56(9):2729–2741.

Wang, L. (2013). Consistency of posterior distributions for heteroscedastic nonparametric regression models. *Communications in Statistics-Theory and Methods*, 42(15):2731–2740.

Wiesenfarth, M., Hisgen, C. M., Kneib, T., and Cadarso-Suarez, C. (2014). Bayesian nonparametric instrumental variables regression based on penalized splines and dirichlet process mixtures. *Journal of Business & Economic Statistics*, 32(3):468–482.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, 23(2):339–362.

Yau, P. and Kohn, R. (2003). Estimation and variable selection in nonparametric heteroscedastic regression. *Statistics and Computing*, 13(3):191–208.

Zheng, Y., Zhu, J., and Roy, A. (2010). Nonparametric Bayesian inference for the spectral density function of a random field. *Biometrika*, 97(1):238–245.