ABSTRACT

SHEN, WEINING. Adaptive Bayesian Function Estimation. (Under the direction of Subhashis Ghosal.)

This dissertation focuses on developing some new Bayesian methodologies for function estimation and studying their theoretical properties. In particular, we investigate conditions under which such methods achieve the optimal posterior rate of convergence even when the smoothness of the underlying function is unknown. Some examples of functions of interest include density function, conditional density function, regression function, and classification function. Although several nonparametric Bayesian models have been developed for many applications, their theoretical properties, such as large-sample convergence properties are often not fully understood. Part of the dissertation focuses on exploring the posterior convergence rates of certain nonparametric Bayesian procedures. We establish rate-adaptive Bayesian procedures in the sense that the optimal minimax rate of estimation can be achieved by using one single prior for the entire smoothness class that the underlying true function belongs to. Such results can be viewed as frequentist large-sample justification, which suggests that by carefully selecting priors, the posterior distribution of the estimator concentrates around the truth at an optimal rate.

In Chapter 2, we consider a multivariate density estimation problem. We show that rate-adaptive Bayesian procedures can be obtained using Dirichlet mixtures of multivariate normal kernels with a prior distribution on the kernels covariance matrix parameter. Locally Hölder smoothness classes and their anisotropic extensions are considered. Our study involves several technical novelties, including a sharp approximation of finitely differentiable multivariate densities by normal mixtures and a new sieve on the space of such densities.

In Chapter 3, we consider a class of finite random series priors for univariate function

estimation problems. The prior is constructed through distributions on the number of basis functions and the associated coefficients. We derive a general result on the construction of an appropriate sieve and obtain adaptive posterior contraction rates. This general result is applied on several statistical problems such as signal processing, density estimation, nonparametric regression, classification, spectral density estimation, functional regression etc. The random series prior can be viewed as an alternative to the commonly used Gaussian process prior, but can be analyzed by relatively simpler techniques and in many cases allows a simpler approach to computation without using Markov chain Monte-Carlo (MCMC) methods.

In Chapter 4, we extend the random series prior to a multivariate setting. In particular, we consider a density regression problem (i.e., estimation of a conditional density function) in a high-dimensional situation, in which the number of covariates is possibly much larger than the sample size. We develop a MCMC-free computing technique to calculate the posterior moments of the conditional density. Adaptive convergence rate is obtained under sparsity conditions that adapts to both the smoothness level and the actual number of covariates that influence the conditional density.

Estimation of derivatives of density functions is another interesting topic as it directly connects to many important statistical quantities such as score functions, Fisher information and hypothesis testing in semiparametric location models. In Chapter 5, we focus on estimating derivatives of a univariate density function using Bayesian kernel methods and B-spline functions. Convergence rates are obtained using new techniques on constructing testing inequalities under L_r-norms, $1 \leq r \leq \infty$.

Adaptive Bayesian Function Estimation

by
Weining Shen

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

Statistics

Raleigh, North Carolina

2013

APPROVED BY:

Eric Laber	Surya T. Tokdar
Huixia Wang	Subhashis Ghosal Chair of Advisory Committee

DEDICATION

To my family

BIOGRAPHY

Weining was born in China. In 2004, he was admitted by Special Class for the Gifted Young (known as SCGY) at University of Science and Technology of China (USTC). There he discovered his interest in using mathematics and statistics to solve real-world problems. After he received a Bachelor degree in statistics in 2008. He attended North Carolina State University to pursue a PhD in statistics.

ACKNOWLEDGEMENTS

To begin with, I would like to express my deepest gratitude to my advisor Dr. Subhashis Ghosal for his generous support throughout my graduate studies, for his patience, motivation, enthusiasm, and immense knowledge. His tremendous work ethics and passion towards academics set me a good example to stay motivated and persistent.

I would also like to extend my appreciation to my committee members, Dr. Eric Laber, Dr. Surya T. Tokdar and Dr. Huixia Wang for their constant encouragement and wonderful advices. In particular, Chapter 2 is based on a joint work with Dr. Tokdar. It has been a great privilege to work with them. Thanks to Dr. Peter Bloomfield, Dr. Howard D. Bondell, Dr. Dennis D. Boos, Dr. David A. Dickey, Dr. Sujit K. Ghosh, Dr. Lexin Li, Dr. Wenbin Lu, Dr. John F. Monahan, Dr. Jason A. Osborne, Dr. Brian J. Reich, Dr. Shan Suthaharan, Dr. Anastasios Tsiatis, Dr. Daowen Zhang and Dr. Hao Zhang for teaching me so many wonderful courses and giving me inspiring comments in my research.

Thanks to my fellow students and friends at NCSU. Life will be much harder without your accompany. I am grateful to Regeneron Pharmaceuticals Inc and First Analytics for offering me intern-ship opportunities. Sincere thanks to my former colleagues for sharing their working and life experiences with me.

I am obliged to the National Science Foundation, the International Society for Bayesian Analysis, and the American Statistical Association for financing my graduate education and conference travel.

Finally, I dedicate this work to my family, your endless support and blessings has always been the source of my strength.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Overview of nonparametric Bayesian

Nonparametric models are widely used in many areas of statistic studies (Wasserman, 2006). Compared with parametric models, which assume that the distributions are based on only a few parameters in a particular functional form, nonparametric models are more flexible in the sense of having a much larger parameter space and relying upon fewer assumptions. Nonparametric inference covers a wide range of different topics such as constructing test procedures that are "distribution-free" at least in the asymptotic sense, estimating functions without any parametric assumptions, computation techniques such as bootstrap etc. In the thesis, we focus on function estimation problems. Some examples of functions of interest include density function, cumulative distribution function, quantile function, regression function, classification function, spectral density, hazard rate function and so on.

The likelihood method has been used extremely often in estimations. However, instead of only relying the data, in many situations there is a need to incorporate outside information in the model along the data, where the Bayesian approach emerges as a useful solution. Moreover,

there are many attractive properties of the Bayesian approach that makes it quite popular. For example, the inference is straightforward once the posterior distribution is obtained. It is simple to obtain and interpret predictions and intervals estimates. With the development of innovative computational algorithms such as Monte-Carlo Markov chain (MCMC) and variational methods, Laplace approximation etc. and the availability of powerful computational resources, most Bayesian methods can now be implemented within a reasonable time.

Nonparametric Bayesian methods have been applied in many complex problems in biomedical, spatial, econometric and other studies. Compared with the parametric Bayesian approach, problems here are more challenging in several aspects. First, the construction of a prior on an infinite-dimensional space is more difficult. A full prior specification means infinite details, which is generally regarded as impossible. A desirable prior should have a large topological support, be computationally tractable, flexible enough to include one's belief and have good asymptotic properties. In many cases such a prior on a function is chosen either via a processes such as a Gaussian process, a Dirichlet process mixture, a Polya tree process, a gamma process etc. or by expanding a function through a basis expansion and assigning priors on the coefficients. Secondly, smart computational techniques are needed, especially in high dimensional and other complicated models. Thirdly, the convergence behavior of posterior under true distributions need more careful and involved investigation. We shall elaborate more in the following sections.

1.2 Common priors

1.2.1 Dirichlet process

We first discuss the Dirichlet process (DP), one of the most classic and popular choices of prior in the infinite-dimensional world. The concept was first introduced by Ferguson (1973). The formal definition is given below

Definition 1. *A random measure P is called a Dirichlet process on a measurable space $\mathfrak{X}$ with parameter α if for every finite measurable partition $\{B_1, \ldots, B_k\}$ of $\mathfrak{X}$, the joint distribution of $(P(B_1), \ldots, P(B_k))$ follows a Dirichlet distribution with parameters $(\alpha(B_1), \ldots, \alpha(B_k))$. We denote this Dirichlet process by $\mathrm{DP}(\alpha)$, where α is a finite measure on $\mathfrak{X}$.*

Dirichlet processes have many nice properties. Conceptually, it can be viewed as an infinite-dimensional generalization of the Dirichlet distribution. The Dirichlet process is conjugate for independent and identically distributed (i.i.d.) observations, which makes posterior computations straightforward. Moreover, it has an interesting stick-breaking representation (Sethuraman, 1994), which leads to some useful MCMC sampling algorithms for priors based on Dirichlet mixture processes. This result turns out to be very helpful in our study in Chapter 2. There are many other interesting properties as listed below.

- (Stick-breaking representation) A Dirichlet process $P \sim \mathrm{DP}(\alpha)$ can be represented as

$$P = \sum_{i=1}^{\infty} \pi_h \delta_{Z_h}, \quad \pi_h = V_h \prod_{j<h} (1 - V_j),\ V_i \sim \mathrm{Beta}(1, |\alpha|),\ Z_h \sim \bar{\alpha}, \tag{1.1}$$

  where $|\alpha| = \alpha(\mathfrak{X})$ and $\bar{\alpha} = \alpha/|\alpha|$.

- (Conjugacy) Given a DP prior $P \sim \mathrm{DP}(\alpha)$, suppose we want to estimate a completely unknown distribution function based on i.i.d observations $X = (X_1, \ldots, X_n)$. Then the

posterior distribution of P given X is $D_{\alpha+\sum_{i=1}^{n}\delta_{X_i}}$, where δ_x is a point mass probability measure on a point x.

- (Discreteness) An interesting property of a DP is that sampling distributions from it are always discrete, even when the base measure is continuous.

1.2.2 Dirichlet process mixture

When estimating a continuous density function, a Dirichlet process prior is not appropriate because it is discrete. Hence a Dirichlet process mixture is usually used (Ferguson, 1983; Lo, 1984) by convoluting a Dirichlet process with a probability density kernel function Ψ parameterized by θ as follows:

$$X_i \;\sim\; p_F(x) = \int \Psi(x,\theta)dF(\theta), i = 1,\ldots,n;\;\; F \sim \mathrm{DP}(\alpha). \tag{1.2}$$

By introducing latent variables $\theta_1,\ldots,\theta_n$, the model can be equivalently written as

$$X_i|\theta_i \sim \Psi(x,\theta_i),\;\; \theta_i|F \overset{\text{i.i.d.}}{\sim} F, i = 1,\ldots,n;\;\; F \sim \mathrm{DP}(\alpha). \tag{1.3}$$

Then by the conjugacy property of a Dirichlet process, we get the posterior of F given $\theta_1,\ldots,\theta_n$ as $F|\theta_1,\ldots,\theta_n \sim \mathrm{DP}(\alpha+\sum_{i=1}^{n}\delta_{\theta_i})$. Hence we obtain the posterior expectation of the density as

$$\mathrm{E}\{p_F(x)|\theta_1,\ldots,\theta_n\} = \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R})+n}\int \Psi(x,\theta)dG(\theta) + \frac{1}{n}\sum_{i=1}^{n}\Psi(x,\theta_i).$$

Similar expressions can be derived for other functionals of p_F. However, in practice, these expressions may be difficult to compute because there are a large number of terms involved.

Therefore, many MCMC computation algorithms have been developed including Gibbs sampling based on the Pólya urn schemes (Escobar and West, 1995; MacEachern and Müller, 1998), methods based on truncation of a Dirichlet process (Ishwaran and James, 2001), retrospective sampling methods (Papaspiliopoulos and Robers, 2008), and slice sampling methods (Walker, 2007).

1.2.3 Gaussian process

Another popular class of priors are constructed via Gaussian processes (Leonard, 1978; Lenk, 1988, 1991). A Gaussian process (GP) can be viewed as an infinite-dimensional generalization of a Gaussian distribution. The formal definition is as follows.

Definition 2. *A stochastic process $\{X_t, t \in T\}$ is called a Gaussian process if for every finite index set $(t_1, \ldots, t_k) \in T$, the joint distribution of $(X_{t_1}, \ldots, X_{t_k})$ is Gaussian.*

The distribution of a GP on $\mathbb{R}^T$ is uniquely determined by its mean function $\mu(t) = \mathrm{E}(X_t)$ and covariance kernel $K(s,t) = \mathrm{E}(X_s X_t)$. We give some examples here for illustration.

Example 1. *Brownian motion*

A Brownian motion B_t is a continuous stochastic process that was first used to model the random moving of a particle in the fluid. A standard Brownian motion on the positive line can be viewed as a Gaussian process with $\mu(t) = 0$ and $K(s,t) = \min(s,t)$.

Example 2. *Fractional Brownian motion*

Fractional Brownian motions generalize a standard Brownian motion without independent increments. Again, it is a Gaussian process with

$$\mu(t) = 0, \;\; K(s,t) = (s^{2\alpha} + t^{2\alpha} - |t-s|^{2\alpha})/2,$$

where $\alpha \in (0, 1)$ is called a Hurst parameter, which controls the correlation of increments of the process. When choosing $\alpha = 1/2$, we obtain a Brownian motion.

Example 3. *Integrated Brownian motion*

In order to smooth a Brownian motion, we take its integral and obtain integrated Brownian motions as a result. For example, the smoothness level of a k-fold integrated Brownian motion is $k + 1/2$. Again, such process is a Gaussian process with

$$\mu(t) = 0, \quad K(s,t) = \int_0^{\min(s,t)} \frac{(s-u)^k (t-u)^k}{(k!)^2} du.$$

Example 4. *Gaussian process with a squared exponential kernel*

Another example of the kernel is squared exponential function $K(s,t) = \exp\{-c(s-t)^2\}$, where c is a hyper-parameter. This kernel is stationary in the sense that it only depends on the difference of the inputs $(s - t)$. Such a kernel has been widely used in the kernel machine learning literature. The resulting Gaussian process has infinitely smooth sample paths.

Example 5. *Finite series expansion*

Consider k i.i.d random variables $Z_1, \ldots, Z_k \sim N(0, 1)$ and some real-valued functions $a_1, \ldots, a_k$. Then the stochastic process $W_t = \sum_{i=1}^k a_i(t) Z_i$ is a Gaussian process with

$$\mu(t) = 0, \quad K(s,t) = \sum_{i=1}^k a_i(s) a_i(t).$$

Gaussian process priors have received a lot of attention in both applications and theoretical studies. There are applications in density estimation problems (Lenk, 1988), regression problems (Neal, 1998), spatial statistics (Handcock and Stein, 1993), fast computations (Tokdar, 2007) and many other areas. Theoretical properties of such models have been well studied by van der Vaart and van Zanten (2007, 2008, 2009); Ghosal and Roy (2006); Tokdar and Ghosh

(2007) and many others. Theoretical properties of a GP can be elegantly expressed in terms of on its reproducing kernel Hilbert space (RKHS). van der Vaart and van Zanten (2009) applied a random rescaling technique on GP priors and showed that such procedures have nice asymptotic properties such as Bayesian adaptation etc.

1.2.4 Tail-free process

A tail-free process is another popular class of processes to construct priors for absolutely continuous probability distributions under certain conditions. Dirichlet process also belongs to this class. Tail-free process priors are constructed based on partitioning a set $\mathfrak{X} \subseteq \mathbb{R}$. Without loss of generality, we consider a binary partition as an example. Let $\{\Gamma_k : k = 1, 2, \ldots\}$ denote nested partitions of $\mathfrak{X}$. For example,

$$\begin{aligned}
\Gamma_1 &= (B_0, B_1); B_0 = B_{00} \cup B_{01}; B_1 = B_{10} \cup B_{11}, \\
\Gamma_2 &= (B_{00}, B_{01}, B_{10}, B_{11}); B_{ij} = B_{ij0} \cup B_{ij1},\ i = 0, 1,\ j = 0, 1, \\
\Gamma_3 &= (B_{000}, B_{001}, B_{010}, B_{011}, B_{100}, B_{101}, B_{110}, B_{111}).
\end{aligned}$$

In other words, every Γ_{k+1} is a refinement of its previous partition Γ_k satisfying the relationship $B_{\boldsymbol{s}} = B_{\boldsymbol{s}0} \cup B_{\boldsymbol{s}1}$, where $\boldsymbol{s}$ is some $\{0, 1\}$ sequence. By defining random variables $V_{\boldsymbol{s}}$ as the probability on a set $B_{\boldsymbol{s}}$ and assume that variables defined in different Γ_k are independent to each other, we may obtain a tail-free random probability measure defined on $\mathfrak{X}$. Define Γ as the collection of $\Gamma_1, \ldots, \Gamma_k$. A formal definition is as follows:

Definition 3. *A random probability measure P on $\mathfrak{X}$ is tail-free with respect to $(\Gamma, \mathfrak{X})$ if there exist nonnegative random variables $\{V_{\boldsymbol{s}} : \boldsymbol{s} \in \{0, 1\}^k, k \geq 1\}$ defined on $B_{\boldsymbol{s}}$ such that*

(i) Classes of random variables $(V_{\boldsymbol{s}} : \boldsymbol{s} \in \{0, 1\}^k)$ are independent to other classes with

different values of k.

(ii) $P(B_{\boldsymbol{s}}) = \prod_{i=1}^{k} V_{s_1 \cdots s_i}$ *for every* $\boldsymbol{s} \in \{0,1\}^k$, $k \geq 1$.

Obviously, this definition can be extended beyond binary partitions. It is also possible to consider varying number of subsets in different partition levels. One of the nicest properties of tail-free process prior is its conjugacy structure. Here we list a few properties, their proofs can be found in Freedman (1963, 1965).

- (Conjugacy) Given i.i.d observations $X_1, \ldots, X_n | P \sim P$, in which P is a tail-free prior with respect to $(\Gamma, \mathfrak{X})$, then the posterior is another tail-free process with respect to the same partition $(\Gamma, \mathfrak{X})$.

- (Continuous or discrete) Given a tail-free process P with respect to $(\Gamma, \mathfrak{X})$. Suppose there is a measure λ such that $\lambda(B) > 0$ for every $B \in \mathcal{H}$. Then P is absolutely continuous with respect to λ with probability zero or one. As a result, P is discrete with probability zero or one.

- (Transformation) The tail-free property is preserved under monotonic transformations.

- (Mixture) The tail-free property is not preserved under mixture operation.

Besides Dirichlet process, another popular tail-free process is called Polya tree process Blackwell and MacQueen (1973); Ferguson (1974). A Polya tree process is defined through binary partitions as follows:

Definition 4. *A random probability measure* P *is a Polya tree process with respect to* $(\Gamma, \mathfrak{X})$ *and parameters* $\boldsymbol{\alpha} = \{\alpha_{\boldsymbol{s}}, \boldsymbol{s} \in \{0,1\}^k, k \geq 1\}$ *if for every* $\boldsymbol{s} \in \{0,1\}^k$ *and* $k \geq 1$, $V_{s0} = P(B_{\boldsymbol{s}0} | B_{\boldsymbol{s}})$ *are mutually independent* $\mathrm{Beta}(\alpha_{\boldsymbol{s}0}, \alpha_{\boldsymbol{s}1})$. *We denote this Polya tree process by* $P \sim \mathrm{PT}(\Gamma, \boldsymbol{\alpha})$.

Unlike the Dirichlet process, a Polya tree process gives rise to continuous distributions. It has been used in nonparametric Bayesian density estimation (Lavine, 1992, 1994). The Polya tree family has a conjugacy property that makes the updating of the posterior distribution simple.

Theorem 1. *Given i.i.d observations* $X_1, \ldots, X_n | P \sim P$ *and* $P \sim \mathrm{PT}(\Gamma, \boldsymbol{\alpha})$*, the posterior process is another Polya tree* $\mathrm{PT}(\Gamma, \boldsymbol{\alpha}^*)$*, where* $\alpha^*_{\boldsymbol{s}} = \alpha_{\boldsymbol{s}} + \sum_{i=1}^{n} \mathbb{1}(X_i \in B_{\boldsymbol{s}})$ *for every* $\boldsymbol{s} \in \{0,1\}^k$*,* $k \geq 1$*.*

However, one limitation of a Polya tree process prior is its dependence on the partition sets. Also, there might be discontinuity issues at the boundary of partition sets, which is not desired in situations such as density estimation. To address this issue, mixtures of Polya trees have been considered in Hanson and Johnson (2002).

1.2.5 Basis expansion

In function estimation problems, it is common to expand the unknown function f by pre-specified basis functions $\{\xi_j\}$: $f = \sum_{j=1}^{J} \theta_j \xi_j$ or through some link function: $f = \Psi(\sum_{j=1}^{J} \theta_j \xi_j)$. Then by assigning prior distributions on J and relevant coefficient θ_j's, we obtain a prior on the function f. Some commonly used bases include polynomials, splines and wavelets. In some situations, in order to induce a valid prior on the function of interest, it may be necessary to consider a specific link function or to put restrictions on the coefficients θ_j. For example, for density estimation problems, in order to make f nonnegative and integrate to one, we may choose splines as basis, restrict all coefficients to be nonnegative and then rescale it as follows

$$f(x) = \frac{\sum_{j=J}^{\infty} \theta_j \xi_j(x)}{\int \sum_{j=J}^{\infty} \theta_j \xi_j(u) du}.$$

This idea applies for some functions with complex forms as well. For example, in Chapter 3.9, we consider a functional regression model where each coefficient is now a functional of time. In Chapter 4, the function we are interested in is a conditional density under a high-dimensional setting.

1.3 Asymptotics

In this section, we discuss the asymptotic properties of Bayesian procedures in a frequentist sense. Denote f as the function of interest. Intuitively, one would expect the posterior concentrate around the true function f_0 if there are infinitely many data available. This property is called posterior consistency and is mathematically defined as

$$\Pi(d(f, f_0) > \epsilon | X_1, \ldots, X_n) \to 0, \text{in } P_{f_0}^n\text{-probability}. \tag{1.4}$$

when the sample size n goes to infinity for some distance d and any predetermined $\epsilon > 0$. For most parametric models, posterior consistency holds when the true parameter is in the support of the prior (Le Cam, 1986). However, for infinite-dimensional problems, such simple results no longer hold. An inconsistency example was given by Freedman (1963). He considered estimating a discrete distribution on natural numbers. While the true distribution is set as geometric distribution with parameter $1/4$, Freedman constructed a prior with positive probability around every neighborhood of the true distribution that leads to a posterior, which surprisingly concentrates near a geometric distribution with parameter $3/4$. Another interesting example was introduced by Kim and Lee (2001) in survival analysis context. They compare two priors with the same mean function equals to the true hazard function. These counterexamples suggest that more conditions are needed.

If the posterior distribution is available in an explicit form, then posterior consistency may be obtained by applying certain moment inequalities such as Chebyshev's inequality. This is the case in some Bayesian survival analysis problems and other places where conjugacy structure is available. However, in most applications this approach is not feasible. A more useful result was established by Schwartz (1965) for dominated families. The theory gives two sufficient conditions for consistency. First, with f_0 being the true function, there should exist a strictly unbiased test for $f = f_0$ against complement of the neighborhood of f_0. This ensures that the type I and type II error probabilities are going to zero exponentially fast. Secondly, the prior should have a positive probability in any neighborhood of the true function in a Kullback-Leibler (KL) divergence sense. Here the KL divergence between two probability densities p and q (with respect to a dominating measure ν) is defined as $K(p, q) = \int p \log(p/q) d\nu$. The second condition is crucial and sometimes is referred to as the Kullback-Leibler property. The KL divergence can be stronger than the natural distance on the parametric space. Unfortunately, Schwartz's testing condition does not hold for many nonparametric problems for stronger topologies unless the parameter space is compact, which is a very strict requirement. A refined result was proposed by Ghosal et al. (1999) using a technique involving sieves. A sieve is defined as a big subset of the parameter space. They showed that by carefully choosing a sieve and then applying Schwartz's techniques on the sieve, posterior consistency can be obtained by bounding the prior probability of the complement of the sieve and the entropy number of the sieve. This idea was initially used by Barron et al. (1999) under stronger conditions using bracketing entropy.

Besides consistency, convergence rate is another issue of extreme importance in asymptotics. It measures the speed of the convergence of posterior f given the data $X_1, \ldots, X_n$ to the true value f_0 under some metric d. A sequence $\epsilon_n \to 0$ is called a rate of convergence if

$$\Pi(d(f, f_0) \geq m_n \epsilon_n | X_1, \ldots, X_n) \to 0 \text{ in } P_{f_0}^n\text{-probability} \tag{1.5}$$

for every sequence $m_n \to \infty$.

If the posterior distribution converges at a rate ϵ_n, then a point estimator can be constructed based on posterior distribution that converges at least as fast as ϵ_n in the frequentist sense (Ghosal et al., 2000). Hence the posterior convergence rate can never be faster than the minimax rate. Achieving minimax rate is the best result one can expect. In many cases, the posterior convergence rate agrees with the minimax rate, sometimes up to a logarithmic factor. A general result on posterior convergence rate is given by Ghosal et al. (2000). Shen and Wasserman (2001) obtained a rate of convergence theorem using a stronger condition to bound entropy integral. Their idea is as follows: Suppose we have i.i.d observations $X_1, \ldots, X_n \sim P_{f_0}$. Define a neighborhood of f_0 as $B = \{f : d(P_f, P_{f_0}) \geq \epsilon_n\}$. Then by showing that the numerator is upper bounded by $e^{-an\epsilon_n^2}$ for some $a > 0$ and the denominator is lower bounded by $e^{-bn\epsilon_n^2}$ for all $b > 0$ in the following expression, they obtain (1.5):

$$\Pi(\theta \in B|X_1, \ldots, X_n) = \frac{\int_B p_f(X_1, \ldots, X_n)/p_{f_0}(X_1, \ldots, X_n)d\Pi(f)}{\int p_f(X_1, \ldots, X_n)/p_{f_0}(X_1, \ldots, X_n)d\Pi(f)}, \tag{1.6}$$

where p_f is defined as the density function of P_f. Compared with the consistency result, stronger conditions are needed to control both numerators and denominators. In particular, the Kullback-Leibler property is replaced by

$$\Pi\{\mathcal{K}(f_0, \epsilon_n)\} \geq e^{-cn\epsilon_n^2} \tag{1.7}$$

for some constant $c > 0$, where $\mathcal{K}(p, \epsilon) = \{f : K(p, f) \leq \epsilon^2, V(p, f) \leq \epsilon^2\}$ and $V(p, q) = \int p(\log p/q)^2$. This condition requires the prior distribution to have sufficient contraction rate around the true point mass measure P_{f_0} in terms of the first and the second moment of the log-likelihood. Such an ϵ_n is called the prior concentration rate. Now, for the numerator, one needs

to find a test that differentiate P_f and P_{f_0} whose type I and type II errors are exponentially small. Such a test is constructed by covering the sieve using balls of size $\epsilon_n/2$ and control the total number of such small balls, see Figure 1.1. The precise requirement can be described by the following entropy condition:

$$\log N(\tilde{\epsilon}_n/2, \mathcal{P}_n, d) \lesssim n\tilde{\epsilon}_n^2 \text{ for some } \tilde{\epsilon}_n > 0. \tag{1.8}$$

The posterior convergence rate is determined by the maximum of prior concentration rate ϵ_n and $\tilde{\epsilon}_n$. Usually these two rates need to be matched to obtain a minimax rate.

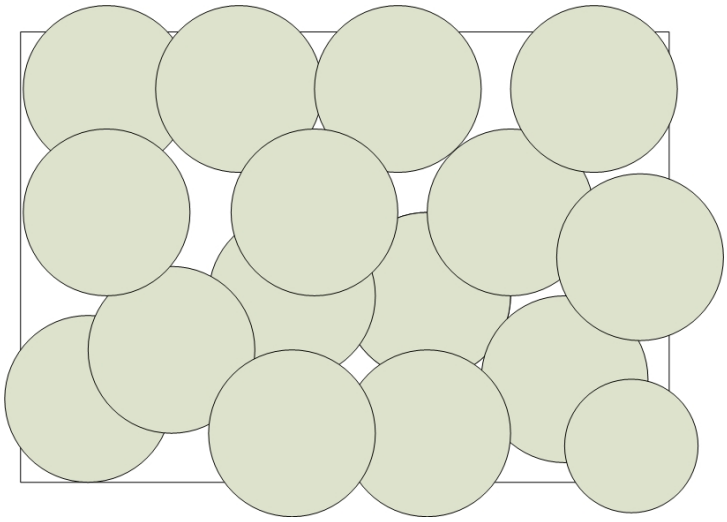


Figure 1.1: Covering balls of a sieve

In conclusion, we summarized a set of sufficient conditions in showing posterior convergence rates of function estimation in the following theorem, which is slightly modified from Theorem 2.1 in Ghosal et al. (2000) and Theorem 4 in Ghosal and van der Vaart (2007a).

Theorem 2. *Let ϵ_n and $\tilde{\epsilon}_n$ be two sequences of positive numbers satisfying $\tilde{\epsilon}_n \leq \epsilon_n$ and*

$\lim_{n\to\infty} n\tilde{\epsilon}_n^2 = \infty$. *Given* n *independent identically distributed (i.i.d.) observations* $X_1, \ldots, X_n$ *from a unknown density function* f_0 *and some prior distribution* Π *on* f*, suppose that there exists a sequence of sieves* $\mathcal{F}_n$ *defined on the space of probability densities, some distance metric* d *and some positive constants* c_1, c_2, c_3, c_4 *such that*

$$\log N(\epsilon_n, \mathcal{F}_n, d) \leq c_1 n\epsilon_n^2, \tag{1.9}$$

$$\Pi(\mathcal{F}_n^c) \leq c_3 e^{-(c_2+4)n\tilde{\epsilon}_n^2}, \tag{1.10}$$

$$\Pi\left\{\mathcal{K}(f_0, \tilde{\epsilon}_n)\right\} \geq c_4 e^{-c_2 n\tilde{\epsilon}_n^2}. \tag{1.11}$$

Then the posterior of f *converges around the true function* f_0 *in a rate* ϵ_n *with respect to* d*.*

1.4 Adaptation

In smoothing problems, it is well known that the optimal convergence rate for the estimation of functions is determined by their smoothness levels. For example, the optimal rate of estimating a univariate α-smooth function is $n^{-\alpha/(2\alpha+1)}$ (Hasminskii, 1978), where n is the sample size. Generally, the prior has to be customized to achieve optimal rate given the knowledge of the smoothness parameter α. However, α is usually unknown in practice. Therefore, it is of interest to investigate if one single choice of prior can lead to posterior distributions that has optimal posterior convergence rates simultaneously for all values of α. Such procedures are called rate-adaptive.

In the literature, the first idea of adaptation is to consider assigning a prior distribution on α and to study the resulting mixture of the posterior. Belitser and Ghosal (2003) considered the problem of estimating a signal with Gaussian white noise and showed that the posterior rate automatically adapts to the unknown smoothness condition if the smoothness parameter

only takes values in a discrete set. Ghosal et al. (2003, 2008) showed that appropriate mixture of certain priors, such as those based on spline expansions, yield optimal posterior rates for a countable range of smoothness parameters for density estimation. A similar work has been done in a nonparametric regression setting using wavelets in Huang (2004). Scricciolo (2006) obtained adaptive rates for density estimation problems when the smoothness parameter belongs to a discrete set. The basic idea behind this approach is to use the optimal dimension $J_{n,\alpha}$ of the model for a given smoothness level α and sample size n obtained from some appropriate rate equation to construct "optimal priors" $\Pi_{n,\alpha}$ for each α, and then mix countably many of them to construct the mixture prior which adapts for all these countably many smoothness levels. Alternatively, van der Vaart and van Zanten (2009) constructed a prior based on a randomly rescaled smooth Gaussian process, which automatically adapts for a continuous range of smoothness parameters.

In this thesis, we prove Bayesian adaptation results following two steps. The first is to construct a sieve that contains the true underlying model while its size is well controlled. The size is usually quantified by the existence of certain tests or in terms of entropy bounds (Ghosal et al., 2000). The other step is to construct an approximation of the true function while its approximation accuracy increases appropriately with the increasing level of smoothness.

1.5 Notations

Throughout the thesis, we use $\mathbb{N} = \{1, 2, \ldots\}$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ and $\Delta_j = \{(x_1, \ldots, x_j) : \sum_{i=1}^{j} x_i = 1, x_1, \ldots, x_j \geq 0\}$. For any $x \in \mathbb{R}$, define $\lfloor x \rfloor$ as the largest integer that is smaller than x. Similarly, define $\lceil x \rceil$ as the smallest integer greater than x. We define $\log_+ x = \max(\log x, 0)$ and $\log_- x = \max(-\log x, 0)$. We define δ_x as the point mass probability distribution at point x. Define an indicator function of a set A by $\mathbb{1}(A)$.

We use $\|\boldsymbol{x}\|_p = \{\sum_{i=1}^d |x_i|^p\}^{1/p}$ for the ℓ_p-norm of a vector $\boldsymbol{x} \in \mathbb{R}^d$; $1 \le p < \infty$ and $\|\boldsymbol{x}\|_\infty = \max_{1\le i\le d} |x_i|$. Moreover, for $p = 2$, we simply write $\|\boldsymbol{x}\|_2$ as $\|\boldsymbol{x}\|$. Define (equivalent classes of) function spaces $L_p = \{f : \|f\|_p < \infty\}$. For a probability measure G, define $\|f\|_{p,G} = \{\int |f(x)|^p dG(x)\}^{1/p}$.

For any $\boldsymbol{a} = (a_1, \ldots, a_d)$ and $\boldsymbol{b} = (b_1, \ldots, b_d)$, define $\boldsymbol{a}^{\boldsymbol{b}} = \prod_{i=1}^d a_i^{b_i}$. Let $\langle a, b\rangle$ denote $a_1b_1 + \cdots + a_db_d$. For a multi-index $\boldsymbol{k} = (k_1, \ldots, k_d) \in \mathbb{N}_0^d$, define $\boldsymbol{k}. = k_1 + \cdots + k_d$, $\boldsymbol{k}! = k_1! \cdots k_d!$ and let $D^{\boldsymbol{k}}$ denote the mixed partial derivative operator $\partial^{\boldsymbol{k}.}/\partial x_1^{k_1} \cdots \partial x_d^{k_d}$.

For any $d \times d$ matrix $A = \big((a_{ij})\big)$, we denote its eigenvalues by $\mathrm{eig}_1(A), \ldots, \mathrm{eig}_d(A)$, its spectral norm by $\|A\|_2 := \max_{x\neq 0} \|Ax\|/\|x\|$ and its max norm by $\|A\|_{\max} = \max_{i,j} |a_{ij}|$.

For any $d \times d$ positive definite real matrix $\boldsymbol{\Sigma}$, let $\phi_{\boldsymbol{\Sigma}}(x)$ denote the d-variate normal density with mean zero and covariance matrix $\boldsymbol{\Sigma}$:

$$\phi_{\boldsymbol{\Sigma}}(x) = (2\pi)^{-d/2} (\det \boldsymbol{\Sigma})^{-1/2} \exp(-\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}/2).$$

For a probability measure F on $\mathbb{R}^d$ and a $d \times d$ positive definite real matrix Σ, the F induced location mixture of $\phi_{\boldsymbol{\Sigma}}$ is denoted $p_{F,\boldsymbol{\Sigma}}$, i.e., $p_{F,\boldsymbol{\Sigma}}(\boldsymbol{x}) = \int \phi_{\boldsymbol{\Sigma}}(\boldsymbol{x} - \boldsymbol{z}) F(d\boldsymbol{z})$, $x \in \mathbb{R}^d$. For a scalar $\sigma > 0$ and any function f on $\mathbb{R}^d$, we denote by $K_\sigma f$ the convolution of f and $\phi_{\sigma^2 \boldsymbol{I}}$, i.e., $(K_\sigma f)(\boldsymbol{x}) = \int \phi_{\sigma^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{z}) f(\boldsymbol{z}) d\boldsymbol{z}$.

For a set T, we denote its cardinality by $|T|$. We say T^* is an ϵ-dispersed subset of T with respect to some metric d if $T^* \subset T$ and $d(t, t') \ge \epsilon$ for all $t \neq t'$, $t, t' \in T^*$. Similarly, we say T^* is an ϵ-net of T with respect to d if for every $t \in T$, there exists a $t' \in T^*$ such that $d(t, t') \le \epsilon$. Then we can define the packing number of T as

$$D(\epsilon, T, d) = \max\{|T^*| : \ T^* \text{ is an } \epsilon\text{-dispersed subset of } \ T \text{ with respect to } d\}$$

and the covering number of T as

$$N(\epsilon, T, d) = \min\{|T^*| : T^* \text{ is an } \epsilon\text{-net of } T \text{ with respect to } d\}.$$

For any $\beta, d > 0$, define a β-Hölder class on a set $\Omega \subset \mathbb{R}^d$, denoted $\mathcal{C}^\beta(\Omega)$, to be the set of all functions $f : \Omega \to \mathbb{R}$ with finite mixed partial derivatives $D^{\boldsymbol{k}} f$, $\boldsymbol{k} \in \mathbb{N}_0^d$, of all orders up to $\boldsymbol{k}_\cdot \leq \beta_0$, where β_0 is the largest integer strictly smaller than β, and for every $\boldsymbol{k} \in \mathbb{N}_0^d$ with $\boldsymbol{k}_\cdot = \beta_0$ satisfying

$$|(D^{\boldsymbol{k}} f)(\boldsymbol{x} + \boldsymbol{y}) - (D^{\boldsymbol{k}} f)(\boldsymbol{x})| \leq C\|\boldsymbol{y}\|^{\beta - \beta_0}, \quad \boldsymbol{x}, \boldsymbol{y} \in \Omega \tag{1.12}$$

for some positive constant C.

We use $\lesssim$ for inequality up to a constant multiple, where the underlying constant of proportionality is universal or not important for our purposes. If two functions f and g satisfy $f \lesssim g \lesssim f$, we shall write $f \asymp g$.

For a density function $f_0(\boldsymbol{x})$ of a probability measure P_0, define $P_0 g = \int g(\boldsymbol{x}) f_0(\boldsymbol{x}) d\boldsymbol{x}$ as the expectation of $g(X)$ under $X \sim f_0$.

The Hellinger distance $h(p, q)$ and the Kullback-Leibler (KL) divergence $K(p, q)$ between two densities p and q are commonly used in statistics. They are respectively defined by $h^2(p, q) = \int (\sqrt{p} - \sqrt{q})^2 d\mu$ and $K(p, q) = \int p \log(p/q) d\mu$. Also define the second order KL divergence by $V(p, q) = \int p \log^2(p/q) d\mu$. We define a KL ball around p with radius ϵ as $\mathcal{K}(p, \epsilon) = \{f : K(p, f) \leq \epsilon^2, V(p, f) \leq \epsilon^2\}$.

1.6 Outline

The rest of the thesis is organized as follows. In Chapter 2, we study a multivariate kernel density estimation problem by using a Dirichlet mixture prior. We obtain adaptive posterior convergence rate when the true density belongs to a Hölder smoothness class or an anisotropic. In Chapter 3, we focus on estimating univariate functions using random series prior for many nonparametric problems. We derive a general result on constructing an appropriate sieve that can later be used to obtain adaptive rates. Then we apply this general result to B-spline expansions and treated a variety of statistical problems such as density estimation, nonparametric regression etc. We also develop a computationally efficient algorithm that allows us to obtain posterior moments without using MCMC methods. We extend this work to a density regression problem under high-dimensional setting, where the number of covariates is possibly much larger than the sample size in Chapter 4. A variable selection prior is discussed and the adaptive posterior convergence rate is obtained under mild sparsity conditions. In Chapter 5, we focus on estimating derivatives of density functions using kernel methods and random series priors. New testing inequalities are developed to obtain posterior rates.

Chapter 2

Multivariate density estimation using Dirichlet mixture of normal prior

2.1 Introduction

In this chapter we focus on a collection of nonparametric Bayesian density estimation methods based on Dirichlet process mixture of multivariate normals priors. Dirichlet process mixture priors (Ferguson, 1983; Lo, 1984) form a cornerstone of nonparametric Bayesian methodology (Escobar and West, 1995; Müller et al., 1996; Müller and Quintana, 2004; Dunson, 2010) and density estimation methods based on these priors are among the first Bayesian nonparametric methods for which consistency (Ghosal et al., 1999; Tokdar, 2006; Wu and Ghosal, 2010) and convergence rate results were obtained (Ghosal and van der Vaart, 2001, 2007b). The rates obtained in Ghosal and van der Vaart (2001, 2007b) requires the knowledge of the smoothness levels of the underlying density, and hence the procedures were not shown to be rate adaptive. In particular, Ghosal and van der Vaart (2001) treated the case of supersmooth normal mixtures while Ghosal and van der Vaart (2007b) dealt with the case of twice continuously dif-

ferentiable density. Two major challenges lie in showing that Dirichlet process mixtures give rate adaptive procedures. The first difficulty is in obtaining adaptive prior concentration rates for mixture priors on density functions. Usually this first step is to obtain an approximation of the true density within the model such that the order of approximation improves with the smoothness level of the true density. However the obvious choice given by the convolution of the true density with the kernel does not suffice because the approximation rate does not improve with increasing smoothness of the true density. The second difficulty is to construct a suitable low-entropy, high-mass sieve on the space of infinite component mixture densities. Such sieve constructions are an integral part of current technical machinery for deriving rates of convergence. The sieves that have been used to study Dirichlet process mixture models (e.g., in Ghosal and van der Vaart, 2007b) do not scale to higher dimensions and lack the ability to adapt to smoothness classes (Wu and Ghosal, 2010). The only paper dealing with rate adaptation for nonparametric mixture of normal prior is Kruijer et al. (2010), where they established rate adaptation for finite normal mixtures in the univariate situation using a technique that was first introduced by Rousseau (2010) in the context of beta mixtures.

We bridge these two gaps and establish rate adaptation properties of a collection of multivariate density estimation methods based on Dirichlet process mixture of normals priors. Our priors include the commonly used specification of mixing over multivariate normal kernels with a location parameter drawn from a Dirichlet process with a Gaussian base measure while using an inverse-Wishart prior on the common covariance matrix parameter of the kernels. Rate adaptation is established with respect to Hölder smoothness classes. In particular, when any density estimation method from our collection is applied to independent observations $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n} \in \mathbb{R}^d$ drawn from a density f_0 which belongs to the smoothness class of locally β-Hölder functions, it is shown to produce a posterior distribution on the unknown density of $\boldsymbol{X_i}$'s that converges to f_0 at a rate $n^{-\beta/(2\beta+d)}(\log n)^t$, where t depends on β, d and tail

properties of f_0. This rate, without the $(\log n)^t$ term, is minimax optimal for the β-Hölder class (Barron et al., 1999). It is further shown that if f_0 is anisotropic with Hölder smoothness coefficients $\beta_1, \ldots, \beta_d$ along the d axes, then the posterior convergence rate is $n^{-\beta_0/(2\beta_0+d)}$ times a $\log n$ factor, where β_0 is the harmonic mean of $\beta_1, \ldots, \beta_d$. Again this rate is minimax optimal for this class of functions (Hoffmann and Lepski, 2002).

To the best of our knowledge, such rate adaptation results are new for any kernel based multivariate density estimation method. The performance of a non-Bayesian, multivariate kernel density estimator depends heavily on the difficult choice of a bandwidth and a smoothing kernel (Scott, 1992). Optimal rates are possible only by using higher order kernels and the choices of bandwidth that require knowing the smoothness level. In contrast our results show that a single Bayesian nonparametric method based on a single choice of Dirichlet process mixture of normal kernels achieves optimal convergence rates universally across all smoothness levels.

2.2 Posterior Convergence Rates for Dirichlet Mixtures

2.2.1 Dirichlet process mixture of multivariate normals prior

Consider drawing inference on an unknown probability density function f on $\mathbb{R}^d$ based on independent observations $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ from f. A nonparametric Bayesian method assigns a prior distribution Π on f and draws inference on f based on the posterior distribution

$$\Pi_n(f \in B \mid \boldsymbol{X_1}, \ldots, \boldsymbol{X_n}) = \frac{\int_B \prod_{i=1}^n f(\boldsymbol{X_i}) d\Pi(f)}{\int \prod_{i=1}^n f(\boldsymbol{X_i}) d\Pi(f)}$$

A Dirichlet process location mixture of normals prior Π is the distribution of a random probability density function $p_{F,\boldsymbol{\Sigma}}$, where $F \sim \mathrm{DP}(\alpha)$ for some finite positive measure α on $\mathbb{R}^d$ and $\boldsymbol{\Sigma} \sim G$, a probability distribution on $d \times d$ positive definite real matrices.

We restrict our discussion to a collection of such prior distributions Π for which the associated $\mathrm{DP}(\alpha)$ and G satisfy the following conditions. Let $|\alpha| = \alpha(\mathbb{R}^d)$ and $\bar{\alpha} = \alpha/|\alpha|$. We assume that $\bar{\alpha}$ has a positive density function on the whole of $\mathbb{R}^d$ and that there exist positive constants $a_1, a_2, a_3, b_1, b_2, b_3, C_1, C_2$ such that

$$1 - \bar{\alpha}([-x, x]^d) \leq b_1 \exp(-C_1 x^{a_1}) \text{ for all sufficiently large } x > 0, \tag{2.1}$$

$$G\{\boldsymbol{\Sigma} : \mathrm{eig}_d(\boldsymbol{\Sigma}^{-1}) \geq x\} \leq b_2 \exp(-C_2 x^{a_2}) \text{ for all sufficiently large } x > 0, \tag{2.2}$$

$$G\{\boldsymbol{\Sigma} : \mathrm{eig}_1(\boldsymbol{\Sigma}^{-1}) < x\} \leq b_3 x^{a_3} \text{ for all sufficiently small } x > 0, \tag{2.3}$$

and that there exist $\kappa, a_4, a_5, b_4, C_3 > 0$ such that for any $0 < s_1 \leq \cdots \leq s_d$ and $t \in (0, 1)$,

$$G\{\boldsymbol{\Sigma} : s_j < \mathrm{eig}_j(\boldsymbol{\Sigma}^{-1}) < s_j(1+t), j = 1, \ldots, d\} \geq b_4 s_1^{a_4} t^{a_5} \exp(-C_3 s_d^{\kappa/2}). \tag{2.4}$$

Our assumption on $\bar{\alpha}$ is analogous to (11) of Kruijer et al. (2010) and holds, for example, when $\bar{\alpha}$ is a Gaussian measure on $\mathbb{R}^d$. Unlike previous treatments of Dirichlet process mixture models (Ghosal and van der Vaart, 2001, 2007b), we allow a fully supported prior on $\boldsymbol{\Sigma}$ including the widely used inverse-Wishart distribution. The following lemma shows that this prior satisfies our assumptions.

Lemma 1. *The inverse-Wishart distribution* $\mathrm{IW}(\nu, \boldsymbol{\Psi})$ *with* ν *degrees of freedom and a scaler matrix* $\boldsymbol{\Psi} = c\boldsymbol{I}$ $(c > 0)$ *satisfies* (2.2)*,* (2.3) *and* (2.4) *with* $\kappa = 2$*.*

Proof. Let $\boldsymbol{\Sigma} \sim \mathrm{IW}(\nu, \boldsymbol{\Psi})$ and suppose $\boldsymbol{\Psi} = \boldsymbol{I}$, i.e., $c = 1$. It is well known that $\mathrm{tr}(\boldsymbol{\Sigma}^{-1}) \sim \chi^2_{\nu d}$, the chi-square distribution with νd degrees of freedom. The cumulative distribution function $F(x; k)$ of χ^2_k satisfies $1 - F(zk; k) \leq \{z \exp(1 - z)\}^{k/2}$ for all $z > 1$. Therefore for all

$x > \nu d$,

$$\mathrm{P}\left\{\mathrm{eig}_d(\mathbf{\Sigma}^{-1}) > x\right\} \leq P\left\{\mathrm{tr}(\mathbf{\Sigma}^{-1}) > x\right\} \leq \left(\frac{x}{\nu d}\right)^{\nu d/2} \exp\{(\nu d - x)/2\} \leq b_2 e^{-C_2 x}$$

for some constants b_2, C_2. Furthermore, the joint probability density of $\mathrm{eig}_1(\mathbf{\Sigma}^{-1}), \ldots, \mathrm{eig}_d(\mathbf{\Sigma}^{-1})$ is

$$f(x_1, \ldots, x_d) = c_{d,\nu} \ \exp(-\sum_{j=1}^{d} x_j/2) \prod_{j=1}^{d} x_j^{(\nu+1-d)/2} \prod_{j<k} (x_k - x_j) \tag{2.5}$$

over the set $\{(x_1, \ldots, x_d) \in (0, \infty)^d : x_1 \leq \cdots \leq x_d\}$, for a known constant $c_{d,\nu}$. Since $\prod_{j<k}(x_k - x_j) \leq \prod_{j<k} x_k = \prod_{k=2}^{d} x_k^{k-1}$, the probability density of $\mathrm{eig}_1(\mathbf{\Sigma}^{-1})$ satisfies

$$f(x_1) \leq c_{d,\nu} x_1^{(\nu+1-d)/2} e^{-x_1/2} \prod_{k=2}^{d} \left\{ \int_0^\infty x_k^{(\nu+1-d)/2+k-1} e^{-x_k/2} dx_k \right\} = \tilde{c}_{d,\nu} x_1^{(\nu+1-d)/2} e^{-x_1/2}$$

for all $x_1 > 0$ and some positive constant $\tilde{c}_{d,\nu}$. Therefore for any $x > 0$

$$\mathrm{P}\left\{\mathrm{eig}_1(\mathbf{\Sigma}^{-1}) < x\right\} \leq \tilde{c}_{d,r} \int_0^x x_1^{(\nu+1-d)/2} dx_1 \leq b_3 x^{a_3}$$

for some positive constants a_3, b_3.

Next, notice that the set on the left hand side of (2.4) contains all $\mathbf{\Sigma}$ which have $\mathrm{eig}_j(\mathbf{\Sigma}^{-1}) \in I_j = (s_j\{1 + (j - 1/2)t/d\}, s_j(1 + jt/d))$ $(j = 1, \ldots, d)$ and for any positive integers $k > j$, $x_j \in I_j$ and $x_k \in I_k$ implies that $x_k - x_j > s_k\{1 + (k - 1/2)t/d\} - s_j(1 + jt/d) \geq s_1 t/(2d)$.

Therefore

$$
\begin{aligned}
&\mathrm{P}\{s_j < \mathrm{eig}_j(\boldsymbol{\Sigma}^{-1}) < s_j(1+t),\ \ j = 1, \ldots, d\} \\
&\quad \geq \int_{I_d} \cdots \int_{I_1} c_{d,\nu}\ \exp\left(-\sum_j x_j/2\right) \prod_{j=1}^{d} x_j^{(\nu+1-d)/2} \prod_{j<k} (x_k - x_j) dx_1 \cdots dx_d \\
&\quad \geq c_{d,\nu} \exp\left(-d s_d\right) s_1^{d(\nu+1-d)/2} \{t/(2d)\}^{d(d-1)/2} \int_{I_d} \cdots \int_{I_1} dx_1 \cdots dx_d \\
&\quad = c_{d,\nu} \exp\left(-d s_d\right) s_1^{d(\nu+1-d)/2} \{t/(2d)\}^{d(d-1)/2} \{s_1 t/(2d)\}^{d},
\end{aligned}
$$

which gives (2.4) for some positive constants a_4, a_5, b_4, C_3.

If $c \neq 1$, the eigenvalues of $\boldsymbol{\Sigma}^{-1}$ scale by a factor c, and hence the same conclusion holds for a different set of constants. □

From a computational point of view, another useful specification is to consider a G that supports only diagonal covariance matrices $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, with each diagonal component independently assigned a prior distribution G_0. By choosing an inverse gamma distribution as G_0, we get a prior G on $\boldsymbol{\Sigma}$ that again satisfies (2.2), (2.3) and (2.4) with $\kappa = 2$. Alternatively, we can take G_0 to be the distribution of the square of an inverse gamma random variable. Such a G_0 leads to a G that satisfies (2.2), (2.3) and (2.4) with $\kappa = 1$. This difference in κ matters, with smaller κ leading to optimal convergence rates for a wider class of true densities.

2.2.2 Locally Hölder classes

Here we introduce a locally Hölder class, which is slightly weaker than the Hölder class definition in Section 1.5. We will use this definition throughout the chapter.

For any $\beta > 0$, $\tau_0 \geq 0$ and any non-negative function L on $\mathbb{R}^d$, define the locally β-Hölder class with envelope L, denoted $\mathcal{C}^{\beta, L, \tau_0}(\mathbb{R}^d)$, to be the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ with finite

mixed partial derivatives $D^{\boldsymbol{k}} f$, $\boldsymbol{k} \in \mathbb{N}_0^d$, of all orders up to $\boldsymbol{k}_\cdot \leq \lfloor \beta \rfloor$, and for every $\boldsymbol{k} \in \mathbb{N}_0^d$ with $\boldsymbol{k}_\cdot = \lfloor \beta \rfloor$ satisfying

$$|(D^{\boldsymbol{k}} f)(\boldsymbol{x}+\boldsymbol{y}) - (D^{\boldsymbol{k}} f)(\boldsymbol{x})| \leq L(\boldsymbol{x}) e^{\tau_0 \|\boldsymbol{y}\|^2} \|\boldsymbol{y}\|^{\beta - \lfloor \beta \rfloor}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d. \tag{2.6}$$

In our discussion, we shall assume that the true density f lies in $\mathcal{C}^{\beta, L, \tau_0}(\mathbb{R}^d)$. This condition is essentially weaker than the one in Kruijer et al. (2010), where $\log f \in \mathcal{C}^{\beta, L, 0}(\mathbb{R})$ is assumed, see Lemma 2.

Lemma 2. *Suppose a probability density function f_0 satisfies the tail condition* (2.9)*, $\log f_0 \in \mathcal{C}^{\beta, Q_1, 0}(\mathbb{R}^d)$ for some polynomial Q_1 with $P_0 |D^k \log f_0|^{(2\beta+\epsilon)/\boldsymbol{k}_\cdot} < \infty$, $\boldsymbol{k} \in \mathbb{N}_0^d$, $\boldsymbol{k}_\cdot \leq \lfloor \beta \rfloor$ and $P_0 Q_1^{(2\beta+\epsilon)/\beta} < \infty$. Additionally, suppose that*

$$\left| \frac{f_0(\boldsymbol{x}+\boldsymbol{y})}{f_0(\boldsymbol{x})} - 1 \right| \leq Q(\boldsymbol{x}) e^{\tau_0 \|\boldsymbol{y}\|^2} \|\boldsymbol{y}\|^{\beta - \lfloor \beta \rfloor}, \text{ for any } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d, \tag{2.7}$$

for some $\tau_1 > 0$ and a function Q satisfying $P_0 Q^2 < \infty$. Then, there exist a $\tau_0 > 0$ and a positive functions $L(\boldsymbol{x})$ such that $f_0 \in \mathcal{C}^{\beta, L, \tau_0}(\mathbb{R}^d)$ and (2.8) *holds.*

Proof. For a multi-index $\boldsymbol{k} \in \mathbb{N}_0^d$, let $\mathcal{P}$ denote the set of all solutions $\{\boldsymbol{m}^{(1)}, \ldots, \boldsymbol{m}^{(q)}\}$ to $\boldsymbol{k} = \boldsymbol{m}^{(1)} + \cdots + \boldsymbol{m}^{(q)}$, $q \geq 1$, $\boldsymbol{m}^{(j)} \in \mathbb{N}_0^d$ with $\boldsymbol{m}_\cdot^{(j)} \geq 1$, $j = 1, \ldots, q$. Existence of $D^{\boldsymbol{k}} f_0$ of all orders $\boldsymbol{k}_\cdot \leq \lfloor \beta \rfloor$ follows from the same property of $\log f_0$. In fact, by the chain rule $D^{\boldsymbol{k}} f_0(\boldsymbol{x}) = f_0(\boldsymbol{x}) \sum_{P \in \mathcal{P}(\boldsymbol{k})} \prod_{\boldsymbol{m} \in P} D^{\boldsymbol{m}} \log f_0(\boldsymbol{x})$, and hence $P_0 |(D^{\boldsymbol{k}} f_0)/f_0|^{(2\beta+\epsilon)/\boldsymbol{k}_\cdot} < \infty$ by an application of the Hölder inequality. Also, because $\log f_0 \in \mathcal{C}^{\beta, Q_1, 0}(\mathbb{R}^d)$ with Q_1 a polynomial, for every $\boldsymbol{k} \in \mathbb{N}_0^d$ with $\boldsymbol{k}_\cdot < \beta$, we can find polynomial $Q_{\boldsymbol{k},1}$ and $Q_{\boldsymbol{k},2}$ such that $|D^{\boldsymbol{k}} \log f_0(\boldsymbol{x})| < Q_{\boldsymbol{k},1}(\boldsymbol{x})$ and $|D^{\boldsymbol{k}} \log f_0(\boldsymbol{x}+\boldsymbol{y}) - D^{\boldsymbol{k}} \log f_0(\boldsymbol{x})| < Q_{\boldsymbol{k},2}(\boldsymbol{x}) e^{\|\boldsymbol{y}\|^2} \|\boldsymbol{y}\|^{\beta - \lfloor \beta \rfloor}$.

Hence, for $\boldsymbol{k}_\cdot = \lfloor \beta \rfloor$,

$$|D^{\boldsymbol{k}} f_0(\boldsymbol{x}+\boldsymbol{y}) - D^{\boldsymbol{k}} f_0(\boldsymbol{x})| \leq |f_0(\boldsymbol{x}+\boldsymbol{y}) - f_0(\boldsymbol{x})| Q_3(\boldsymbol{x}) + f_0(\boldsymbol{x}) Q_4(\boldsymbol{x}) e^{\tau_2 \|\boldsymbol{y}\|^2} \|\boldsymbol{y}\|^{\beta - \lfloor \beta \rfloor}$$

for some polynomials Q_3 and Q_4 and a $\tau_2 > 0$. Therefore $f_0 \in \mathcal{C}^{\beta, L, \tau_0}$ for $\tau_0 = \max(\tau_1, \tau_2)$ and $L(\boldsymbol{x}) = f_0(\boldsymbol{x})\{Q(\boldsymbol{x}) Q_3(\boldsymbol{x}) + Q_4(\boldsymbol{x})\}$. Because of the tail condition on f_0, for any polynomial $\tilde{Q}$ and $a > 0$, $P_0 |\tilde{Q}|^a < \infty$. And so $P_0 (L/f_0)^{2+\epsilon/\beta} < \infty$ by Hölder's inequality and the assumption on Q. □

Without (2.7), the assumptions made on f_0 in the above lemma match one to one with conditions (C1)–(C3) of Kruijer et al. (2010). The additional assumption (2.7) is a mild one and is satisfied by densities with tails exactly as in the bound (2.9) below with $\tau \leq 2$, and also by finite mixtures of such densities.

2.2.3 Convergence rates results

Let Π be a Dirichlet process mixture prior as defined in Section 2.2.1 and let $\Pi_n(\cdot \mid \boldsymbol{X_1}, \ldots, \boldsymbol{X_n})$ denote the posterior distribution based on n observations $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ modeled as $\boldsymbol{X_i} \sim f$, $f \sim \Pi$. It is known that the minimax rate associated with a β-Hölder class is $n^{-\beta/(2\beta+d)}$. We obtain posterior convergence rate for this class as $n^{-\beta/(2\beta+d)}$ up to a factor a power of $\log n$. A formal result requires some additional conditions on f_0, as summarized in the theorem below.

Theorem 3. *Suppose that $f_0 \in \mathcal{C}^{\beta, L, \tau_0}(\mathbb{R}^d)$ is a probability density function satisfying*

$$\int f_0 \left(|D^{\boldsymbol{k}} f_0| / f_0 \right)^{(2\beta+\epsilon)/\boldsymbol{k}_\cdot} < \infty, \quad \boldsymbol{k} \in \mathbb{N}_0^d, \boldsymbol{k}_\cdot \leq \lfloor \beta \rfloor, \qquad \int f_0 \left(L/f_0 \right)^{(2\beta+\epsilon)/\beta} < \infty \tag{2.8}$$

for some $\epsilon > 0$. Also suppose there are positive constants a, b, c, τ such that

$$f_0(\boldsymbol{x}) \leq c \exp(-b\|\boldsymbol{x}\|^{\tau}), \;\; \|\boldsymbol{x}\| > a. \tag{2.9}$$

For the prior Π constructed in Section 2.2.1, the posterior converges around f_0 in the Hellinger or the L_1-metric with a rate $\epsilon_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$, where

$$t > \{d^*(1 + 1/\tau + 1/\beta) + 1\}/(2 + d^*/\beta), \;\; d^* = \max(d, \kappa).$$

When $\kappa = 1$, the rate in Theorem 3 equals the optimal rate $n^{-\beta/(2\beta+d)}$ up to a factor $\log n$. However, the commonly used inverse Wishart specification of G leads to $\kappa = 2$, and hence Theorem 3 gives the optimal rate only for $d \geq 2$. We shall later see that κ has a bigger impact on rates of convergence for anisotropic densities.

Our result also applies for a finite mixture prior specification Π_{FM}, where the density function f is represented by $f(\boldsymbol{x}) = \sum_{h=1}^{H} \omega_h \phi_{\boldsymbol{\Sigma}}(\boldsymbol{x} - \boldsymbol{\mu_h})$ and priors are assigned on H, $\boldsymbol{\Sigma}$, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_H)$ and $\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_H}$. We assume $\boldsymbol{\Sigma} \sim G$, which satisfies (2.2), (2.3) and (2.4), and that there exist positive constants a_4, b_4, b_5, b_6, b_7, C_4, C_5, C_6, C_7 such that

$$b_4 \exp\{-C_4 x (\log x)^{\tau_1}\} \leq \Pi(H \geq x) \leq b_5 \exp\{-C_5 x (\log x)^{\tau_1}\}$$

for sufficiently large $x > 0$, while for every fixed $H = h$,

$$\Pi(\boldsymbol{\mu_i} \notin [-x, x]^d) \leq b_6 \exp(-C_6 x^{a_4}), \;\; \text{for sufficiently large } x > 0, \;\; i = 1, \ldots, h,$$
$$\Pi(\|\boldsymbol{\omega} - \boldsymbol{\omega_0}\| \leq \epsilon) \geq b_7 \exp\{-C_7 h \log(1/\epsilon)\}, \;\; \text{for all } 0 < \epsilon < 1/h \text{ and all } \boldsymbol{\omega_0} \in \Delta_h.$$

Theorem 4 summarizes our findings for a finite mixture prior.

Theorem 4. *Suppose that $f_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ is a probability density function satisfying* (2.8) *and* (2.9) *for some positive constants a, b, c, τ, ϵ. For a finite mixture prior Π_{FM} as in above, the posterior converges around f_0 in the Hellinger or the L_1-metric with a rate $\epsilon_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$ for every $t > \{d^*(1+1/\tau+1/\beta)+1\}/(2+d^*/\beta)+\max\{0,(1-\tau_1)/2\}$, where $d^* = \max(d,\kappa)$.*

We prove both theorems by verifying a set of sufficient conditions presented in Theorem 2. In the following two sections, we will present relevant results in Theorems 6 and 7. The proofs of these two theorems are given in Section 2.6.

2.3 Prior thickness results

Functions in $\mathcal{C}^{\beta,L,\tau_0}$ can be approximated by mixtures of $\phi_{\sigma^2\boldsymbol{I}}$ with an accuracy that improves with β. We establish this through the following constructions and lemma, which are adapted from Lemma 3.4 of de Jonge and van Zanten (2010) and univariate approximation results of Kruijer et al. (2010).

For each $\boldsymbol{k} \in \mathbb{N}_0^d$, let $m_{\boldsymbol{k}}$ denote the $\boldsymbol{k}$-th moment $m_{\boldsymbol{k}} = \int \boldsymbol{y}^{\boldsymbol{k}}\phi_1(\boldsymbol{y})d\boldsymbol{y}$ of the standard normal distribution on $\mathbb{R}^d$. For $\boldsymbol{n} \in \mathbb{N}_0^d$, define two sequences of numbers through the following recursion. If $\boldsymbol{n}_\cdot = 1$, set $c_{\boldsymbol{n}} = 0$ and $d_{\boldsymbol{n}} = -m_n/n!$, and for $\boldsymbol{n}_\cdot \geq 2$ define

$$c_{\boldsymbol{n}} = -\sum_{\substack{\boldsymbol{n}=\boldsymbol{l}+\boldsymbol{k} \\ \boldsymbol{l}_\cdot \geq 1,\ \boldsymbol{k}_\cdot \geq 1}} \frac{(-1)^{\boldsymbol{k}_\cdot}}{\boldsymbol{k}!} m_{\boldsymbol{k}} d_{\boldsymbol{l}}, \qquad d_{\boldsymbol{n}} = \frac{(-1)^{\boldsymbol{n}_\cdot} m_{\boldsymbol{n}}}{\boldsymbol{n}!} + c_{\boldsymbol{n}}. \tag{2.10}$$

Given $\beta > 0$, $\sigma > 0$, define a transform $T_{\beta,\sigma}$ on $f : \mathbb{R}^d \to \mathbb{R}$ with derivatives up to order $\lfloor\beta\rfloor$ by

$$T_{\beta,\sigma} f = f - \sum_{\substack{\boldsymbol{k} \in \mathbb{N}_0^d \\ 1 \leq \boldsymbol{k}_\cdot \leq \lfloor\beta\rfloor}} d_{\boldsymbol{k}} \sigma^{\boldsymbol{k}_\cdot} D^{\boldsymbol{k}} f. \tag{2.11}$$

Lemma 3. *For any $\beta, \tau_0 > 0$ there is a positive constant M_β such that any $f \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$, it satisfies $|\{K_\sigma(T_{\beta,\sigma}f) - f\}(\boldsymbol{x})| < M_\beta L(\boldsymbol{x})\sigma^\beta$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and all $\sigma \in (0, 1/(2\tau_0)^{1/2})$.*

Proof. For any $f \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$, consider the multivariate Taylor expansion

$$f(\boldsymbol{x}-\boldsymbol{y}) - f(\boldsymbol{x}) = \sum_{1 \le \boldsymbol{k}_\cdot \le \lfloor\beta\rfloor} \frac{(-\boldsymbol{y})^{\boldsymbol{k}_\cdot}}{\boldsymbol{k}!}(D^{\boldsymbol{k}}f)(\boldsymbol{x}) + R(\boldsymbol{x},\boldsymbol{y})$$

with the residual satisfying $|R(\boldsymbol{x},\boldsymbol{y})| \le K_1 L(\boldsymbol{x}) \exp(\tau_0\|\boldsymbol{y}\|^2)\|\boldsymbol{y}\|^\beta$ for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and for a universal constant K_1. Therefore for any $\sigma \in (0, 1/(2\tau_0)^{1/2})$,

$$\begin{aligned}
&\{K_\sigma(T_{\beta,\sigma}f) - f\}(\boldsymbol{x}) \\
&\quad = \int \phi_{\sigma^2\boldsymbol{I}}(\boldsymbol{y})\{f(\boldsymbol{x}-\boldsymbol{y}) - f(\boldsymbol{x})\}d\boldsymbol{y} - \sum_{2 \le \boldsymbol{k}_\cdot \le \lfloor\beta\rfloor} d_{\boldsymbol{k}}\sigma^{\boldsymbol{k}_\cdot}\{K_\sigma(D^{\boldsymbol{k}}f)\}(\boldsymbol{x}) \\
&\quad = \int \phi_{\sigma^2\boldsymbol{I}}(\boldsymbol{y})R(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{y} \\
&\qquad + \sum_{2 \le \boldsymbol{k}_\cdot \le \lfloor\beta\rfloor} \sigma^{\boldsymbol{k}_\cdot}\left[\frac{(-1)^{\boldsymbol{k}_\cdot} m_{\boldsymbol{k}}}{\boldsymbol{k}!}(D^{\boldsymbol{k}}f)(\boldsymbol{x}) - d_{\boldsymbol{k}}\{K_\sigma(D^{\boldsymbol{k}}f)\}(\boldsymbol{x})\right]. \qquad (2.12)
\end{aligned}$$

The first term of (2.12) is bounded by $K_2 L(\boldsymbol{x})\sigma^\beta$ for some universal constant K_2. If $\beta \le 2$ then the second term of (2.12) does not exist and we get a proof with $M_\beta = K_2$. For $\beta > 2$ we use induction on $\lfloor\beta\rfloor$.

From (2.10) we can rewrite the second term of (2.12) as

$$\sum_{2 \le \boldsymbol{k}_\cdot \le \lfloor\beta\rfloor} \left[\frac{(-1)^{\boldsymbol{k}_\cdot} m_{\boldsymbol{k}}\sigma^{\boldsymbol{k}_\cdot}}{\boldsymbol{k}!}\{D^{\boldsymbol{k}}f - K_\sigma(D^{\boldsymbol{k}}f)\}(\boldsymbol{x}) - c_{\boldsymbol{k}}\sigma^{\boldsymbol{k}_\cdot}\{K_\sigma(D^{\boldsymbol{k}}f)\}(\boldsymbol{x})\right]. \qquad (2.13)$$

For each $1 \le \boldsymbol{k}_\cdot \le \lfloor\beta\rfloor$, the induction hypothesis implies that $D^{\boldsymbol{k}}f \in \mathcal{C}^{\beta-\boldsymbol{k}_\cdot,L,\tau_0}(\mathbb{R}^d)$ and

$$D^{\boldsymbol{k}}f - K_\sigma(D^{\boldsymbol{k}}f) = \{D^{\boldsymbol{k}}f - K_\sigma T_{\beta-\boldsymbol{k}_\cdot,\sigma}(D^{\boldsymbol{k}}f)\} + K_\sigma\{T_{\beta-\boldsymbol{k}_\cdot,\sigma}(D^{\boldsymbol{k}}f) - D^{\boldsymbol{k}}f\} \qquad (2.14)$$

with $|\{D^{\boldsymbol{k}}f - K_\sigma T_{\beta-\boldsymbol{k}_\cdot,\sigma}(D^{\boldsymbol{k}}f)\}(\boldsymbol{x})| \leq M_{\beta-\boldsymbol{k}_\cdot} L(\boldsymbol{x})\sigma^{\beta-\boldsymbol{k}_\cdot}$, for all $x \in \mathbb{R}^d$. This establishes the claim with $M_\beta = K_2 + \sum_{2\leq \boldsymbol{k}_\cdot \leq \lfloor\beta\rfloor}(m_k/\boldsymbol{k}!)M_{\beta-\boldsymbol{k}_\cdot}$ because

$$\begin{aligned}
&\sum_{2\leq \boldsymbol{k}_\cdot \leq \lfloor\beta\rfloor} \left[\frac{(-1)^{\boldsymbol{k}_\cdot} m_{\boldsymbol{k}}\sigma^{\boldsymbol{k}_\cdot}}{\boldsymbol{k}!}\{T_{\beta-\boldsymbol{k}_\cdot,\sigma}(D^{\boldsymbol{k}}f) - D^{\boldsymbol{k}}f\} - c_{\boldsymbol{k}}\sigma^{\boldsymbol{k}_\cdot} D^k f\right] \\
&\qquad = \sum_{2\leq \boldsymbol{k}_\cdot \leq \lfloor\beta\rfloor} \left\{ \frac{(-1)^{\boldsymbol{k}_\cdot} m_{\boldsymbol{k}}\sigma^{\boldsymbol{k}_\cdot}}{\boldsymbol{k}!} \sum_{1\leq \boldsymbol{j}_\cdot \leq \lfloor\beta\rfloor - \boldsymbol{k}_\cdot} d_{\boldsymbol{j}}\sigma^{\boldsymbol{j}_\cdot} D^{\boldsymbol{k}+\boldsymbol{j}} f - c_{\boldsymbol{k}}\sigma^{\boldsymbol{k}_\cdot} D^{\boldsymbol{k}} f \right\} \\
&\qquad = \sum_{3\leq \boldsymbol{n}_\cdot \leq \lfloor\beta\rfloor} \left\{ \sum_{\substack{\boldsymbol{n}=\boldsymbol{l}+\boldsymbol{k} \\ \boldsymbol{l}_\cdot \geq 1, \boldsymbol{k}_\cdot \geq 2}} \frac{(-1)^{\boldsymbol{k}_\cdot}}{\boldsymbol{k}!} m_{\boldsymbol{k}} d_{\boldsymbol{l}} - c_{\boldsymbol{n}} \right\} \sigma^{\boldsymbol{n}_\cdot} D^{\boldsymbol{n}} f = 0
\end{aligned}$$

identically, by definitions of $c_{\boldsymbol{n}}$ and $d_{\boldsymbol{n}}$. □

Lemma 3 applies to any functions $f \in \mathcal{C}^{\beta,L,\tau_0}$, not necessarily a probability density, and the mixing function $T_{\beta,\sigma}f$ need not be a density in that it may be negative. Fortunately, when f is a probability density, we can derive a density h_σ from $T_{\beta,\sigma}f$ so that $K_\sigma h_\sigma$ provides a σ^β order approximation to f. The construction of h_σ can be viewed as a multivariate extension of results in Section 3 of Kruijer et al. (2010). The main difference is that we establish approximation results under the Hellinger distance and apply Taylor expansions on f_0 instead of $\log f_0$, which lead to a more elegant proof.

Theorem 5. *Let $f_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ be a probability density function and write $f_\sigma = T_{\beta,\sigma}f_0$. Suppose that f_0 satisfies* (2.8) *for some $\epsilon > 0$. Then there exist $s_0 > 0, K > 0$ such that for any $0 < \sigma < s_0$, $g_\sigma = f_\sigma + \frac{1}{2}f_0\,\mathbb{1}\{f_\sigma < (1/2)f_0\}$ is a non-negative function with $\int g_\sigma(\boldsymbol{x})d\boldsymbol{x} < \infty$ and the density $h_\sigma = g_\sigma/\int g_\sigma(\boldsymbol{x})d\boldsymbol{x}$ satisfies $d_H^2(f_0, K_\sigma h_\sigma) \leq K\sigma^{2\beta}$.*

Proof. Fix $s_0 \in (0, 1/(2\tau_0)^{1/2})$ such that

$$\sum_{1\leq \boldsymbol{k}_\cdot \leq \lfloor\beta\rfloor} |d_{\boldsymbol{k}}||\log\sigma|^{-\boldsymbol{k}_\cdot/2} < 1/2 \text{ and } \sigma^\epsilon|\log\sigma|^{(2\beta+\epsilon)/2} < 1$$

for all $0 < \sigma < s_0$. For any $\sigma \in (0, s_0)$ define

$$A_\sigma = \left\{ \boldsymbol{x} : \frac{|D^{\boldsymbol{k}} f_0(\boldsymbol{x})|}{f_0(\boldsymbol{x})} \leq \sigma^{-\boldsymbol{k}_\cdot} |\log \sigma|^{-\boldsymbol{k}_\cdot/2}, \boldsymbol{k}_\cdot \leq \lfloor \beta \rfloor, \ \ \frac{L(\boldsymbol{x})}{f_0(\boldsymbol{x})} \leq \sigma^{-\beta} |\log \sigma|^{-\beta/2} \right\}$$

and notice that by Markov's inequality

$$\begin{aligned}
P_0(A_\sigma^c) &\leq \sum_{\boldsymbol{k}_\cdot \leq \lfloor \beta \rfloor} P_0 \left\{ \frac{|D^{\boldsymbol{k}} f_0(\boldsymbol{X})|}{f_0(\boldsymbol{X})} > \sigma^{-\boldsymbol{k}_\cdot} |\log \sigma|^{-\boldsymbol{k}_\cdot/2} \right\} + P_0 \left\{ \frac{L(\boldsymbol{X})}{f_0(\boldsymbol{X})} > \sigma^{-\beta} |\log \sigma|^{-\beta/2} \right\} \\
&= \sum_{\boldsymbol{k}_\cdot \leq \lfloor \beta \rfloor} P_0 \left\{ \left(|D^{\boldsymbol{k}} f_0| / f_0 \right)^{(2\beta+\epsilon)/\boldsymbol{k}_\cdot} > \sigma^{-(2\beta+\epsilon)} |\log \sigma|^{-(2\beta+\epsilon)/2} \right\} \\
&\qquad + P_0 \left\{ (L/f_0)^{(2\beta+\epsilon)/\beta} > \sigma^{-(2\beta+\epsilon)} |\log \sigma|^{-(2\beta+\epsilon)/2} \right\} \\
&\leq \sigma^{2\beta+\epsilon} |\log \sigma|^{(2\beta+\epsilon)/2} \left\{ \sum_{\boldsymbol{k}_\cdot \leq \lfloor \beta \rfloor} P_0 \left(|D^{\boldsymbol{k}} f_0| / f_0 \right)^{(2\beta+\epsilon)/\boldsymbol{k}_\cdot} + P_0 \left(L/f_0 \right)^{(2\beta+\epsilon)/\beta} \right\},
\end{aligned}$$

which is bounded by $K_1 \sigma^{2\beta}$ for some constant K_1. Also, for any $x \in A_\sigma$,

$$|(f_\sigma - f_0)(\boldsymbol{x})| \leq \sum_{1 \leq \boldsymbol{k}_\cdot \leq \lfloor \beta \rfloor} |d_{\boldsymbol{k}}| \sigma^{\boldsymbol{k}_\cdot} |D^{\boldsymbol{k}} f_0(\boldsymbol{x})| \leq f_0(\boldsymbol{x}) \sum_{1 \leq \boldsymbol{k}_\cdot \leq \lfloor \beta \rfloor} |d_{\boldsymbol{k}}| |\log \sigma|^{-\boldsymbol{k}_\cdot/2} \leq \frac{1}{2} f_0(\boldsymbol{x}).$$

Consequently, $f_\sigma \geq f_0/2$ on A_σ. Because of integrability conditions on $D^{\boldsymbol{k}} f_0 / f_0$, it turns out that in calculating $\int D^{\boldsymbol{k}} f_0(\boldsymbol{x}) d\boldsymbol{x}$ for any $1 \leq \boldsymbol{k}_\cdot \leq \lfloor \beta \rfloor$, one can integrate under the derivative and conclude that $\int D^{\boldsymbol{k}} f_0(\boldsymbol{x}) d\boldsymbol{x} = 0$ as f_0 is a density. So $\int f_\sigma(\boldsymbol{x}) d\boldsymbol{x} = 1$ and for some constant K_2 and all $\sigma < s_0$,

$$1 \leq \int g_\sigma(\boldsymbol{x}) d\boldsymbol{x} \leq 1 + \frac{1}{2} \int f_0(\boldsymbol{x}) \mathbb{1}\{f_\sigma(\boldsymbol{x}) < f_0(\boldsymbol{x})/2\} d\boldsymbol{x} \leq 1 + \frac{1}{2} P_0(A_\sigma^c) \leq 1 + K_2 \sigma^{2\beta}.$$

So $\int g_\sigma(\boldsymbol{x}) d\boldsymbol{x} < \infty$ and h_σ is a well defined probability density function on $\mathbb{R}^d$.

To prove the final result of Theorem 5, denote $r_\sigma = \frac{1}{2} f_0 \, \mathbb{1}\{f_\sigma < (1/2) f_0\}$ and $c_\sigma =$

$\int g_\sigma(\boldsymbol{x})d\boldsymbol{x}$ and note that for $a > 0, b > 0$, we have $(\sqrt{a} - \sqrt{b})^2 = (a-b)^2/(\sqrt{a} + \sqrt{b})^2 \leq (a-b)^2/(a+b)$ and hence

$$\begin{aligned}
d_H^2(f_0, K_\sigma h_\sigma) &\leq \int \frac{(f_0 - K_\sigma h_\sigma)^2(\boldsymbol{x})}{f_0(\boldsymbol{x}) + (K_\sigma h_\sigma)(\boldsymbol{x})} d\boldsymbol{x} \\
&= \frac{1}{c_\sigma} \int \frac{(c_\sigma f_0 - K_\sigma g_\sigma)^2(\boldsymbol{x})}{c_\sigma f_0(\boldsymbol{x}) + (K_\sigma g_\sigma)(\boldsymbol{x})} d\boldsymbol{x} \\
&\leq 3 \int \frac{(c_\sigma - 1)^2 f_0^2(\boldsymbol{x}) + (f_0 - K_\sigma f_\sigma)^2(\boldsymbol{x}) + (K_\sigma r_\sigma)^2(\boldsymbol{x})}{c_\sigma f_0(x) + (K_\sigma g_\sigma)(\boldsymbol{x})} d\boldsymbol{x} \\
&\leq 3 \left\{ \int (c_\sigma - 1)^2 f_0(\boldsymbol{x}) d\boldsymbol{x} + \int \frac{(f_0 - K_\sigma f_\sigma)^2(\boldsymbol{x})}{f_0(\boldsymbol{x})} d\boldsymbol{x} + \int \frac{(K_\sigma r_\sigma)^2(\boldsymbol{x})}{(K_\sigma g_\sigma)(\boldsymbol{x})} d\boldsymbol{x} \right\} \\
&\leq 3 \left\{ K_2^2 \sigma^{4\beta} + M_\beta^2 \sigma^{2\beta} P_0 \left(L/f_0\right)^2 + \int (K_\sigma r_\sigma)(\boldsymbol{x}) d\boldsymbol{x} \right\}, \qquad (2.15)
\end{aligned}$$

because $1 \leq c_\sigma \leq 1 + K_2 \sigma^{2\beta}$, $|(f_0 - K_\sigma f_\sigma)(\boldsymbol{x})| < M_\beta L(x) \sigma^\beta$ and $K_\sigma r_\sigma \leq K_\sigma g_\sigma$ since $r_\sigma \leq g_\sigma$. By Jensen's inequality $P_0(L/f_0)^2 \leq \{P_0(L/f_0)^{(2\beta+\epsilon)/\beta}\}^{\beta/(\beta+\epsilon/2)} < \infty$. Also,

$$\begin{aligned}
\int (K_\sigma r_\sigma)(\boldsymbol{x}) d\boldsymbol{x} &= \frac{1}{2} \int \int \phi_{\sigma^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{y}) f_0(\boldsymbol{y}) \mathbb{1}\{f_\sigma(\boldsymbol{y}) < f_0(\boldsymbol{y})/2\} dx dy \\
&= \frac{1}{2} \int f_0(\boldsymbol{y}) \mathbb{1}\{f_\sigma(\boldsymbol{y}) < f_0(\boldsymbol{y})/2\} d\boldsymbol{y} \\
&\leq P_0(A_\sigma^c) \\
&\leq K_1 \sigma^{2\beta}.
\end{aligned}$$

□

The next result trades g_σ for a compactly supported density h_σ whose convolution with $\phi_{\sigma^2 I}$ inherits the same order σ^β approximation to f_0. We need the tail condition (2.9) on f_0 to obtain a suitable compact support.

Proposition 1. *Let $f_0 \in \mathcal{C}^{\beta, L, \tau_0}(\mathbb{R}^d)$ be a probability density function satisfying* (2.8) *and* (2.9) *for some positive constants ϵ, a, b, c, τ. For any $\sigma > 0$, define $E_\sigma = \{\boldsymbol{x} \in \mathbb{R}^d : f_0(\boldsymbol{x}) \geq$*

$\sigma^{(4\beta+2\epsilon+8)/\delta}\}$. *Then there exist* $s_0, a_0, B_0, K_0 > 0$ *such that for every* $0 < \sigma < s_0$, $P_0(E_\sigma^c) \leq B_0\sigma^{4\beta+2\epsilon+8}$, $E_\sigma \subset \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq a_\sigma\}$ *where* $a_\sigma = a_0\{\log(1/\sigma)\}^\tau$ *and there is a probability density* $\tilde{h}_\sigma$ *with support inside* $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq a_\sigma\}$ *satisfying* $d_H(f_0, K_\sigma\tilde{h}_\sigma) \leq K_0\sigma^\beta$.

Proof. Define g_σ and h_σ as in the statement of Theorem 5. This theorem implies that there are $s_1, K > 0$ such that $d_H^2(f_0, K_\sigma h_\sigma) \leq K\sigma^{2\beta}$ for all $0 < \sigma < s_1$. The tail condition on f_0 implies existence of a small $\delta > 0$ such that B_0, which is defined as $P_0(f_0^{-\delta})$, satisfies $B_0 < \infty$. Let $s_2 \in (0, 1/(2\tau_0)^{1/2})$ be such that $\{(4\beta + 2\epsilon + 8)/(b\delta)\}\log(1/s_2) > \max\{(1/b)\log c, a^\tau/2\}$. Set $s_0 = \min(s_1, s_2)$ and pick any $\sigma \in (0, s_0)$. Define $E_\sigma = \{\boldsymbol{x} \in \mathbb{R}^d : f_0(\boldsymbol{x}) \geq \sigma^{(4\beta+2\epsilon+8)/\delta}\}$ and $a_\sigma = a_0\log(1/\sigma)^{1/\tau}$ with $a_0 = \{(8\beta + 4\epsilon + 16)/(b\delta)\}^{1/\tau}$. Then $a_\sigma > a$ and $E_\sigma \subset \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq a_\sigma\}$.

By Markov's inequality,

$$P_0(E_\sigma^c) = P_0\{f_0(\boldsymbol{x})^{-\delta} > \sigma^{-(4\beta+2\epsilon+8)}\} \leq B_0\sigma^{4\beta+2\epsilon+8} \leq B_0\sigma^{2\beta+\epsilon}$$

and consequently by (2.8) and applications of Hölder's inequality

$$\begin{aligned}
&\int_{E_\sigma^c} g_\sigma(\boldsymbol{x})d\boldsymbol{x} \\
&\quad\leq \frac{3}{2}\int_{E_\sigma^c} f_0(\boldsymbol{x})d\boldsymbol{x} + \sum_{\boldsymbol{k}_\cdot=1}^{\lfloor\beta\rfloor} \sigma^{\boldsymbol{k}_\cdot}|d_{\boldsymbol{k}}| \int_{E_\sigma^c} |D^{\boldsymbol{k}} f_0(\boldsymbol{x})|d\boldsymbol{x} \\
&\quad\leq \frac{3}{2}P_0(E_\sigma^c) + \sum_{\boldsymbol{k}_\cdot=1}^{\lfloor\beta\rfloor} \sigma^{\boldsymbol{k}_\cdot}|d_{\boldsymbol{k}}| \left\{P_0\left(|D^{\boldsymbol{k}} f_0|/f_0\right)^{(2\beta+\epsilon)/\boldsymbol{k}_\cdot}\right\}^{\boldsymbol{k}_\cdot/2\beta+\epsilon} P_0(E_\sigma^c)^{(2\beta+\epsilon-\boldsymbol{k}_\cdot)/(2\beta+\epsilon)},
\end{aligned}$$

which is bounded by $B_1\sigma^{2\beta+\epsilon}$ for some constant B_1 that does not depend on σ. Hence

$$\int_{E_\sigma^c} h_\sigma(x)dx \leq \int_{E_\sigma^c} g_\sigma(x)dx \leq B_1\sigma^{2\beta+\epsilon}.$$

Define $\tilde{h}_\sigma$ to be the restriction of h_σ to E_σ, i.e., $\tilde{h}_\sigma(\boldsymbol{x}) = h_\sigma(\boldsymbol{x})\mathbb{1}(\boldsymbol{x} \in E_\sigma)/\int_{E_\sigma} h(\boldsymbol{x})d\boldsymbol{x}$. Then

$$d_H(K_\sigma h_\sigma, K_\sigma \tilde{h}_\sigma) \leq d_H(h_\sigma, \tilde{h}_\sigma) = \left[2 - 2\left\{\int_{E_\sigma} h_\sigma(\boldsymbol{x})d\boldsymbol{x}\right\}^{1/2}\right]^{1/2} = O(\sigma^{\beta+\epsilon/2}).$$

This completes the proof because $d_H(f_0, K_\sigma \tilde{h}_\sigma) \leq d_H(f_0, K_\sigma h_\sigma) + d_H(K_\sigma h_\sigma, K_\sigma \tilde{h}_\sigma)$. □

Proposition 1 paves the way to calculating prior thickness around f_0 because the probability density $K_\sigma \tilde{h}_\sigma$ can be well approximated by densities $p_{F,\Sigma}$ with $(F, \boldsymbol{\Sigma})$ chosen from a suitable set. Toward this we present the final theorem of this section and a proof of it that overlaps with Section 9 of Ghosal and van der Vaart (2007b). However, our proof requires new calculations to handle a non-compactly supported f_0 and a matrix valued $\boldsymbol{\Sigma}$.

Theorem 6. *Let $f_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ be a bounded probability density function satisfying* (2.8) *and* (2.9) *for some positive constants ϵ, a, b, c, τ, and let $F \sim \mathrm{DP}(\alpha)$ and $\boldsymbol{\Sigma} \sim G$ independently. Then for some $A, C > 0$ and all sufficiently large n,*

$$\mathrm{P}\left\{(F, \boldsymbol{\Sigma}) : P_0 \log \frac{f_0}{p_{F,\boldsymbol{\Sigma}}} \leq A\tilde{\epsilon}_n^2, \;\; P_0\left(\log \frac{f_0}{p_{F,\boldsymbol{\Sigma}}}\right)^2 \leq A\tilde{\epsilon}_n^2\right\} \geq e^{-Cn\tilde{\epsilon}_n^2} \tag{2.16}$$

where $\tilde{\epsilon}_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$ with any $t \geq \{d^*(1 + 1/\tau + 1/\beta) + 1\}/(2 + d^*/\beta)$.

Proof. Let δ, s_0, a_0, K_0 be as in Proposition 1. Consider n large enough so that $\tilde{\epsilon}_n < s_0^\beta$. Fix $\sigma^\beta = \tilde{\epsilon}_n\{\log(1/\tilde{\epsilon}_n)\}^{-1}$ and as in Proposition 1 define $E_\sigma = \{\boldsymbol{x} \in \mathbb{R}^d : f_0(\boldsymbol{x}) \geq \sigma^{(4\beta+2\epsilon+8)/\delta}\}$ and $a_\sigma = a_0\{\log(1/\sigma)\}^{1/\tau}$. Recall that $P_0(E_\sigma^c) \leq B_0\sigma^{4\beta+2\epsilon+8}$ for some constant B_0 and $E_\sigma \subset \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq a_\sigma\}$. Apply Proposition 1 to find $\tilde{h}_\sigma$ with support E_σ such that $d_H(f_0, K_\sigma \tilde{h}_\sigma) \leq K_0\sigma^\beta$. Find $b_1 > \max(1, 1/2\beta)$ such that $\tilde{\epsilon}_n^{b_1}\{\log(1/\tilde{\epsilon}_n)\}^{5/4} \leq \tilde{\epsilon}_n$.

By Corollary 1, there is a discrete probability measure $F_\sigma = \sum_{j=1}^{N} p_j \delta_{z_j}$, with at most

$$N \leq D_0 \sigma^{-d} \{\log(1/\sigma)\}^{d/\tau} \{\log(1/\tilde{\epsilon}_n)\}^d \leq D_1 \sigma^{-d} \{\log(1/\tilde{\epsilon}_n)\}^{d+d/\tau}$$

many support points inside $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq a_\sigma\}$, with at least $\sigma\tilde{\epsilon}_n^{2b_1}$ separation between any $z_i \neq z_j$ such that $d_H(K_\sigma \tilde{h}_\sigma, K_\sigma F_\sigma) \leq A_1 \tilde{\epsilon}_n^{b_1} \{\log(1/\tilde{\epsilon}_n)\}^{1/4}$ for some constants A_1, D_1.

Place disjoint balls U_j with centers at $\boldsymbol{z_1}, \ldots, \boldsymbol{z_N}$ with diameter $\sigma\tilde{\epsilon}_n^{2b_1}$ each. Extend the collection $\{U_1, \ldots, U_N\}$ to a partition $\{U_1, \ldots, U_K\}$ of $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq a_\sigma\}$ such that each U_j, $j = N+1, \ldots, K$, has a diameter at most σ. This can be done with $K \leq D_2 \sigma^{-d} \{\log(1/\tilde{\epsilon}_n)\}^{d+d/\tau}$ for some constant D_2. Further extend this to a partition $U_1, \ldots, U_M$ of $\mathbb{R}^d$ such that $a_1(\sigma\tilde{\epsilon}_n^{2b_1})^d \leq \alpha(U_j) \leq 1$ for all $j = 1, \ldots, M$ for some constant a_1. We can still have

$$M \leq D_3 \sigma^{-d} \{\log(1/\tilde{\epsilon}_n)\}^{d+d/\tau} \leq D_4 \tilde{\epsilon}_n^{-d/\beta} \{\log(1/\tilde{\epsilon}_n)\}^{sd}$$

with $s = 1 + \beta^{-1} + \tau^{-1}$, for some constants D_3, D_4. None of these constants depends on n or σ.

Define $p_j = 0$, $j = N+1, \ldots, M$. Let $\mathcal{P}_\sigma$ denote the set of probability measures F on $\mathbb{R}^d$ with $\sum_{j=1}^{M} |F(U_j) - p_j| \leq 2\tilde{\epsilon}_n^{2db_1}$ and $\min_{1 \leq j \leq M} F(U_j) \geq \tilde{\epsilon}_n^{4db_1}/2$. Note that

$$M\tilde{\epsilon}_n^{2db_1} \leq D_4 [\tilde{\epsilon}_n^{b_1 - 1/(2\beta)} \{\log(1/\tilde{\epsilon}_n)\}^{s/2}]^{2d} \leq 1,$$
$$\min_{1 \leq j \leq M} \alpha(U_j)^{1/2} \geq a_1^{1/2} \tilde{\epsilon}_n^{2db_1} \{\tilde{\epsilon}_n^{b_1 - 1/(2\beta)} \log(1/\tilde{\epsilon}_n)\}^{-d} \geq (a_1/D_4)^{1/2} \tilde{\epsilon}_n^{2db_1},$$

provided n has been chosen large enough. By Lemma 10 of Ghosal and van der Vaart (2007b),

$$\mathrm{P}(\mathcal{P}_\sigma) \geq C_1 \exp\{-c_1 M \log(1/\tilde{\epsilon}_n)\} \geq C_1 \exp[-c_2 \tilde{\epsilon}_n^{-d/\beta} \{\log(1/\tilde{\epsilon}_n)\}^{sd+1}]$$

for some constants C_1, c_2 that depend on $\alpha(\mathbb{R}^d)$, a_1, D_4, d and b_1. Also, let $\mathcal{S}_\sigma$ denote the set of all $d \times d$ non-singular matrices Σ such that all eigenvalues of Σ^{-1} lie between σ^{-2} and $\sigma^{-2}(1+\sigma^{2\beta})$. By (2.4),

$$G(\mathcal{S}_\sigma) \geq \sigma^{D_5} \exp(-D_6/\sigma^\kappa) \geq C_3 \exp[-c_3 \tilde{\epsilon}_n^{-\kappa/\beta} \{\log(1/\tilde{\epsilon}_n)\}^{s\kappa+1}]$$

for some constants C_3, c_3. Any $\mathbf{\Sigma} \in \mathcal{S}_\sigma$ satisfies

$$\det(\mathbf{\Sigma}^{-1}) \geq \sigma^{-2d}, \;\; \boldsymbol{y}^T \mathbf{\Sigma}^{-1} \boldsymbol{y} \leq 2\|\boldsymbol{y}\|^2/\sigma^2$$

for any $\boldsymbol{y} \in \mathbb{R}^d$ and

$$|\mathrm{tr}(\sigma^2 \mathbf{\Sigma}^{-1}) - d - \log\det(\sigma^2 \mathbf{\Sigma}^{-1})| < d\sigma^{2\beta}.$$

Apply Lemma 4 with $V_i = U_i$, $i = 1, \ldots, N$ and $V_0 = \cup_{j>N} U_j$ to conclude that for any $F \in \mathcal{P}_\sigma$, $d_H(K_\sigma F_\sigma, K_\sigma F) \leq A_2 \tilde{\epsilon}_n^{b_1}$ for some universal constant A_2 and hence

$$\begin{aligned} d_H(f_0, K_\sigma F) &\leq d_H(f_0, K_\sigma \tilde{h}_\sigma) + d_H(K_\sigma \tilde{h}_\sigma, K_\sigma F_\sigma) + d_H(K_\sigma F_\sigma, \phi_{\sigma^2 I} * F) \\ &\leq K_0 \sigma^\beta + A_1 \tilde{\epsilon}_n^{b_1} \{\log(1/\tilde{\epsilon}_n)\}^{1/4} + A_2 \tilde{\epsilon}_n^{b_1} \leq A_3 \sigma^\beta \end{aligned}$$

for some constant A_3. Therefore, for any $F \in \mathcal{P}_\sigma$, $\mathbf{\Sigma} \in \mathcal{S}_\sigma$,

$$d_H(f_0, p_{F,\Sigma}) \leq d_H(f_0, K_\sigma F) + d_H(p_{F,\sigma^2 I}, p_{F,\Sigma}) \leq A_4 \sigma^\beta$$

for some constant A_4 because

$$d_H(p_{F,\sigma^2 I}, p_{F,\Sigma}) \leq |\mathrm{tr}(\sigma^2 \mathbf{\Sigma}^{-1}) - d - \log\det(\sigma^2 \mathbf{\Sigma}^{-1})|^{1/2}$$

for any F. Moreover, for every $\boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{x}\| < a_\sigma$,

$$\frac{p_{F,\boldsymbol{\Sigma}}(\boldsymbol{x})}{f_0(\boldsymbol{x})} \geq \frac{K_1}{\sigma^d} \int_{\|\boldsymbol{x}-\boldsymbol{z}\| \leq \sigma} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{\sigma^2}\right) F(d\boldsymbol{z}) \geq \frac{K_2}{\sigma^d} F(U_{J(\boldsymbol{x})}) \geq K_3 \frac{\tilde{\epsilon}_n^{4db_1}}{\sigma^d}$$

for some constants K_1, K_2, K_3, where $J(\boldsymbol{x})$ denotes the index $j \in \{1, \ldots, K\}$ for which $\boldsymbol{x} \in U_j$. The penultimate inequality follows because $U_{J(\boldsymbol{x})}$ with diameter no larger than σ must be a subset of a ball of radius σ around $\boldsymbol{x}$. Also, for any $\boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{x}\| > a_\sigma$,

$$\frac{p_{F,\Sigma}(\boldsymbol{x})}{f_0(\boldsymbol{x})} \geq \frac{K_1}{\sigma^d} \int_{\|\boldsymbol{z}\| \leq a_\sigma} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{\sigma^2}\right) F(d\boldsymbol{z}) \geq \frac{K_4}{\sigma^d} \exp(-4\|\boldsymbol{x}\|^2/\sigma^2)$$

for some constant K_4 because $\|\boldsymbol{x}-\boldsymbol{z}\|^2 \leq 2\|\boldsymbol{x}\|^2 + 2\|\boldsymbol{z}\|^2 \leq 4\|\boldsymbol{x}\|^2$ and

$$F(\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq a_\sigma\}) \geq 1 - 2\tilde{\epsilon}_n^{2db_1}.$$

Set $\lambda = K_3 \tilde{\epsilon}_n^{4db_1}/\sigma^d$ and notice $\log(1/\lambda) \leq K_5 \log(1/\tilde{\epsilon}_n)$ for some constant K_5. For any $F \in \mathcal{P}_\sigma$, $\boldsymbol{\Sigma} \in \mathcal{S}_\sigma$,

$$\begin{aligned} P_0\left\{\left(\log \frac{f_0}{p_{F,\boldsymbol{\Sigma}}}\right)^2 \mathbb{1}\left(\frac{p_{F,\boldsymbol{\Sigma}}}{f_0} < \lambda\right)\right\} &\leq \frac{K_6}{\sigma^4} \int_{\|\boldsymbol{x}\| > a_\sigma} \|\boldsymbol{x}\|^4 f_0(\boldsymbol{x}) d\boldsymbol{x} \\ &\leq \frac{K_6}{\sigma^4} (P_0 \|\boldsymbol{X}\|^8)^{1/2} P_0(E_\sigma^c)^{1/2} \leq K_7 \sigma^{2\beta+\epsilon} \end{aligned}$$

for some constant K_7 since $P_0\|\boldsymbol{X}\|^m < \infty$ for all $m > 0$ because of the tail condition (2.9). Given n sufficiently large, we have $\lambda < e^{-1}$ and hence

$$\log \frac{f_0}{p_{F,\boldsymbol{\Sigma}}} \mathbb{1}\left(\frac{p_{F,\boldsymbol{\Sigma}}}{f_0} < \lambda\right) \leq \left(\log \frac{f_0}{p_{F,\boldsymbol{\Sigma}}}\right)^2 \mathbb{1}\left(\frac{p_{F,\boldsymbol{\Sigma}}}{f_0} < \lambda\right).$$

Therefore $P_0\{\log(f_0/p_{F,\boldsymbol{\Sigma}}) \mathbb{1}(p_{F,\boldsymbol{\Sigma}}/f_0 < \lambda)\} \leq K_7 \sigma^{2\beta+\epsilon}$. Now apply Lemma 5 to conclude

that

$$\max\left\{\int f_0 \log\frac{f_0}{p_{F,\mathbf{\Sigma}}}, \int f_0\left(\log\frac{f_0}{p_{F,\mathbf{\Sigma}}}\right)^2\right\} \leq K_8 \log\left(\frac{1}{\lambda}\right)^2 \sigma^{2\beta} \leq K_9\sigma^{2\beta}\left(\log\frac{1}{\tilde{\epsilon}_n}\right)^2 \leq A\tilde{\epsilon}_n^2$$

for some positive constant A. Therefore

$$\begin{aligned}\mathrm{P}\left\{P_0\log\frac{f_0}{p_{F,\mathbf{\Sigma}}} \leq A\tilde{\epsilon}_n^2, P_0\left(\log\frac{f_0}{p_{F,\mathbf{\Sigma}}}\right)^2 \leq A\tilde{\epsilon}_n^2\right\} &\geq \mathrm{P}(\mathcal{P}_\sigma)G(\mathcal{S}_\sigma) \\ &\geq C_4\exp\left[-c_4\tilde{\epsilon}_n^{-d^*/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{sd^*+1}\right].\end{aligned}$$

This gives (2.16) provided $\tilde{\epsilon}_n^{-d^*/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{sd^*+1} \leq n\tilde{\epsilon}_n^2$. With $\tilde{\epsilon}_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$, the condition is satisfied if $t \geq (sd^*+1)/(2+d^*/\beta)$. □

2.4 Sieve construction

In the following proposition, based on the stick-breaking representation of a Dirichlet process, we give an explicit definition of the sieve and derive upper bounds for its entropy and the prior probability of its complement. This result serves as the main tool in obtaining adaptive posterior convergence rates.

Proposition 2. *Fix $\epsilon, a, \sigma_0 > 0$ and integers $M, H \geq d$. Define*

$$\mathcal{Q} = \left\{ p_{F,\mathbf{\Sigma}} \text{ with } F = \sum_{h=1}^{\infty}\pi_h\delta_{z_h} : \begin{array}{l} z_h \in [-a,a]^d, h \leq H; \quad \sum_{h>H}\pi_h < \epsilon; \\ \sigma_0^2 \leq \mathrm{eig}_j(\mathbf{\Sigma}) < \sigma_0^2\left(1+\epsilon^2/d\right)^M, \quad j = 1,\ldots,d \end{array} \right\} \tag{2.17}$$

and let Π denote the Dirichlet mixture prior defined in Section 2.2.1. Then:

(i) For some constant K and ρ defined as either the Hellinger or the L_1-metric,

$$\log N(\epsilon, \mathcal{Q}, \rho) \leq K\left\{dH \log\left\{a/(\sigma_0\epsilon)\right\} - H\log\epsilon + \log M + M\epsilon^2\right\}.$$

(ii) With the constants as defined in (2.1)–(2.4)*, we have*

$$\begin{aligned}\Pi(\mathcal{Q}^c) \;\leq\; & b_1 H\exp\{-C_1 a^{a_1}\} + \{(e|\alpha|/H)\log(1/\epsilon)\}^H \\ & + b_2\exp\{-C_2\sigma_0^{-2a_2}\} + b_3\sigma_0^{-2a_3}(1+\epsilon^2/d)^{-2Ma_3}.\end{aligned}$$

Proof. Let $\hat{R}$ be a $(\sigma_0\epsilon)$-net of $[-a,a]^d$, $\hat{S}$ be an ϵ-net of the H-simplex Δ_H and $\hat{\mathcal{O}}$ be an δ-net of $\mathcal{O}_d$, the group of $d\times d$ orthogonal matrices equipped with the spectral norm $\|\cdot\|_2$, where $\delta = \epsilon^2/\{3d(1+\epsilon^2/d)^M\}$. It is known that the cardinality of $\hat{R}$ is bounded by a constant multiple of $\{a/(\sigma_0\epsilon)\}^d$, the cardinality of $\hat{S}$ is bounded by a constant multiple of ϵ^{-H} and the cardinality of $\hat{\mathcal{O}}$ is bounded by a constant multiple of $\delta^{-d(d-1)/2}$ (van der Vaart and Wellner, 1996, Chapter 2.7).

Pick any $p_{F,\boldsymbol{\Sigma}} \in \mathcal{Q}$, with $F = \sum_{h=1}^{\infty} z_h\delta_{z_h}$ and let the spectral decomposition of $\boldsymbol{\Sigma}^{-1}$ be $\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$ where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1,\ldots,\lambda_d)$ and $\boldsymbol{P}$ is an orthogonal matrix. Find $\hat{z}_1,\ldots,\hat{z}_H \in \hat{R}$, $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1,\ldots,\hat{\pi}_H) \in \hat{S}$, $\hat{\boldsymbol{P}} \in \hat{\mathcal{O}}$ and $\hat{m}_1,\ldots,\hat{m}_d \in \{1,\ldots,M\}$ such that

$$\max_{1\leq h\leq H}\|z_h - \hat{z}_h\| < \sigma_0\epsilon, \tag{2.18}$$

$$\sum_{h=1}^{H}|\tilde{\pi}_h - \hat{\pi}_h| < \epsilon, \text{ where } \tilde{\pi}_h = \frac{\pi_h}{1-\sum_{l>H}\pi_l}, 1\leq h\leq H, \tag{2.19}$$

$$\|\boldsymbol{P} - \hat{\boldsymbol{P}}\|_2 \leq \epsilon^2, \tag{2.20}$$

$$\hat{\lambda}_j = \{\sigma_0^2(1+\epsilon^2/d)^{\hat{m}_j - 1}\}^{-1} \text{ satisfies } 1\leq \hat{\lambda}_j/\lambda_j < 1+\epsilon^2/d,\;\; j=1,\ldots,d. \tag{2.21}$$

Take $\hat{F} = \sum_{h=1}^{H} \hat{\pi}_h \delta_{\hat{z}_h}$ and $\hat{\boldsymbol{\Sigma}} = (\hat{\boldsymbol{P}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{P}}^T)^{-1}$ where $\hat{\boldsymbol{\Lambda}} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_d)$. Also define $\tilde{\boldsymbol{\Sigma}} = (\hat{\boldsymbol{P}}\boldsymbol{\Lambda}\hat{\boldsymbol{P}}^T)^{-1}$ and $\boldsymbol{Q} = \hat{\boldsymbol{P}}^T\boldsymbol{P}$. By the triangle inequality

$$\|p_{F,\boldsymbol{\Sigma}} - p_{\hat{F},\hat{\boldsymbol{\Sigma}}}\|_1 \leq \|p_{F,\boldsymbol{\Sigma}} - p_{F,\hat{\boldsymbol{\Sigma}}}\|_1 + \|p_{F,\hat{\boldsymbol{\Sigma}}} - p_{\hat{F},\hat{\boldsymbol{\Sigma}}}\|_1. \tag{2.22}$$

The first term on the right hand side can be bounded by

$$\int \|\phi_{\boldsymbol{\Sigma}}(\cdot - \boldsymbol{z}) - \phi_{\hat{\boldsymbol{\Sigma}}}(\cdot - \boldsymbol{z})\|_1 dF(\boldsymbol{z}) = \|\phi_{\boldsymbol{\Sigma}} - \phi_{\hat{\boldsymbol{\Sigma}}}\|_1 \leq \|\phi_{\boldsymbol{\Sigma}} - \phi_{\tilde{\boldsymbol{\Sigma}}}\|_1 + \|\phi_{\tilde{\boldsymbol{\Sigma}}} - \phi_{\hat{\boldsymbol{\Sigma}}}\|_1.$$

By Kemperman's inequality (Kemperman, 1969), the total variation distance is bounded by $\sqrt{2}$ times the square root of the Kullback–Leibler divergence, we have

$$\|\phi_{\tilde{\boldsymbol{\Sigma}}} - \phi_{\hat{\boldsymbol{\Sigma}}}\|_1 \leq \{\mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\Sigma}}) - \log\det(\hat{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\Sigma}}) - d\}^{1/2}.$$

But

$$\mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\Sigma}}) = \mathrm{tr}(\hat{\boldsymbol{\Lambda}}\boldsymbol{\Lambda}^{-1}) = \sum_{j=1}^{d} \hat{\lambda}_j/\lambda_j < d + \epsilon^2$$

and $\det(\hat{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\Sigma}}) = \prod_{j=1}^{d}(\hat{\lambda}_j/\lambda_j) > 1$. Thus $\|\phi_{\tilde{\boldsymbol{\Sigma}}} - \phi_{\hat{\boldsymbol{\Lambda}}}\|_1 \leq \epsilon$. For the other term, we have

$$\|\phi_{\boldsymbol{\Sigma}} - \phi_{\tilde{\boldsymbol{\Sigma}}}\|_1 \leq \{\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}) - \log\det(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}) - d\}^{1/2} = \{\mathrm{tr}(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T\boldsymbol{\Lambda}^{-1} - \boldsymbol{I})\}^{1/2}$$

because $\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T\hat{\boldsymbol{P}}\boldsymbol{\Lambda}^{-1}\hat{\boldsymbol{P}}^T$ has determinant one and trace equal to that of $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T\boldsymbol{\Lambda}^{-1}$. Write $\boldsymbol{Q} = \boldsymbol{I} + \boldsymbol{B}$. Then

$$\|\boldsymbol{B}\|_{\max} \leq \|\boldsymbol{B}\|_2 = \|\hat{\boldsymbol{P}}^T\boldsymbol{P} - \boldsymbol{I}\|_2 = \|\boldsymbol{P} - \hat{\boldsymbol{P}}\|_2 \leq \delta$$

and hence

$$\begin{aligned}\mathrm{tr}(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T\boldsymbol{\Lambda}^{-1} - \boldsymbol{I}) &= \mathrm{tr}(\boldsymbol{B} + \boldsymbol{\Lambda}\boldsymbol{B}^T\boldsymbol{\Lambda}^{-1} + \boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}^T\boldsymbol{\Lambda}^{-1}) \\ &\leq 3d\|\boldsymbol{B}\|_{\max}\frac{\max(\lambda_1,\ldots,\lambda_d)}{\min(\lambda_1,\ldots,\lambda_d)} \leq \epsilon^2.\end{aligned}$$

Therefore the first term on the right hand side of (2.22) is bounded by 2ϵ. The last term of (2.22) equals to

$$\begin{aligned}&\left\|\sum_{h>H}\pi_h\phi_{\hat{\boldsymbol{\Sigma}}}(\cdot - \boldsymbol{z_h}) + \sum_{h=1}^{H}\pi_h\{\phi_{\hat{\boldsymbol{\Sigma}}}(\cdot - \boldsymbol{z_h}) - \phi_{\hat{\boldsymbol{\Sigma}}}(\cdot - \hat{\boldsymbol{z}}_{\boldsymbol{h}})\} + \sum_{h=1}^{H}(\pi_h - \hat{\pi}_h)\phi_{\hat{\boldsymbol{\Sigma}}}(\cdot - \hat{\boldsymbol{z}}_{\boldsymbol{h}})\right\|_1 \\ &\quad\leq \sum_{h>H}\pi_h + \sum_{h=1}^{H}\pi_h\|\phi_{\hat{\boldsymbol{\Sigma}}}(\cdot - \boldsymbol{z_h}) - \phi_{\hat{\boldsymbol{\Sigma}}}(\cdot - \hat{\boldsymbol{z}}_{\boldsymbol{h}})\|_1 + \sum_{h=1}^{H}|\pi_h - \hat{\pi}_h|.\end{aligned}$$

The first term above is smaller than ϵ and so is the second term because

$$\|\phi(\cdot - \boldsymbol{z_h}) - \phi(\cdot - \hat{\boldsymbol{z}}_{\boldsymbol{h}})\|_1 \leq \left(\frac{2}{\pi}\right)^{1/2}\|\hat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{z_h} - \hat{\boldsymbol{z}}_{\boldsymbol{h}})\| \leq \epsilon.$$

The last term is smaller than or equal to

$$(1 - \sum_{h>H}\pi_h)\sum_{h=1}^{H}|\tilde{\pi}_h - \hat{\pi}_h| + \sum_{h>H}\pi_h\sum_{h=1}^{H}\hat{\pi}_h \leq 2\epsilon.$$

Thus a 6ϵ-net of $\mathcal{Q}$, in the L_1-metric, can be constructed with $\hat{p} = p_{\hat{F},\hat{\boldsymbol{\Sigma}}}$ as above. The total number of such $\hat{p}$ is bounded by a multiple of $\{a/(\sigma_0\epsilon)\}^{dH}\,\epsilon^{-H}\delta^{-d(d-1)/2}M^d$. This proves the first assertion with $\rho = \|\cdot\|_1$, because $M\log(1+\epsilon^2/d) \lesssim M\epsilon^2$ and the constant factor by 6 can be absorbed in the bound. The same obtains when ρ is the Hellinger metric because it is bounded by the square-root of the L_1-metric.

For the second assertion, we consider the stick-breaking representation of a Dirichlet pro-

cess $F \sim \mathcal{D}_\alpha$ as described in Section 1.2.1:

$$F = \sum_{h=1}^{\infty} \pi_h \delta_{Z_h}, \quad \pi_h = V_h \prod_{j<h} (1 - V_j), \tag{2.23}$$

where $\{V_h, h \geq 1\}$ are independent beta distributed random variables with parameters 1 and $|\alpha| = \alpha(\mathbb{R}^d)$, $\{Z_h, h \geq 1\}$ are independently distributed according to the probability measure $\bar{\alpha} = \alpha/|\alpha|$ and these two sets of random variables are mutually independent. Hence $p_{F,\mathbf{\Sigma}} = \sum_{h=1}^{\infty} \pi_h \phi_\Sigma(\cdot - Z_h)$ with π_h and Z_h as described in (2.23). Therefore

$$\begin{aligned} \Pi(\mathcal{Q}^c) \;\leq\; & H\bar{\alpha}([-a,a]^d)^c + \mathrm{P}\left(\sum_{h>H} \pi_h > \epsilon\right) + \mathrm{P}\left\{\mathrm{eig}_d(\mathbf{\Sigma}^{-1}) > \sigma_0^{-2}\right\} \\ & + \mathrm{P}\left\{\mathrm{eig}_1(\mathbf{\Sigma}^{-1}) \leq \sigma_0^{-2}\left(1 + \frac{\epsilon^2}{d}\right)^{-M}\right\}. \end{aligned} \tag{2.24}$$

The first term is bounded by $b_1 H \exp(-C_1 a^{a_1})$ by assumption on α. As $W = -\sum_{h=1}^{H} \log(1 - V_h)$ is gamma distributed with parameters H and $|\alpha|$, we have

$$\mathrm{P}\left(\sum_{h>H} \pi_h > \epsilon\right) = \mathrm{P}\left(W < \log\frac{1}{\epsilon}\right) \leq \frac{(-|\alpha| \log \epsilon)^H}{\Gamma(H+1)} \leq \left(\frac{e|\alpha|}{H} \log\frac{1}{\epsilon}\right)^H$$

because $\Gamma(H+1) \geq (H/e)^H$. The last two terms are bounded by a multiple of

$$b_2 \exp\{-C_2 \sigma_0^{-2a_2}\} + b_3 \sigma_0^{-2a_3} \left(1 + \epsilon^2/d\right)^{-Ma_3}.$$

This proves the second assertion. □

The sieve defined here can easily adapt to different rates of convergence of the form $\epsilon_n = n^{-\gamma}(\log n)^{(d+1+s)/2}$ for $0 < \gamma \leq 1/2$ and $s > 0$. The extreme case $\gamma = 1/2$ corresponds to the class of Gaussian mixtures (Ghosal and van der Vaart, 2001). For a β-Hölder class convergence

rate we need to work with $\gamma = \beta/(2\beta + d^*)$. The following theorem makes this precise.

Theorem 7. *Fix a $\gamma \in (0, 1/2)$ and a pair of numbers $t > t_0 \geq (d+1)/2$. For $n \geq 1$, take $\epsilon_n = n^{-\gamma}(\log n)^t$, $\tilde{\epsilon}_n = n^{-\gamma}(\log n)^{t_0}$ and define $\mathcal{F}_n$ as $\mathcal{Q}$ in* (2.17) *with $\epsilon = \epsilon_n$, $H = \lfloor n\epsilon_n^2/(\log n)\rfloor$, $M = a^{a_1} = \sigma_0^{-2a_2} = n$. Then $\mathcal{F}_n$ satisfies* (1.9) *and* (1.10) *for all large n for some $c_1, c_3 > 0$ and every $c_2 > 0$.*

Proof. By Proposition 2,

$$\begin{aligned}\log N(\bar{\epsilon}_n, \mathcal{F}_n, \rho) &\leq K\{dn^{1-2\gamma}(\log n)^{2t} + n^{1-2\gamma}(\log n)^{2t} + \log n + n^{1-2\gamma}(\log n)^{2t}\} \\ &\leq c_1 n^{1-2\gamma}(\log n)^{2t} = c_1 n\epsilon_n^2\end{aligned}$$

for some $c_1 > 0$ and hence (1.9) holds. By the second assertion of the same proposition,

$$\begin{aligned}(\mathcal{D}_\alpha \times G)(\mathcal{F}_n^c) &\leq b_1 n^{1-2\gamma}(\log n)^{2t-1}e^{-b_1 n} + n^{-(1-2\gamma)n^{1-2\gamma}(\log n)^{2t-1}} \\ &\quad + \quad b_2 e^{-C_2 n} + b_3 n^{a_3/a_2} e^{-2a_3 n\log(1+\bar{\epsilon}_n^2/d)} \\ &\leq c_3 e^{-(1-2\gamma)n^{1-2\gamma}(\log n)^{2t}} \leq c_3 e^{-(c_2+4)n^{1-2\gamma}(\log n)^{2t_0}}\end{aligned}$$

for all large n, some $c_3 > 0$ and every $c_2 > 0$. □

2.5 Supplementary results

Theorem 8. *Let P_0 be a probability measure on $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq a\} \subset \mathbb{R}^d$. For any $\varepsilon > 0$ and $\sigma > 0$, there is a discrete probability measure F_σ on $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq a\}$ with at most $N_{\sigma,\varepsilon} = D[\{(a/\sigma) \vee 1\}\log(1/\varepsilon)]^d$ support points such that $\|p_{P_0,\sigma} - p_{F_\sigma,\sigma}\|_\infty \lesssim \varepsilon/\sigma^d$ and $\|p_{P_0,\sigma} - p_{F_\sigma,\sigma}\|_1 \lesssim \varepsilon\{\log(1/\varepsilon)\}^{1/2}$, for some universal constant D.*

Proof. The proof is an extension of Lemma 2 of Ghosal and van der Vaart (2007b) and Lemma 3.1 of Ghosal and van der Vaart (2001) to d dimensions. We consider a larger support $[-a, a]^d$ rather than $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq a\}$. We can partition $[-a, a]^d$ into k disjoint, consecutive regions $I_1, \ldots, I_k$ of size σ^d and k' final pieces $I_{k+1}, \ldots, I_{k+k'}$ of size smaller than σ^d, where $k = (\lfloor 2a/\sigma \rfloor)^d$ and $k' = (\lfloor 2a/\sigma \rfloor + 1)^d - (\lfloor 2a/\sigma \rfloor)^d$. We may write $P_0 = \sum_{i=1}^{k+k'} P_0(I_i) P_{0,i}$, where each $P_{0,i}$ is a probability measure concentrated on I_i. Hence $p_{P_0,\sigma} = \sum_{i=1}^{k+k'} P_0(I_i) p_{P_{0,i},\sigma}$. Let $\boldsymbol{Z_i}$ be a random variable distributed according to $P_{0,i}$, and for $\boldsymbol{a_i}$ the left endpoint of I_i, let G_i be the law of $(\boldsymbol{Z_i} - \boldsymbol{a_i})/\sigma$. Thus G_i is a probability measure defined on $[0, 1]^d$ for $i = 1, \ldots, k$ and on subsets of $[0, 1]^d$ for $i = (k + 1), \ldots, (k + k')$.

Now, apply the same arguments in Lemma 3.1 of Ghosal and van der Vaart (2001) under a multi-dimensional setting, for any probability distribution Q on $[0, 1]^d$, there exists a discrete distribution Q' with at most $\{(2t - 2)^d + 1\}$ support points such that

$$\begin{aligned} \|p_{Q,1} - p_{Q',1}\|_\infty &\lesssim \epsilon \\ \|p_{Q,1} - p_{Q',1}\|_1 &\lesssim \epsilon\{\log(1/\epsilon)\}^{1/2}. \end{aligned} \tag{2.25}$$

The number of support points is determined by matching up the mixed moments $z_1^{l_1} z_2^{l_2} \cdots z_d^{l_d}$ for every $1 \leq l_i \leq 2t - 2$, $i = 1, \ldots, d$, for $t = \lfloor 2a/\sigma \rfloor$. This power of d propagate all through the require extensions. Now, for every G_i, we can find a G_i' such that (2.25) holds for $Q = G_i$ and $Q' = G_i'$. Let $P_{0,i}'$ has the law of $\boldsymbol{a_i} + \sigma W_i'$ if W_i' has law G_i' and set $F_\sigma = \sum_{i=1}^{k+k'} P_0(I_i) P_{0,i}'$, we have

$$\begin{aligned} \|p_{P_{0,i},\sigma} - p_{P_{0,i}',\sigma}\|_\infty &= \sigma^{-d} \|p_{Q,1} - p_{Q',1}\|_\infty \lesssim \varepsilon/\sigma^d, \\ \|p_{P_{0,i},\sigma} - p_{P_{0,i}',\sigma}\|_1 &= \|p_{Q,1} - p_{Q',1}\|_1 \lesssim \varepsilon\{\log(1/\varepsilon)\}^{1/2}, \end{aligned} \tag{2.26}$$

and hence

$$\begin{aligned} \|p_{P_0,\sigma} - p_{F_\sigma,\sigma}\|_\infty &\lesssim \varepsilon/\sigma^d, \\ \|p_{P_0,\sigma} - p_{F_\sigma,\sigma}\|_1 &\lesssim \varepsilon\{\log(1/\varepsilon)\}^{1/2}. \end{aligned} \tag{2.27}$$

The total number of support points of F_σ is bounded by

$$(k+k')\{\log(1/\epsilon)\}^d \lesssim [\{(a/\sigma)\vee 1\}\log(1/\varepsilon)]^d.$$

Hence the proof is complete. □

Corollary 1. *Let P_0 be a probability measure on $\{\boldsymbol{x}\in\mathbb{R}^d : \|\boldsymbol{x}\|\le a\}$. For any $\varepsilon>0$ and $\sigma>0$, there is a discrete probability measure F_σ^* on $\{\boldsymbol{x}\in\mathbb{R}^d : \|\boldsymbol{x}\|\le a\}$ with at most $N_{\sigma,\varepsilon} = D[\{(a/\sigma)\vee 1\}\log(1/\varepsilon)]^d$ support points from the set*

$$\{(n_1,\ldots,n_p)\sigma\varepsilon : n_i\in\mathbb{Z}, |n_i| < \lceil a/(\sigma\varepsilon)\rceil, i=1,\ldots,p\},$$

such that

$$\|p_{P_0,\sigma} - p_{F_\sigma^*,\sigma}\|_\infty \lesssim \varepsilon/\sigma^d, \quad \|p_{P_0,\sigma} - p_{F_\sigma^*,\sigma}\|_1 \lesssim \varepsilon\{\log(1/\varepsilon)\}^{1/2}.$$

Proof. First get F_σ as in Theorem 8 and then move each of its support points to the nearest point on the grid $\{(n_1,\ldots,n_p)\sigma\varepsilon : n_i\in\mathbb{Z}, |n_i| < \lceil a/(\sigma\varepsilon)\rceil, i=1,\ldots,p\}$ to get F_σ^*. These moves cost at most a constant times ϵ^2/σ^d to the supremum distance and at most a constant times ϵ to the L_1 distance. □

Lemma 4. *Let $V_0, V_1, \ldots, V_N$ be a partition of $\mathbb{R}^d$ and $F' = \sum_{j=1}^N p_j\delta_{z_j}$ a probability measure*

on $\mathbb{R}^d$ *with* $\boldsymbol{z_j} \in V_j$, $j = 1, \ldots, N$. *Then, for any probability measure* F *on* $\mathbb{R}^d$, *and any* $\sigma > 0$,

$$\|p_{F,\sigma} - p_{F',\sigma}\|_\infty \lesssim \frac{1}{\sigma^{d+1}} \max_{1 \le j \le N} \operatorname{diam}(V_j) + \frac{1}{\sigma^d} \sum_{j=1}^N |F(V_j) - p_j|,$$

$$\|p_{F,\sigma} - p_{F',\sigma}\|_1 \lesssim \frac{1}{\sigma} \max_{1 \le j \le N} \operatorname{diam}(V_j) + \sum_{j=1}^N |F(V_j) - p_j|,$$

where $\operatorname{diam}(A) = \sup\{\|\boldsymbol{z_1} - \boldsymbol{z_2}\| : \boldsymbol{z_1}, \boldsymbol{z_2} \in A\}$ *denotes the diameter of a set* A.

Proof. The proof is an extension of Lemma 5 of Ghosal and van der Vaart (2007b) to d dimensions. We bound $p_{F,\sigma} - p_{F',\sigma}$ by

$$\begin{aligned} \int_{V_0} \phi_{\sigma^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{z}) dF(\boldsymbol{z}) \quad &+ \quad \sum_{j=1}^N \int_{V_j} (\phi_{\sigma^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{z}) - \phi_{\sigma^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{z_j}) dF(\boldsymbol{z}) \\ &+ \quad \sum_{j=1}^N \phi_{\sigma^2 \boldsymbol{I}}(\boldsymbol{x} - \boldsymbol{z_j})(F(V_j) - p_j). \end{aligned}$$

The result now follows because

$$F(V_0) = 1 - \sum_{j=1}^N F(V_j) \le \sum_{j=1}^N |F(V_j) - p_j|, \tag{2.28}$$

$$\|\phi_{\sigma^2 \boldsymbol{I}}\|_\infty \lesssim \sigma^{-d}, \tag{2.29}$$

$$\|\phi_{\sigma^2 \boldsymbol{I}}\|'_\infty \lesssim \sigma^{-(d+1)}, \tag{2.30}$$

$$\|\phi_{\sigma^2 \boldsymbol{I}}(\cdot - \boldsymbol{z}) - \phi_{\sigma^2 \boldsymbol{I}}(\cdot - \boldsymbol{z'})\|_1 \lesssim \sigma^{-1} \|\boldsymbol{z} - \boldsymbol{z'}\|. \tag{2.31}$$

□

Lemma 5. *There is a* $\lambda_0 \in (0, 1)$ *such that for any two probability measures* P, Q *with densities*

p, q and any $\lambda \in (0, \lambda_0)$

$$P \log \frac{p}{q} \leq d_H^2(p,q) \left(1 + 2 \log \frac{1}{\lambda}\right) + 2P \left\{ \left(\log \frac{p}{q} \right) \mathbb{1} \left(\frac{q}{p} \leq \lambda \right) \right\},$$
$$P \left(\log \frac{p}{q} \right)^2 \leq d_H^2(p,q) \left\{ 12 + 2 \left(\log \frac{1}{\lambda} \right)^2 \right\} + 8P \left\{ \left(\log \frac{p}{q} \right)^2 \mathbb{1} \left(\frac{q}{p} \leq \lambda \right) \right\}.$$

Proof. Our proof follows the argument presented in the proof of Lemma 7 of Ghosal and van der Vaart (2007b). The function $r : (0, \infty) \to \mathbb{R}$ defined implicitly by

$$\log x = 2(\sqrt{x} - 1) - r(x)(\sqrt{x} - 1)^2$$

is non-negative and decreasing, and there exists a $\lambda_0 > 0$ such that $r(x) \leq 2 \log(1/x)$ for all $x \in (0, \lambda_0)$. Using these properties and $d_H^2(p,q) = -2P\{\sqrt{q/p} - 1\}$ we obtain

$$\begin{aligned} P \log \frac{p}{q} &= d_H^2(p,q) + P \left\{ r \left(\frac{q}{p} \right) \left(\sqrt{\frac{q}{p}} - 1 \right)^2 \right\} \\ &\leq d_H^2(p,q) + r(\lambda) d_H^2(p,q) + P \left\{ r \left(\frac{q}{p} \right) \mathbb{1} \left(\frac{q}{p} < \lambda \right) \right\} \\ &\leq d_H^2(p,q) + 2 \left(\log \frac{1}{\lambda} \right) d_H^2(p,q) + 2P \left\{ \left(\log \frac{p}{q} \right) \mathbb{1} \left(\frac{q}{p} < \lambda \right) \right\} \end{aligned}$$

for any $\lambda < \lambda_0$, proving the first inequality of the Lemma.

To prove the other inequality, note that $|\log x| \leq 2|\sqrt{x} - 1|$, $x \geq 1$ and so

$$P \left\{ \left(\log \frac{p}{q} \right)^2 \mathbb{1} \left(\frac{q}{p} \geq 1 \right) \right\} \leq 4P \left(\sqrt{\frac{q}{p}} - 1 \right)^2 = 4 d_H^2(p,q).$$

On the other hand,

$$
\begin{aligned}
P\left\{\left(\log\frac{p}{q}\right)^2 \mathbb{1}\left(\frac{q}{p}\le 1\right)\right\} & \\
&\le 8P\left(\sqrt{\frac{q}{p}}-1\right)^2 + 2P\left\{r^2\left(\frac{q}{p}\right)\left(\sqrt{\frac{q}{p}}-1\right)^4 \mathbb{1}\left(\frac{q}{p}\le 1\right)\right\} \\
&\le 8d_H^2(p,q) + 2r^2(\lambda)P\left(\sqrt{\frac{q}{p}}-1\right)^2 + 2P\left\{r^2\left(\frac{q}{p}\right)\mathbb{1}\left(\frac{q}{p}\le \lambda\right)\right\} \\
&\le 8d_H^2(p,q) + 2\left(\log\frac{1}{\lambda}\right)^2 d_H^2(p,q) + 8P\left\{\left(\log\frac{p}{q}\right)^2 \mathbb{1}\left(\frac{q}{p}\le \lambda\right)\right\}.
\end{aligned}
$$

This completes the proof. □

Lemma 6. *Let $\mathcal{A}$ be an index set and $\mathfrak{X}$ be a metric space. Suppose that $\{p_\alpha\}_{\alpha\in\mathcal{A}}$ and $\{q_\alpha\}_{\alpha\in\mathcal{A}}$ are collections of probability density functions on $\mathfrak{X}$ with respect to a dominating measure ν. Then for any probability measure G on $\mathcal{A}$, $d_H^2(\int p_\alpha dG, \int q_\alpha dG) \le \int d_H^2(p_\alpha, q_\alpha)dG$. In particular, for any three densities p, q and ϕ on $\mathbb{R}^d$, $d_H(\phi * p, \phi * q) \le d_H(p,q)$.*

Proof. By the Cauchy–Schwartz inequality,

$$
\begin{aligned}
1-\frac{1}{2}\int d_H^2(p_\alpha,q_\alpha)dG &= \int\int \{p_\alpha(x)q_\alpha(x)\}^{1/2}\nu(dx)G(d\alpha) \\
&\le \int\left\{\int p_\alpha(x)G(d\alpha)\int q_\alpha(x)G(d\alpha)\right\}^{1/2}\nu(dx) \\
&= 1-\frac{1}{2}d_H^2\left(\int p_\alpha dG, \int q_\alpha dG\right).
\end{aligned}
$$

This gives the first result. The second assertion holds by choosing $\mathcal{A} = \mathfrak{X} = \mathbb{R}^d$, $p_\alpha(x) = p(x-\alpha)$, $q_\alpha(x) = q(x-\alpha)$ and $G(d\alpha) = \phi(\alpha)d\alpha$. □

2.6 Proof of Theorems 3 and 4

Proof of Theorem 3. In Theorem 6 we show that (1.11) holds for the Dirichlet mixture prior Π defined in Section 2.2.1 with $\tilde{\epsilon}_n = n^{-\beta/(2\beta+d^*)}(\log n)^{t_0}$ where $t_0 = \{d^*(1+1/\tau+1/\beta)+1\}/(2+d^*/\beta)$. In Theorem 7 we show that (1.9), (1.10) hold with $\tilde{\epsilon}_n$ as before and $\epsilon_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$ for every $t > t_0$. □

Proof of Theorem 4. The proof is similar to that of Theorem 3. We define a sieve by

$$\mathcal{F}_n = \left\{ f(\boldsymbol{x}) = \sum_{h=1}^{H} \omega_h \phi_\Sigma(\boldsymbol{x}-\mu_h) : \begin{array}{l} \mu_h \in [-a,a]^d, h \le H; \\ \sigma_0^2 \le \mathrm{eig}_j(\Sigma) < \sigma_0^2 \left(1+\epsilon^2/d\right)^M,\ j=1,\ldots,d \end{array} \right\},$$

where $H = \lfloor n\epsilon_n^2/(\log n) \rfloor$, $M = a^{a_1} = \sigma_0^{-2a_2} = n$. Note that Proposition 2 can still be applied with a little modification. Define $\tilde{\epsilon}_n = n^{-\beta/(2\beta+d^*)}(\log n)^{t_0}$ where $t_0 = \{d^*(1+\tau^{-1}+\beta^{-1})+1\}/(2+d^*\beta^{-1})$ and $\epsilon_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$ for every $t > t_0 + \max\{0,(1-\tau_1)/2\}$. Then (1.9) holds because

$$\begin{aligned} \log N(\epsilon_n, \mathcal{F}_n, \rho) &\le K\{dH\log n + H\log n + \log n + n\epsilon_n^2\} \\ &\le K\{dn\epsilon_n^2 + \log n + n\epsilon_n^2\} \\ &\le c_1 n\epsilon_n^2 \end{aligned} \tag{2.32}$$

for some constants $c_1, K > 0$. Also, (1.10) holds since

$$\begin{aligned} \Pi_{\mathrm{FM}}(\mathcal{F}_n^c) &\le H\Pi(\mu_i \notin [-a,a]^d, 1 \le i \le H) + b_5 \exp\{-C_2 H(\log H)^{\tau_1}\} \\ &\quad + b_2 \exp\{-C_2 n\} + b_3 n^{a_3/a_2}(1+\epsilon_n^2/d)^{-2na3} \\ &\le c_3 \exp\{-(c_2+4)n\tilde{\epsilon}_n^2\} \end{aligned} \tag{2.33}$$

for some constants $c_2, c_3, C_1, C_2 > 0$.

For (1.11), note that the finite mixture approximation used in Theorem 6 is still valid here. Therefore $\Pi_{\text{FM}}\{\mathcal{K}(f_0, \tilde{\epsilon}_n)\} \geq \mathrm{P}_{F_\epsilon}(\mathcal{P}_\sigma)G(\mathcal{S}_\sigma)$, where $\mathcal{P}_\sigma$ and $\mathcal{S}_\sigma$ are defined in the proof of Theorem 6. Observe $\mathrm{P}_{F_\epsilon}(\mathcal{P}_\sigma) \gtrsim \exp\{-C_3 M \log(1/\tilde{\epsilon}_n)\}$ by the assumption on Π_{FM}, where M is the number of support points of the finite mixture and satisfies $M \leq \tilde{\epsilon}_n^{-d/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{sd}$, $s = 1 + \beta^{-1} + \tau^{-1}$ according to the same arguments in the proof of Theorem 6. Also, we are using the same prior on Σ, hence the calculation of $G(\mathcal{S}_\sigma) \gtrsim \exp\{-C_4\tilde{\epsilon}_n^{-\kappa/\beta}(\log(1/\tilde{\epsilon}_n))^{s\kappa+1)}\}$ remains the same. As a result, for some constants $C_3, C_4, C_5, c_4 > 0$,

$$\begin{aligned}\Pi_{\text{FM}}\{\mathcal{K}(f_0, \tilde{\epsilon}_n)\} &\gtrsim \exp\{-C_3\tilde{\epsilon}_n^{-d/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{sd}\log(1/\tilde{\epsilon}_n) - C_4\tilde{\epsilon}_n^{-\kappa/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{s\kappa+1}\} \\ &\geq c_4\exp\{-C_5 n\tilde{\epsilon}_n^2\}\end{aligned}$$

given $\tilde{\epsilon}_n^{-d^*/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{sd^*+1} \leq n\tilde{\epsilon}_n^2$. The result follows. □

2.7 Anisotropic Hölder functions

Anisotropic functions are those that have different orders of smoothness along different axes. The isotropic result presented before gives adaptive rates corresponding to the least smooth direction. Sharper results can be obtained by explicitly factoring in the anisotropy. For a $\beta > 0$, an $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in (0, \infty)^d$ with $\boldsymbol{\alpha}. = d$ and an $L : \mathbb{R}^d \to (0, \infty)$ satisfying $L(\boldsymbol{x} + \boldsymbol{y}) \leq L(\boldsymbol{x})\exp(\tau_0\|\boldsymbol{y}\|_1^2)$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and some $\tau_0 > 0$, the $\boldsymbol{\alpha}$-anisotropic β-Hölder class with envelope L is defined as the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ that have continuous mixed partial derivatives $D^{\boldsymbol{k}}f$ of all orders $\boldsymbol{k} \in \mathbb{N}_0^d$, $\beta - \alpha_{\max} \leq \langle \boldsymbol{k}, \boldsymbol{\alpha}\rangle < \beta$, with

$$|D^{\boldsymbol{k}}f(\boldsymbol{x} + \boldsymbol{y}) - D^{\boldsymbol{k}}f(\boldsymbol{x})| \leq L(\boldsymbol{x})e^{\tau_0\|\boldsymbol{y}\|_1^2}\sum_{j=1}^{d}|y_j|^{\min(\beta/\alpha_j - k_j, 1)}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d,$$

where $\alpha_{\max} = \max(\alpha_1, \ldots, \alpha_d)$. We denote this set of functions by $\mathcal{C}^{\boldsymbol{\alpha},\beta,L,\tau_0}(\mathbb{R}^d)$. Here β refers to the mean smoothness and $\boldsymbol{\alpha}$ refers to the anisotropy index. An $f \in \mathcal{C}^{\boldsymbol{\alpha},\beta,L,\tau_0}$ has partial derivatives of all orders up to $\lfloor \beta_j \rfloor$ along axis j where $\beta_j = \beta/\alpha_j$, and β is the harmonic mean $d/(\beta_1^{-1} + \cdots + \beta_d^{-1})$ of these axial smoothness coefficients. In the special case of $\boldsymbol{\alpha} = (1, \ldots, 1)$, the anisotropic set $\mathcal{C}^{\boldsymbol{\alpha},\beta,L,\tau_0}(\mathbb{R}^d)$ equals the isotropic set $\mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$.

Theorem 9. *Suppose that $f_0 \in \mathcal{C}^{\boldsymbol{\alpha},\beta,L,\tau_0}(\mathbb{R}^d)$ is a probability density function satisfying*

$$\int f_0 \left(|D^{\boldsymbol{k}} f_0| / f_0 \right)^{(2\beta+\epsilon)/\langle \boldsymbol{k}, \boldsymbol{\alpha} \rangle} < \infty, \;\; \boldsymbol{k} \in \mathbb{N}_0^d, \langle \boldsymbol{k}, \boldsymbol{\alpha} \rangle < \beta, \qquad \int f_0 \left(L/f_0 \right)^{(2\beta+\epsilon)/\beta} < \infty$$

for some $\epsilon > 0$ and that (2.9) *holds for some constants $a, b, c, \tau > 0$. If Π is as in Section 2.2.1, then the posterior convergence rate at f_0 in the Hellinger or the L_1-metric is $\epsilon_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$, where $t \geq \{d^*(1+\tau^{-1}+\beta^{-1})+1\}/(2+d^*/\beta)$, and $d^* = \max(d, \kappa\alpha_{\max})$.*

The proof is similar to those of the results presented in Section 2.6, except that to obtain an approximation to f_0, we replace the single bandwidth σ with bandwidth σ^{α_j} along the j-th axis. An f_0 satisfying the conditions of the above theorem also satisfies the conditions of Theorem 3 with smoothness index $\beta/\alpha_{\max}$, which is strictly smaller than β as long as not all α_j are equal to 1. Therefore when the true density is anisotropic, Theorem 9 indeed leads to a sharper convergence rate result.

Proof. For any $\sigma > 0$, define the transformation $T_{\boldsymbol{\alpha},\beta,\sigma}$ on $\mathcal{C}^{\boldsymbol{\alpha},\beta,L,\tau_0}(\mathbb{R}^d)$ as

$$T_{\boldsymbol{\alpha},\beta,\sigma} f = f - \sum_{\boldsymbol{k} \in \mathbb{N}_0^d:\, 1 \leq \langle \boldsymbol{k}, \boldsymbol{\alpha} \rangle < \beta} d_k \sigma^{\langle \boldsymbol{k}, \boldsymbol{\alpha} \rangle} f. \tag{2.34}$$

Also define $K_{\boldsymbol{\alpha},\sigma} f$ as the convolution of f and the normal density with mean zero and variance $\mathrm{diag}(\sigma^{2\alpha_1}, \ldots, \sigma^{2\alpha_d})$. The anisotropic analog of Lemma 3 is that there exists a constant $M_{\boldsymbol{\alpha},\beta}$

such that for any $f \in \mathcal{C}^{\boldsymbol{\alpha},\beta,L,\tau_0}$, any $\sigma \in (0, 1/(2\tau_0)^{1/2\alpha\max})$ and all $\boldsymbol{x} \in \mathbb{R}^d$,

$$|\{K_{\boldsymbol{\alpha},\sigma}(T_{\boldsymbol{\alpha},\beta,\sigma}f) - f\}(\boldsymbol{x})| < M_{\boldsymbol{\alpha},\beta}L(\boldsymbol{x})\sigma^\beta.$$

This follows along the lines of our proof of Lemma 3 starting from the anisotropic Taylor approximation

$$f(\boldsymbol{x}+\boldsymbol{y}) - f(\boldsymbol{x}) = \sum_{1 \le \langle \boldsymbol{k},\boldsymbol{\alpha}\rangle < \beta} \frac{(-\boldsymbol{y})^{\boldsymbol{k}}}{\boldsymbol{k}!}(D^{\boldsymbol{k}}f)(\boldsymbol{x}) + R(\boldsymbol{x},\boldsymbol{y}),$$

where the residual $R(\boldsymbol{x},\boldsymbol{y})$ in absolute value is bounded by a sum over terms of the form

$$\begin{aligned}
&\frac{|\boldsymbol{y}|^{\boldsymbol{k}}}{\boldsymbol{k}!}|(D^{\boldsymbol{k}}f)(x_1,\ldots,x_{j-1},x_j+\xi_j,x_{j+1}+y_{j+1},\ldots,x_d+y_d) \\
&\qquad\qquad - (D^{\boldsymbol{k}}f)(x_1,\ldots,x_{j-1},x_j,x_{j+1}+y_{j+1},\ldots,x_d+y_d)| \\
&\le L(\boldsymbol{x})\exp(\tau_0\|\boldsymbol{y}\|_1^2)|\boldsymbol{y}|^{\boldsymbol{k}}|y_j|^{\min(\beta/\alpha_j - k_j, 1)}/\boldsymbol{k}!
\end{aligned}$$

with j such that $\beta > \langle k,\alpha\rangle > \beta - \alpha_j$. Consequently, for some constant K_1,

$$\int |R(\boldsymbol{x},\boldsymbol{y})|\phi_{\mathrm{diag}(\sigma^{2\boldsymbol{\alpha}})}(\boldsymbol{y})d\boldsymbol{y} \le K_1 L(\boldsymbol{x})\sigma^\beta.$$

The rest of the argument in our proof of Lemma 3 goes through. The pointwise error bound between f_0 and $K_{\boldsymbol{\alpha},\sigma}(T_{\boldsymbol{\alpha},\beta,\sigma}f)$ then leads to exact analogs of Theorem 5 and Proposition 1, giving a $\tilde{h}_\sigma$ with support inside $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \le a_0\{\log(1/\sigma)\}^\tau\}$ satisfying $d_H(f_0, K_{\alpha,\sigma}\tilde{h}_\sigma) \le K_0\sigma^\beta$ for some constant K_0. Next the arguments in the proof of Theorem 6 can be replicated, with $\mathcal{P}_\sigma$ built around a discrete $F_\sigma = \sum_{j=1}^N p_j\delta_{z_j}$ with $N \le D_1\sigma^{-d}\{\log(1/\tilde{\epsilon}_n)\}^{d+d/\tau}$ support points such that $d_H(K_{\alpha,\sigma}\tilde{h}_\sigma, K_{\alpha,\sigma}F_\sigma) \le A_1\tilde{\epsilon}_n^{b_1}\{\log(1/\tilde{\epsilon}_n)\}^{1/4}$. We also need to define $\mathcal{S}_\sigma$ as the set of

$\boldsymbol{\Sigma}$ such that

$$\sigma^{-2\alpha_j} \leq \mathrm{eig}_j(\boldsymbol{\Sigma}^{-1}) \leq \sigma^{-2\alpha_j}(1+\sigma^{2\boldsymbol{\beta}})$$

for each $j = 1, \ldots, d$. Then

$$G(\mathcal{S}_\sigma) \geq C_3 \exp[-c_3 \tilde{\epsilon}_n^{-\kappa\alpha_{\max}/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{s\kappa+1}],$$

which contributes the $\kappa\alpha_{\max}$ term in $d^* = \max(d, \kappa\alpha_{\max})$. □

With the standard inverse Wishart prior G, we have $\kappa = 2$ and consequently the optimal rate $n^{-\beta/(2\beta+d)}$ is recovered up to a $\log n$ factor only when $\alpha_{\max} \leq d/2$. Therefore in a two dimensional case only the isotropic case is addressed and for higher dimensions we get optimal results for a limited amount of anisotropy. But, when $\kappa \leq 1$, as in the case of a diagonal $\boldsymbol{\Sigma}$ with squared inverse gamma diagonal components, Theorem 9 provides optimal rates for any dimension and any degree of anisotropy because $\alpha_{\max}$ can never exceed d.

Chapter 3

Univariate function estimation using random series prior

3.1 Introduction

In the nonparametric Bayesian literature, a prior on a function is usually constructed via a stochastic process, (e.g., a Gaussian process prior in van der Vaart and van Zanten (2008, 2009)) or by expanding a function in a basis of functions or by convoluting a kernel with a random measure. In this chapter, we focus on the basis expansion approach by putting a prior on the coefficients of basis functions and the number of basis functions. There are several advantages of using a prior directly on model dimension J rather than on the smoothness level as in the construction of an adaptive procedure described in Ghosal et al. (2008). First, this bypasses the need for specifying the optimal dimension $J_{n,\alpha}$ in practice, which are given by the posterior convergence theorem only up to a constant multiple. In a sense, assigning a prior on J is more natural from a Bayesian point of view and is a common method used by practitioners (Denison et al., 1998). More importantly, assigning a prior directly on J allows us to obtain adaptation

for all smoothness levels in an interval rather than for only a countable number of smoothness levels.

A similar idea was used in Babenko and Belitser (2010) for the infinite dimensional normal model, which is equivalent to a white noise model. They introduced the idea of the oracle dimension for every parameter value defined by the minimizer of risk in a class of estimation problems induced by the dimension. The oracle dimension is used to slice the parameter space into different smoothness levels. They assigned a prior distribution on the oracle dimension and the projection of the infinite dimensional mean vector on the finite dimensional oracle. They showed that the risk of the Bayes estimator satisfies some desirable oracle inequalities, which lead to a complete adaptation of the Bayes estimator. Interestingly, the oracle inequality gives adaptation simultaneously for many different families of smoothness classes such as Sobolev ellipsoids or hyper-rectangles. Although a random series with a random number of terms has been used by practitioners in various applied models, theoretical properties of such priors have been studies only very recently. Rivoirard and Rousseau (2012) exclusively considered the density estimation problem using a wavelet basis. de Jonge and van Zanten (2012) considered a general class of inference problems using spline basis and Gaussian priors on coefficients, and hence the resulting priors are mixtures of finite dimensional Gaussian processes.

We suggest a unified treatment for almost any curve estimation problem that leads to adaptation. We formulate one general theorem in an abstract setting suitable as a prelude for many different inference problems where we allow arbitrary basis functions and arbitrary multivariate distributions on the coefficients of the expansion. Thus the resulting process induced on the function need not be Gaussian, and can accommodate a variety of functions starting from one with a bounded support to one with a heavy tail. The resulting rate obtained in the abstract theorem depends on the smoothness of the underlying function, approximation ability of the basis expansion used, tail of the prior distribution on the coefficients, prior on the number of

terms in the series expansion, prior concentration rate and the metrics being used. The general theorem then gives rise to adaptive posterior convergence rates for different estimation problems. We illustrate the implication of the abstract theorem for the white noise model, density estimation on compact interval or the real line, nonparametric normal, Poisson and binary regression, spectral density estimation of a stationary time series and some functional regression models.

We show that the use of B-splines leads to MCMC-free computation if the prior on the coefficients is chosen appropriately. B-splines have been well studied by mathematicians (de Boor, 2001) and have been used in statistics as well; see Truong et al. (2005) for details. Nonnegativity, near orthogonality, summation to one and the possibility of choosing arbitrarily high smoothness level of B-spline functions are collectively the reasons for the popularity of the B-spline basis. The idea is to approximate a function that is of interest as a linear combination of the spline functions. Then the estimation of the function becomes equivalent to the estimation of the coefficients in the B-spline basis expansion (Truong et al., 2005). In some cases, we do not approximate the true function directly. Instead, we consider a transformation that is needed to satisfy certain constraints. For example, for density estimation, an exponential transformation along with a normalization step is often used as the functions are required to be nonnegative and integrate to one (Stone, 1990; Ghosal et al., 2000).

We also discuss adaptation on Sobolev and Besov spaces besides commonly used Hölder spaces to quantify smoothness. For nonparametric normal regression, we also remove a commonly used condition that the variance is bounded away form zero. Further, for the density estimation problem, we allow a flexible general link function instead of only the exponential link. Such a flexible link function approach is also carried out for binary, Poisson regression and spectral density estimation problems. By restricting coefficients to appropriate subsets like $(0, 1)$ or $(0, \infty)$, we may not even need to use a link function. We show that the additional

flexibility can be utilized to develop MCMC-free computing strategies.

The finite random series prior may be regarded as a valuable alternative to the Gaussian process prior. The Gaussian process has received a lot of attention in the literature on all aspects, from modeling a prior distribution (Leonard, 1978; Lenk, 1988) to computation (Tokdar, 2007) to asymptotics (Tokdar and Ghosh, 2007; Ghosal and Roy, 2006; Choi and Schervish, 2007; van der Vaart and van Zanten, 2007, 2008, 2009; Castillo, 2008, 2012), to applications in spatial statistics (Banerjee et al., 2008) and elsewhere. Asymptotic properties of posterior distributions based on Gaussian process priors are primarily driven by the structure of its reproducing kernel Hilbert space. While the elegant result of van der Vaart and van Zanten (2009) established that appropriately randomly rescaled Gaussian processes lead to posterior that automatically adapts to the unknown smoothness, in this paper we show that the same property also holds for random series priors using relatively elementary techniques. Computationally, Gaussian processes are relatively difficult to deal with. Except for Gaussian Markov random fields for which the integrated nested Laplace approximation method has been developed (Rue et al., 2009), the general approach to computation is to approximate the given Gaussian process by one that is generated by finitely many normal variables, obtained by conditioning the original process at a number of knots, which needs to be sufficiently large (Tokdar, 2007). The approach adaptively chooses the knots through a reversible jump Markov chain Monte-Carolo (RJMCMC). While the RJMCMC can also be used for random series prior, we shall describe a method that uses a conjugate-like prior for the model and hence avoids the use of MCMC as the posterior can be represented analytically. When the sample size n is relatively small (e.g. $n = 10$), the exact values of posterior moments can be computed. When the sample size is large, we use a direct sampling strategy. Thus at least conceptually, the random series prior gives rise to a more straightforward approach to computation. It may be noted that Gaussian process and random series priors are intimately related in two ways — a normal prior on the

coefficients of a random series gives a Gaussian process (conditional on the number of terms) while the Karhunen-Loève expansion of a Gaussian process expresses itself as a random series with basis consisting of eigenfunctions of the covariance kernel of the Gaussian process. Thus random series may be regarded as a more general and flexible alternative to Gaussian processes that allows a more straightforward approach to computation and asymptotics.

3.2 General results

3.2.1 Main theorem

We consider a random variable J taking values in $\mathbb{N}$. For each $J \in \mathbb{N}$, we consider a triangular array of linearly independent real-valued functions $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_J)^T$ defined on Ω_0. Note that the resulting basis functions may change from one stage to the next, although we did not make that explicit in our notation. A joint prior Π is assigned on J and the coefficients of basis functions $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^T$ such that the following conditions hold:

(A1) The prior for J satisfies $\Pi(J > j) \leq A(j)$ and $\Pi(j \leq J < c_1 j) \geq B(j)$ when j is sufficiently large for some constant $c_1 > 1$. The functions $A(j)$ and $B(j)$ are assumed to be nonnegative and strictly decreasing to 0 when $j \to \infty$.

(A2) Given J, we consider a J-dimensional joint distribution as the prior for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^T$ satisfying

$$\Pi(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|_2 \leq \epsilon) \geq \exp\{-c_2 J \log(1/\epsilon)\}$$

for every $\|\boldsymbol{\theta_0}\|_\infty \leq H$, where c_2 is some positive constant, H is chosen sufficiently large and $\epsilon > 0$ is sufficiently small. Also, we assume an exponential decay condition given by $\Pi(\boldsymbol{\theta} \notin [-M, M]^J) \leq J \exp\{-CM^{t_3}\}$ for sufficiently large M and some constant

$t_3 > 0$.

Remark 1. Examples of priors satisfying (A2) include independent gamma, exponential distributions assigned on each element of $\boldsymbol{\theta}$ and multivariate normal and Dirichlet distributions provided the parameters lie in a fixed compact set; see Lemma 6.1 of Ghosal et al. (2000) for the last conclusion.

The priors on J and $\boldsymbol{\theta}$ are allowed to depend on n provided that the constants appearing in (A1) and (A2) are free of n. However, to simplify the notation, we shall drop the subscript n from Π. In the following, with some abuse of notation, we shall use Π as the prior distribution on J and $\boldsymbol{\theta}$ as well as for the induced prior distribution on functions $\boldsymbol{\theta}^T\boldsymbol{\xi}$.

We consider two distance metrics d_1 and d_2 on functions belonging to Ω satisfying the following conditions:

$$d_1(\boldsymbol{\theta_1}^T\boldsymbol{\xi}, \boldsymbol{\theta_2}^T\boldsymbol{\xi}) \leq a(J)\|\boldsymbol{\theta_1} - \boldsymbol{\theta_2}\|_2, \tag{3.1}$$

$$d_2(\boldsymbol{\theta_1}^T\boldsymbol{\xi}, \boldsymbol{\theta_2}^T\boldsymbol{\xi}) \leq b(J)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \tag{3.2}$$

for some positive increasing functions $a(\cdot)$, $b(\cdot)$ and $\boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \mathbb{R}^J$ for every possible value of $J \in \mathbb{N}$.

Now we state the main theorem, which can be regarded as a master theorem where the required conditions for posterior convergence rates for various inference problems are established, in a manner similar to Theorem 2.1 of van der Vaart and van Zanten (2008) and Theorem 3.1 of van der Vaart and van Zanten (2009).

Theorem 10. *Let $\epsilon_n \geq \bar{\epsilon}_n$ be two sequence of positive numbers satisfying $\epsilon_n \to 0$ and $n\bar{\epsilon}_n^2 \to \infty$ 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 $\boldsymbol{\theta}_{0,j} \in \mathbb{R}^j$ for any $j \in \mathbb{N}$, such*

that the following conditions hold for some positive constants a_1, a_1', a_2:

$$e(\bar{J}_n) \le \bar{\epsilon}_n, \tag{3.3}$$

$$d_2(w_0, \boldsymbol{\theta}_{0,j}^T \boldsymbol{\xi}) \le e(j), \tag{3.4}$$

$$A(J_n) \le a_1 \exp\{-(a_2+4)n\bar{\epsilon}_n^2\}, \tag{3.5}$$

$$\log\{1/B(\bar{J}_n)\} + c_1 c_2 \bar{J}_n \log(2b(c_1\bar{J}_n)/\bar{\epsilon}_n) \le a_2 n\bar{\epsilon}_n^2, \tag{3.6}$$

$$J_n \left\{\log J_n + \log a(J_n) + \log M_n + \log(1/\epsilon_n)\right\} \le n\epsilon_n^2, \tag{3.7}$$

$$J_n \exp\{-CM_n^{t_3}\} \le a_1' \exp\{-(a_2+4)n\bar{\epsilon}_n^2\}. \tag{3.8}$$

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

$$\begin{aligned}
\log D(\epsilon_n, \mathcal{W}_{J_n,M_n}, d_1) &\le n\epsilon_n^2, &(3.9)\\
\Pi(W \notin \mathcal{W}_{J_n,M_n}) &\le (a_1 + a_1')\exp\{-(a_2+4)n\bar{\epsilon}_n^2\}, &(3.10)\\
-\log \Pi\{w = \boldsymbol{\theta}^T\boldsymbol{\xi} : d_2(w_0, w) \le \bar{\epsilon}_n\} &\le a_2 n\bar{\epsilon}_n^2. &(3.11)
\end{aligned}$$

Proof. We first verify (3.9), using the definition of packing number, the assumptions on M_n, $J_n \ge 2$, the fact that $a(\cdot)$ is increasing and (3.1), we obtain:

$$\begin{aligned}
&\log D(\epsilon_n, \mathcal{W}_{J_n,M_n}, d_1)\\
&\le \log\Big(\sum_{j=1}^{J_n} D(\epsilon_n/a(j), \{\boldsymbol{\theta} \in \Omega_j, \|\boldsymbol{\theta}\|_\infty \le M_n\}, \|\cdot\|_2)\Big)\\
&\le \log\Big[J_n\Big\{\frac{\sqrt{J_n}M_n a(J_n)}{\epsilon_n}\Big\}^{J_n}\Big]\\
&\le J_n(\log J_n + \log M_n + \log a(J_n) + \log(1/\epsilon_n)) \le n\epsilon_n^2. \qquad (3.12)
\end{aligned}$$

Next, to verify (3.10), observe the following:

$$\begin{aligned}\Pi(w \notin \mathcal{W}_{J_n,M_n}) &\leq \Pi(J > J_n) + \sum_{j=1}^{J_n} \Pi(\boldsymbol{\theta} \notin [-M_n, M_n]^j)\Pi(J = j) \\ &\leq A(J_n) + J_n \exp\{-CM_n^{t_3}\} \\ &\leq (a_1 + a_1') \exp\{-(a_2 + 4)n\bar{\epsilon}_n^2\}. \end{aligned} \tag{3.13}$$

For (3.11), using (3.3) and (3.4), for every $j \geq \bar{J}_n$, there exists a $\boldsymbol{\theta}_{0,j}$ such that $d_2(w_0, \boldsymbol{\theta}_{0,j}^T\boldsymbol{\xi}) \leq e(j) \leq e(\bar{J}_n) \leq \bar{\epsilon}_n$. Hence, for these $\boldsymbol{\theta} \in \mathbb{R}^j$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{0,j}\|_2 \leq \bar{\epsilon}_n/b(j)$,

$$d_2(w_0, \boldsymbol{\theta}^T\boldsymbol{\xi}) \leq d_2(w_0, \boldsymbol{\theta}_{0,j}^T\boldsymbol{\xi}) + d_2(\boldsymbol{\theta}_{0,j}^T\boldsymbol{\xi}, \boldsymbol{\theta}^T\boldsymbol{\xi}) \leq 2\bar{\epsilon}_n.$$

Therefore,

$$\begin{aligned}\Pi\{w = \boldsymbol{\theta}^T\boldsymbol{\xi} : d_2(w_0, w) \leq 2\bar{\epsilon}_n\} &\geq \Pi(\bar{J}_n \leq J \leq c_1\bar{J}_n)\Pi(\boldsymbol{\theta} \in \mathbb{R}^J : \|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|_2 \leq \bar{\epsilon}_n/b(J)) \\ &\geq B(\bar{J}_n)\Pi(\boldsymbol{\theta} \in \mathbb{R}^{c_1\bar{J}_n} : \|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|_2 \leq \bar{\epsilon}_n/b(c_1\bar{J}_n)) \\ &\geq B(\bar{J}_n) \exp\Big\{ - c_1c_2\bar{J}_n \log\Big(\frac{b(c_1\bar{J}_n)}{\bar{\epsilon}_n}\Big)\Big\} \end{aligned} \tag{3.14}$$

for c_1, c_2 defined by (A1) and (A2). By taking the negative of the logarithm on both sides, substituting $\bar{\epsilon}_n$ by $\bar{\epsilon}_n/2$, we get

$$-\log \Pi\{w = \boldsymbol{\theta}^T\boldsymbol{\xi} : d_2(w_0, w) \leq \bar{\epsilon}_n\} \leq \log\left\{\frac{1}{B(\bar{J}_n)}\right\} + c_1c_2\bar{J}_n \log\left(\frac{2b(c_1\bar{J}_n)}{\bar{\epsilon}_n}\right). \tag{3.15}$$

By (3.6), the result follows. □

Remark 2. Conditions (3.3) and (3.5) in Theorem 10 require sufficiently large $J_n, \bar{J}_n$ in order to have sufficiently good approximation to w_0. In other words, J_n controls the bias of the

model. Meanwhile, Conditions (3.6), (3.7) and (3.8) state that $J_n, \bar{J}_n$ should not be too large if the complexity of the model is to be controlled. When studying Bayesian asymptotic properties, a balance between bias and complexity needs to be established to obtain the optimal posterior convergence.

3.2.2 Posterior convergence rates

The general theorem can be further simplified for a particular collection of basis functions. We assume the approximation error is $e(J) \asymp J^{-\alpha}(\log J)^s$ for functions belonging to a Hölder class and $s \geq 0$. Such choice includes B-splines, wavelets, Fourier series and many other commonly used base. Let $d_1 = d_2$ as the L_2-distance or the L_∞-distance. Assume that the prior on J satisfies the following tail bounds:

(A3) The prior for J satisfies $\Pi(J > j) \leq \exp\{-c_3 j \log^{t_1} j\}$ and $\Pi(j \leq J < c_1 j) \geq \exp\{-c_4 j \log^{t_2} j\}$ for some constants $c_3, c_4 > 0$, $0 \leq t_1 \leq t_2 \leq 1$ when j is sufficiently large.

Remark 3. Geometric, Poisson and negative binomial distributions on J satisfy Condition (A3) respectively with $t_1 = t_2 = 0$, $t_1 = t_2 = 1$ and $t_1 = t_2 = 0$.

Now we can obtain the posterior convergence rate for various inference problems.

Theorem 11. *Suppose that we have independent observations X_i following some distributions with density $p_{i,w,\boldsymbol{\eta}}$, $i = 1, \ldots, n$, where $\boldsymbol{\eta}$ is a nuisance parameter taking values in some Euclidean space Ω_1. Assume that $w_0 \in \mathcal{C}^\alpha(\Omega_0)$ and $\boldsymbol{\eta_0}$ are true values of w and $\boldsymbol{\eta}$. Let $\epsilon_n \geq \bar{\epsilon}_n$ be two sequence of positive numbers satisfying $\epsilon_n \to 0$ and $n\bar{\epsilon}_n^2 \to \infty$ as $n \to \infty$. Assume that we are using basis functions $\boldsymbol{\xi}$ such that for all sufficiently large J, there exists a $\boldsymbol{\theta_0} \in \mathbb{R}^J$ and*

$\|\boldsymbol{\theta_0}\|_\infty \leq H$ *satisfying*

$$d(w_0, \boldsymbol{\theta_0}^T \boldsymbol{\xi}) \leq C_0 J^{-\alpha} (\log J)^s, \ s \geq 0 \tag{3.16}$$

$$d(\boldsymbol{\theta_1}^T \boldsymbol{\xi}, \boldsymbol{\theta_2}^T \boldsymbol{\xi}) \leq J^{K_0} \|\boldsymbol{\theta_1} - \boldsymbol{\theta_2}\|_2, \ K_0 \geq 0 \tag{3.17}$$

where $\boldsymbol{\theta_1}$ *and* $\boldsymbol{\theta_2}$ *are arbitrary vectors in* $\mathbb{R}^J$*. Let* J_n*,* $\bar{J}_n \geq 2$ *and* M_n *be sequences of positive numbers such that the following conditions hold:*

$$\bar{J}_n^{-\alpha} (\log \bar{J}_n)^s \leq \bar{\epsilon}_n, \tag{3.18}$$

$$c_3 J_n \log^{t_1} J_n \geq 6n\bar{\epsilon}_n^2, \tag{3.19}$$

$$\log J_n + 6n\epsilon_n^2 \leq c M_n^{t_3}, \tag{3.20}$$

$$\Pi(\boldsymbol{\eta} : \|\boldsymbol{\eta} - \boldsymbol{\eta_0}\| < \bar{\epsilon}_n) \geq \exp\{-n\bar{\epsilon}_n^2\}, \tag{3.21}$$

$$J_n\{(K_0 + 1) \log J_n + \log M_n + \log(1/\epsilon_n)\} \leq n\epsilon_n^2, \tag{3.22}$$

$$\bar{J}_n\{c_4 \log^{t_2} \bar{J}_n + c_1 c_2 K_0 \log(\bar{J}_n) + c_1 c_2 \log(1/\bar{\epsilon}_n)\} \leq 2n\bar{\epsilon}_n^2, \tag{3.23}$$

$$\max_{1 \leq i \leq n} \{K(p_{i,w_0,\boldsymbol{\eta_0}}, p_{i,w,\boldsymbol{\eta}}), V(p_{i,w_0,\boldsymbol{\eta_0}}, p_{i,w,\boldsymbol{\eta}})\} \lesssim d^2(w_0, w) + \|\boldsymbol{\eta} - \boldsymbol{\eta_0}\|_2^2, \tag{3.24}$$

provided the right hand side is sufficiently small. Assume that the prior on J *and* $\boldsymbol{\theta}$ *satisfy conditions (A2) and (A3). Further, assume that there exists* $\mathcal{F}_n \subseteq \Omega_1$ *such that*

$$\Pi(\boldsymbol{\eta} \notin \mathcal{F}_n) \leq \exp\{-6n\bar{\epsilon}_n^2\}, \tag{3.25}$$

$$\log D(\epsilon_n, \mathcal{F}_n, \|\cdot\|_2) \leq n\epsilon_n^2. \tag{3.26}$$

Then for every $L_n \to \infty$*, we have that*

$$\Pi(h_n(w, w_0) > L_n \epsilon_n | X_1, \ldots, X_n) \to 0, \textit{ in } P_{w_0,\boldsymbol{\eta_0}}^{(n)}\textit{-probability}, \tag{3.27}$$

where h_n is defined by $h_n^2(w_1, w_2) = \frac{1}{n}\sum_{i=1}^{n} h^2(p_{i,w_1}, p_{i,w_2})$, the average squared Hellinger distance, and $P_{w_0,\boldsymbol{\eta_0}}^{(n)}$ is the true distribution of $(X_1, \ldots, X_n)$.

Proof. In order to obtain the posterior convergence rate, we verify the following conditions according to Theorem 2:

$$\log D(\epsilon_n, \mathcal{W}_{J_n,M_n} \times \mathcal{F}_n, h_n) \leq b_1 n\epsilon_n^2, \tag{3.28}$$

$$\Pi(w \notin \mathcal{W}_{J_n,M_n}) + \Pi(\boldsymbol{\eta} \notin \mathcal{F}_n) \leq b_3 \exp\{-(b_2+4)n\bar{\epsilon}_n^2\}, \tag{3.29}$$

$$\Pi(\mathcal{K}(w_0, \bar{\epsilon}_n)) \geq b_4 \exp\{-b_2 n\bar{\epsilon}_n^2\}, \tag{3.30}$$

where $\mathcal{W}_{J_n,M_n}$ is defined in Theorem 10 and b_1, b_2, b_3, b_4 are some positive constants. Note that the conditions in Theorem 10 are satisfied for $a_1 = a_1' = 1$ and $a_2 = 2$ in the following way:

- Relation (3.3) holds because of the definition of $e(\cdot)$ and (3.18);
- Relation (3.4) holds because of the approximation ability of the basis $\boldsymbol{\xi}$ in (3.16);
- Note that $A(x) = \exp\{-c_3 x \log^{t_1} x\}$, hence by (3.19), (3.5) holds;
- By Condition (A3) and (3.17), we have $B(x) = \exp\{-c_4 x \log^{t_2} x\}$ and $b(x) = x^{K_0}$. Thus (3.6) holds for $a_2 = 2$ because of (3.23).
- Relation (3.7) holds because $a(x) = x^{K_0}$ and (3.22);
- Relation (3.8) holds for $a_1' = 1$ and $a_2 = 2$ by (3.20).

Therefore, we have

$$\log D(\epsilon_n, \mathcal{W}_{J_n,M_n}, d) \leq n\epsilon_n^2,$$

$$\Pi(W \notin \mathcal{W}_{J_n,M_n}) \leq 2\exp\{-6n\bar{\epsilon}_n^2\},$$

$$\Pi\{w = \boldsymbol{\theta}^T\boldsymbol{\xi} : d(w_0, w) \leq \bar{\epsilon}_n\} \geq \exp\{-n\bar{\epsilon}_n^2\}.$$

Because $h^2(p_{i,w_0,\boldsymbol{\eta_0}}, p_{i,w,\boldsymbol{\eta}}) \leq d^2(w_0, w) + \|\boldsymbol{\eta_0} - \boldsymbol{\eta}\|_2^2$, we have

$$\log D(\epsilon_n, \mathcal{W}_{J_n,M_n} \times \mathcal{F}_n, h) \leq \log D(\epsilon_n, \mathcal{W}_{J_n,M_n}, d) + \log D(\epsilon_n, \mathcal{F}_n, \|\cdot\|_2) \leq 2n\epsilon_n^2. \quad (3.31)$$

Also, relation (3.29) holds for $b_2 = 2$ and $b_3 = 3$. For (3.30), observe

$$\Pi(\mathcal{K}(w_0, \bar{\epsilon}_n)) \geq \Pi(d(w_0, w) \leq \bar{\epsilon}_n)\Pi(\|\boldsymbol{\eta} - \boldsymbol{\eta_0}\| < \bar{\epsilon}_n) \geq \exp\{-2n\bar{\epsilon}_n^2\}. \quad (3.32)$$

Hence (3.30) holds for $b_4 = 1$. □

Remark 4. Relation (3.17) holds for B-splines, polynomials and Fourier series basis with $K_0 = 1/2$ when d is the L_2-distance and $K_0 = 1$ when d is the supremum distance. For wavelets, (3.17) holds with $K_0 = 1$ for both the L_2-distance and the supremum distance. This is because

$$\|(\boldsymbol{\theta_1} - \boldsymbol{\theta_2})^T\boldsymbol{\xi}\|_p \leq \sum_{j=1}^{J} |\theta_{1j} - \theta_{2j}| \max_{1\leq j\leq J} \|\boldsymbol{\xi}_j\|_p \leq \sqrt{J}\|\boldsymbol{\theta_1} - \boldsymbol{\theta_2}\|_2 C_{p,J}$$

for $C_{p,J} = \max_{1\leq j\leq J} \|\boldsymbol{\xi}_j\|_p$ and $1 \leq p \leq \infty$. For B-splines, polynomials and Fourier series basis, $C_{p,J} \asymp 1$ when $p = 2$ and $C_{p,J} \asymp \sqrt{J}$ when $p = \infty$. For wavelets, $C_{p,J} \asymp \sqrt{J}$ for $p = 2$ or ∞.

In some applications, $\boldsymbol{\eta}$ may be vacuous, i.e., not present at all.

Theorem 11 suggests that in order to obtain adaptive rates, it is crucial to

(i) choose sequences J_n, $\bar{J}_n$, ϵ_n, M_n in the rate equations (3.18)–(3.23);

(ii) bound the KL-divergence by d^2, where d is either the L_2 or the L_∞-distance in applica-

tions.

Bounding the KL-divergence can be very different for various statistical problems, while the choices of J_n and $\bar{J}_n$ are common for a set of basis functions.

Next, we give some examples to illustrate the choice of the sequences.

Example 6. *Fourier trigonometric series*

Choose the basis $\{\cos 2\pi jx, \sin 2\pi jx,\ j \in \mathbb{N}\}$ in $L_2(0,1)$. Then for a function $w_0 \in \mathcal{C}^\alpha(0,1)$, we have $e(J) \asymp (\log J)^s/J^\alpha$, where $s = 0$ if d is L_2, and $s = 1$ if d is the supremum distance (Jackson, 1930). The rate calculation proceeds in the following way: Relations (3.18) and (3.23) imply $\bar{J}_n^{-\alpha}(\log n)^s \lesssim \bar{\epsilon}_n$ and $\bar{J}_n \log n \lesssim n\bar{\epsilon}_n^2$. Hence $(\log n)^{s+\alpha} \asymp \bar{\epsilon}_n(n\bar{\epsilon}_n^2)^\alpha$, $\bar{\epsilon}_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{(s+\alpha)/(2\alpha+1)}$ and $\bar{J}_n \asymp n^{1/(2\alpha+1)}(\log n)^{(2s-1)/(2\alpha+1)}$. Now using (3.19), we have $J_n \log^{t_1} n \gtrsim n\bar{\epsilon}_n^2$. Hence we choose $J_n \asymp n^{1/(2\alpha+1)}(\log n)^{(2s+2\alpha)/(2\alpha+1)-t_1}$. Note that (3.22) implies $J_n \log n \lesssim n\epsilon_n^2$. As a result, we choose $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{(s+\alpha)/(2\alpha+1)+(1-t_1)/2}$. For (3.20), we may choose $M_n^{t_3} \asymp n$, then the relationship holds for all $\alpha > 0$.

Example 7. *Bernstein polynomials*

We consider the Bernstein polynomial prior proposed by Petrone (1999a,b). Consider a continuously differentiable density function w_0 with bounded second derivative, the approximation property of Bernstein polynomials to w_0 is $e(J) = C/J$ for some universal constant C and d as the supremum distance (Lorenz, 1953). We can choose $\bar{J}_n = (n/\log n)^{1/3}$, $J_n = n^{1/3}(\log n)^{2/3-t_1}$, $\bar{\epsilon}_n = (n/\log n)^{-1/3}$ and $M_n = n^{1/t_3}$. The rate ϵ_n is $n^{-1/3}(\log n)^{1/3+(1-t_1)/2}$, which has the same polynomial power as given in Ghosal (2001). In fact, for any $0 \le \alpha \le 2$, the approximation rate of Bernsten polynomials is $J^{-\alpha/2}$ and the resulting posterior convergence rate is $n^{-\alpha/2(\alpha+1)}(\log n)^{\alpha/2(\alpha+1)}$; see Kruijer and van der Vaart (2008). The poor convergence rate stems from the poor approximation rate of Bernsten polynomials. Kruijer and van der Vaart (2008) used coarsened Bernstein polynomials and showed that for any $f \in \mathcal{C}^\alpha[0,1]$

with $0 \leq \alpha \leq 1$, the approximation rate with J undetermined parameters is $J^{-\alpha}$. If we choose $\bar{J}_n \asymp (n/\log n)^{-\alpha/(2\alpha+1)}$, then the rate is $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$, which adapts to the optimal rate in the range $0 \leq \alpha \leq 1$.

Example 8. *Polynomial basis*

Consider the orthogonal polynomials as the approximation tool for $w_0 \in \mathcal{C}^\alpha(0,1)$, we have the same approximation result with Fourier series under the L_2- or the L_∞-metrics (e.g., Theorem 6.1 of Hesthaven et al. (2007)). Hence the choice of J_n, M_n and rates are exactly the same with Example 6. The rate is $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)}(\log n)^{(s+\alpha)/(2\alpha+1)+(1-t_1)/2}$, where $s = 0$ if d is L_2 and $s = 1$ when d is the supremum distance.

Example 9. *B-splines*

If we choose the B-spline functions as the basis, then for $w_0 \in \mathcal{C}^\alpha(0,1)$, we have $e(J) \asymp J^{-\alpha}$ for d the supremum distance or the L_2-distance. Then (3.18) and (3.23) imply $\bar{J}_n^{-\alpha} \lesssim \bar{\epsilon}_n$ and $\bar{J}_n \log n \lesssim n\bar{\epsilon}_n^2$, hence $\bar{J}_n = (n/\log n)^{1/(2\alpha+1)}$ and $\bar{\epsilon}_n = (n/\log n)^{-\alpha/(2\alpha+1)}$. Now using (3.19), we have $J_n \log^{t_1} n \gtrsim n\bar{\epsilon}_n^2$, hence we choose $J_n = n^{1/(2\alpha+1)}(\log n)^{2\alpha/(2\alpha+1)-t_1}$. So the rate ϵ_n can be chosen as $n^{-\alpha/(2\alpha+1)}$ multiplied by some power of $\log n$, where the power depends on the statistical problem. A brief introduction to splines is given in Appendix A.1.

Example 10. *Wavelets*

We consider a multiresolution truncated wavelet series on $(0,1)$:

$$\sum_{k\in\mathbb{Z}} \alpha_k \phi_k(x) + \sum_{j=0}^{j_1} \sum_{k\in\mathbb{Z}} \beta_{jk}\psi_{jk}(x), \tag{3.33}$$

where ϕ is the father wavelet, ψ is the mother wavelet, $\phi_k(x) = \phi(x-k)$, $\psi_{jk}(x) = 2^{j/2}\psi(2^j(x-k))$, $k \in \mathbb{N}$, $j \in \mathbb{N}_0$ and $j_1 \in \mathbb{N}$. We put priors on j_1 and wavelet coefficients α and β. It has been shown that, for $w_0 \in \mathcal{C}^\alpha$, the approximation error is $e(j_1) = 2^{-j_1\alpha}$ for d chosen as the L_2-

distance (Mallat, 1989). Meanwhile, if w_0 has compact support, the number of nonzero coefficients is $O(2^{j_1})$. Hence we apply Theorem 11 for $J = 2^{j_1}$ and choose $\bar{J}_n = (n/\log n)^{1/(2\alpha+1)}$, $J_n = n^{1/(2\alpha+1)}(\log n)^{2\alpha/(2\alpha+1)-t_1}$, $M_n = n^{1/t_3}$ and $\bar{\epsilon}_n = (n/\log n)^{-\alpha/(2\alpha+1)}$. Doing the same calculation as in Example 6, the resulting rate ϵ_n is $n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$. This rate coincides with the optimal rate for white noise models obtained in Lian (2011) and for density estimation and regression models obtained in Rivoirard and Rousseau (2012).

Example 11. *Multivariate B-splines*

Theorem 10 can be used in multi-dimensional situation as well. Consider the tensor-product B-splines (Schumaker, 2007) as a basis in $\mathcal{C}^\alpha[0,1]^d$. Then we have $e(J) \asymp J^{-\alpha/d}$ for d as the supremum distance or the L_2-distance, where $J = K^d$, and K is the number of univariate B-spline functions used in making the tensor products. Apply Theorem 11 with $\bar{J}_n = (n/\log n)^{1/(2\alpha+d)}$, $\bar{\epsilon}_n = (n/\log n)^{-\alpha/(2\alpha+d)}$, $M_n = n^{1/t_3}$ to obtain the rate ϵ_n as $n^{-\alpha/(2\alpha+d)}$ multiplied by some power of $\log n$, where the power depends on the statistical problem. A brief introduction to tensor product B-splines is given in Appendix A.2.

In the posterior convergence rate theorem, the basis $\boldsymbol{\xi}$ will be assumed to satisfy one of the following three conditions: For a function $f \in \mathcal{C}^\alpha(0,1)$, there exists $\boldsymbol{\theta} \in \mathbb{R}^J$ such that

(B1) $\|f - \boldsymbol{\theta}^T\boldsymbol{\xi}\|_\infty \le CJ^{-\alpha}$;

(B2) $\|f - \boldsymbol{\theta}^T\boldsymbol{\xi}\|_\infty \le CJ^{-\alpha}\log J$;

(B3) $\|f - \boldsymbol{\theta}^T\boldsymbol{\xi}\|_2 \le CJ^{-\alpha}$.

By the above discussions, B-splines and wavelets satisfy (B1) (and hence (B2) also), polynomials and Fourier series satisfy (B2) and all of them satisfy (B3).

3.2.3 Approximation ability of B-splines

B-spline functions can be used to approximate a function appearing in the underlying data generating scheme (e.g., the true density). The approximation ability of linear combinations of B-spline functions is determined by the smoothness level α of the target function and the number of spline basis functions J, provided that $q \geq \alpha$. Denote $\boldsymbol{B}$ as the column vector of B-spline basis functions. In the following discussion, we assume that q is chosen large enough such that $q \geq \alpha$ holds for every statistical problem we are interested in.

Lemma 7. *(a) For any function $f \in \mathcal{C}^{\alpha}(0,1)$, $0 < \alpha \leq q$, there exists $\boldsymbol{\theta} \in \mathbb{R}^J$ and a constant $C > 0$ such that*

$$\|f - \boldsymbol{\theta}^T \boldsymbol{B}\|_{\infty} \leq C J^{-\alpha} \|f^{(\alpha)}\|_{\infty}. \tag{3.34}$$

(b) Further, if $f > 0$ and $J > J_0$, where J_0 is a sufficiently large constant that only depends on q, we can choose every element of $\boldsymbol{\theta}$ to be positive.
(c) If $0 < f < 1$ and $J > J_0$, where J_0 is a sufficiently large constant that only depends on q, we can choose every element of $\boldsymbol{\theta}$ to be between 0 and 1.
(d) Define $B_i^* = B_i / \int_0^1 B_i(x)dx$ for $i = 1, \ldots, J$ and $\boldsymbol{B}^*$ as the column vector of B_i^*s. If f is a positive density function, then for sufficiently large J, there exists $\boldsymbol{\theta} \in \Delta_J$ and a constant $C > 0$ such that

$$\|f - \boldsymbol{\theta}^T \boldsymbol{B}^*\|_{\infty} \leq C J^{-\alpha} \|f^{(\alpha)}\|_{\infty}. \tag{3.35}$$

Proof. The first part is a well-known spline approximation result (de Boor, 2001).

For the second assertion, find $\epsilon > 0$ such that $f \geq \epsilon$ pointwise. Using Corollaries 4 and 6 in Chapter 11 of de Boor (2001), for each θ_i, there exists a universal constant C_1 that depends

only on q, such that

$$|\theta_i - c| \leq C_1 \sup\{|f(x) - c| : x \in [t_{i+1}, t_{i+q-1}]\}$$

for and any choice of the constant c; here t_{i+1} and t_{i+q-1} are $(i+1)$-th and $(i+q-1)$-th knots. Choose $c = \inf\{f(x) : x \in [t_{i+1}, t_{i+q-1}]\} \geq \epsilon$, and note that the infimum is attained somewhere in $[t_{i+1}, t_{i+q-1}]$, say at t^*. By the smoothness condition on f, we have

$$\sup\{|f(x) - c| : x \in [t_{i+1}, t_{i+q-1}]\} \leq C_2 |x - t^*|^{\min(\alpha,1)} \leq C_2 (q/J)^{\min(\alpha,1)}$$

for some constant $C_2 > 0$. Choose $J > q(C_1 C_2/\epsilon)^{\max(1/\alpha,1)}$, then $\theta_i > c - C_1(q/J)^{\min(\alpha,1)} \geq 0$.

Part (c) is a consequence of (b) by considering $1 - f > 0$.

For Part (d), by (b), we know there exists a $\boldsymbol{\eta_1} \in (0,\infty)^J$ such that $\|f - \boldsymbol{\eta_1}^T \boldsymbol{B}\| \lesssim J^{-\alpha}$. Define $\eta_{2,i} = \eta_{1,i} \int_0^1 B_i(x)dx$ for $i = 1, \ldots, J$. Then $\|f - \boldsymbol{\eta_2}^T \boldsymbol{B^*}\|_\infty \lesssim J^{-\alpha}$. By integration, we have $|1 - \sum_{i=1}^J \eta_{2,i}| \lesssim J^{-\alpha}$. Choose $\boldsymbol{\theta} = \boldsymbol{\eta_2}/\|\boldsymbol{\eta_2}\|_1 \in \Delta_J$. Note that $\|\boldsymbol{\theta} - \boldsymbol{\eta_2}\|_1 \lesssim J^{-\alpha}$. Since $\|\boldsymbol{\theta}^T \boldsymbol{B} - \boldsymbol{\eta_2}^T \boldsymbol{B}\|_\infty \leq \|\boldsymbol{\theta} - \boldsymbol{\eta_2}\|_1 \|\boldsymbol{B}\|_\infty$ and every B_i^* is uniformly bounded above, we have

$$\|f - \boldsymbol{\theta}^T \boldsymbol{B^*}\|_\infty \leq \|f - \boldsymbol{\eta_2}^T \boldsymbol{B^*}\|_\infty + \|\boldsymbol{\theta} - \boldsymbol{\eta_2}\|_\infty \max_{1 \leq i \leq J} B_i^*(x) \lesssim J^{-\alpha}.$$

This completes the proof. □

Remark 5. In Part (b), the strictness of the condition $f > 0$ is crucial. If we approximate only a nonnegative function f using a nonnegative coefficient vector $\boldsymbol{\theta}$, then the approximation error is only $O(J^{-1})$ (de Boor and Daniel, 1974), which does not adapt to the smoothness level beyond 1.

3.3 Gaussian white noise model

We consider a Gaussian white noise model

$$dX(t) = f(t)dt + \frac{\sigma}{\sqrt{n}}dW(t),\ 0 \le t \le 1,\ \sigma > 0, \tag{3.36}$$

where $X(t)$ is the observed signal, $f(t)$ is the unknown signal and $W(t)$ is a standard Wiener process. Let ϕ_i, $i = 1, 2, \ldots$ be any orthonormal basis in $L_2[0, 1]$. Assume that $f \in L_2[0, 1]$. Then this problem can be transformed into the estimation of the mean $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ for an infinite-dimensional normal distribution as follows:

$$X_i = \theta_i + \frac{\epsilon_i}{\sqrt{n}},\ i = 1, 2, \ldots; \tag{3.37}$$

here X_i's are independent observations, ϵ_i's are i.i.d white noise variables that follows the normal distribution $\mathrm{N}(0, \sigma^2)$. Pinsker (1980) showed that the minimax convergence rate is $n^{-\alpha/(2\alpha+1)}$ for α-smoothness functions with respect to ℓ_2-metric. In a Bayesian study, minimax rate was obtained by Freedman (1999) and Zhao (2000). Adaptive rate was obtained by Belitser and Ghosal (2003) for a discrete collection of smoothness parameters α. Babenko and Belitser (2010) considered putting a prior on the oracle J, which is defined as the best cut-off $\theta_i = 0$ for all $i > J$ such that the risk of $\{X_i \mathbb{1}(i \le J) : i \in \mathbb{N}\}$ is minimized. They showed that such an oracle estimator is minimax for a general smoothness class and hence obtained adaptation results as well.

Assume that the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ belongs to a Sobolev type class $\mathcal{S}_\alpha(A) = \{\boldsymbol{\theta} : \sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2 \le A\}$ for some $\alpha, A > 0$. Here α can be viewed as a smoothness parameter. We consider a cut-off $J > 0$ such that for all $i > J$, θ_i's are set to 0 in the prior. The approximation

error of using the truncation any $\boldsymbol{\theta} \in \mathcal{S}_\alpha(A)$ is $e(J)$, given by

$$e(J)^2 \asymp \sum_{i=J+1}^{\infty} \theta_i^2 \leq \sum_{i=J+1}^{\infty} \frac{\theta_i^2 i^{2\alpha}}{J^{2\alpha}} \leq \frac{A}{J^{2\alpha}}. \tag{3.38}$$

This is equivalent to using the finite random series prior $\sum_{j=1}^{J} \theta_j \phi_j(t)$ for $f(t)$. The posterior convergence rate is obtained as follows:

Theorem 12. *Suppose that the true mean $\boldsymbol{\theta_0} \in \mathcal{S}_\alpha(A)$ for some $\alpha, A > 0$. Assume that the prior satisfies Conditions (A2) and (A3). Then the posterior distribution of $\boldsymbol{\theta}$ converges at the rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$ with respect to the ℓ_2-distance.*

Proof. For two normal likelihood of X's given true value of $(\theta_1, \theta_2, \ldots)$ as $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, a simple calculation (see Belitser and Ghosal, 2003) bounds their KL-divergences as following

$$\begin{aligned} K(p_{X,\boldsymbol{\theta}}, p_{X,\boldsymbol{\beta}}) &\lesssim \|\boldsymbol{\theta} - \boldsymbol{\beta}\|_2^2 \\ V(p_{X,\boldsymbol{\theta}}, p_{X,\boldsymbol{\beta}}) &\lesssim \|\boldsymbol{\theta} - \boldsymbol{\beta}\|_2^2 + \|\boldsymbol{\theta} - \boldsymbol{\beta}\|_2^4. \end{aligned} \tag{3.39}$$

Further, $h^2(p_{X,\boldsymbol{\theta}}, p_{X,\boldsymbol{\beta}}) \gtrsim \|\boldsymbol{\theta} - \boldsymbol{\beta}\|_2^2$ if the left hand side is sufficiently small. Hence by using Theorem 11 for

$$\bar{J}_n = (n/\log n)^{1/(2\alpha+1)}, \;\; J_n = n^{1/(2\alpha+1)}(\log n)^{2\alpha/(2\alpha+1)-t_1},$$

$$\bar{\epsilon}_n = (n/\log n)^{-\alpha/(2\alpha+1)}, \;\; \epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2},$$

$\boldsymbol{\eta}$ as a vacuous parameter and d as the ℓ_2-distance, we obtain the result. □

Remark 6. It is possible to treat unknown σ by assigning a prior on σ as in (3.52)–(3.54). The posterior convergence rate will still be $n^{-\alpha/(2\alpha+1)}$ multiplied by a logarithmic factor provided

the prior density on σ is positive in a neighborhood of σ_0 and has sharp tails on both sides; see Section 3.6.

3.4 Density estimation

3.4.1 Density on the unit interval

In this section, we apply Theorem 11 to density estimation problems. Frequentist optimal rate of convergence $n^{-\alpha/(2\alpha+1)}$ has been obtained for the maximum likelihood estimators in Hasminskii (1978). Stone (1990) constructed sieve MLE based on the log-spline models. A Bayesian log-spline model has been studied in Ghosal et al. (2000), where the optimal posterior convergence rate $n^{-\alpha/(2\alpha+1)}$ was obtained. When α is unknown, the posterior convergence rate $n^{-\alpha/(2\alpha+1)}$, possibly up to an additional logarithmic factor, is established in Ghosal et al. (2003, 2008).

We first consider estimating a density function f_0 that is defined on the unit interval $[0, 1]$. A Bayesian estimator of f can be constructed by using basis functions through a nonnegative, monotonic link function Ψ, i.e., $p_{\boldsymbol{\theta}} = \Psi(\boldsymbol{\theta}^T\boldsymbol{\xi})/\int_0^1 \Psi(\boldsymbol{\theta}^T\boldsymbol{\xi})(x)dx$ for $\boldsymbol{\theta} \in \mathbb{R}^J$ and J is given a prior on $\mathbb{N}$. If we choose Ψ as the exponential function and $\boldsymbol{\xi}$ as the B-spline, then it gives the log-spline model. If Ψ is a polynomial, then $p_{\boldsymbol{\theta}}$ is a rational function of $\boldsymbol{\theta}$. We can also choose Ψ as the identity function, and restrict the prior for $\boldsymbol{\theta}$ on Δ_J when using the normalized B-spline basis, by Lemma 10, part (d). By using the identity link function and B-splines, it is possible to avoid the normalization altogether; see Section 3.4.3.

Theorem 13. *Consider n independent, identically distributed observations $X_1, \ldots, X_n$ from a true density f_0. Assume that there exists a known nonnegative, monotonic, Lipschitz continuous link function Ψ such that $\Psi^{-1}(f_0) \in \mathcal{C}^\alpha(0, 1)$ and f_0 takes values in $[m, M]$ for some positive*

constants m and M. Also let the prior satisfy Conditions (A2) and (A3).

(a) If the basis $\boldsymbol{\xi}$ satisfies Condition (B1), then the posterior distribution converges at rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$ with respect to the Hellinger distance.

(b) If the basis $\boldsymbol{\xi}$ satisfies Condition (B2), then the posterior distribution converges at rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{(\alpha+1)/(2\alpha+1)+(1-t_1)/2}$ with respect to the Hellinger distance.

(c) If Ψ is the identity function and the normalized B-spline basis $\{B_1^*, \ldots, B_J^*\}$ is used, then the prior on $\boldsymbol{\theta}$ can be restricted to Δ_J to obtain the convergence rate in (a).

Proof. We first consider Part (a) when d is the supremum distance. Given J, there exists $\boldsymbol{\theta_0} \in \mathbb{R}^J$ such that $\|\boldsymbol{\theta_0}^T\boldsymbol{\xi} - \Psi^{-1}(f_0)\|_\infty \lesssim J^{-\alpha}$. By the Lipschitz continuity of Ψ, it follows that for every x, $f_0(x) - \Psi(\boldsymbol{\theta_0}^T\boldsymbol{\xi})(x)$ is uniformly bounded by a multiple of $J^{-\alpha}$. Integrating with respect to x, we obtain

$$\left|1 - \int_0^1 \Psi(\boldsymbol{\theta_0}^T\boldsymbol{\xi})(x)dx\right| \leq \int_0^1 |f_0(x) - \Psi(\boldsymbol{\theta_0}^T\boldsymbol{\xi})(x)|dx \lesssim J^{-\alpha}.$$

Hence by the triangle inequality, $\|f_0 - p_{\boldsymbol{\theta_0}}\|_\infty \lesssim J^{-\alpha}$. Therefore, $\|f_0 - p_{\boldsymbol{\theta}}\|_\infty \lesssim J^{-\alpha} + J^{K_0}\|\boldsymbol{\theta_0} - \boldsymbol{\theta}\|_2$. Using the fact that f_0 is positive, it follows that

$$h^2(f_0, p_{\boldsymbol{\theta}}) = \int \frac{|f_0 - p_{\boldsymbol{\theta}}|^2}{(\sqrt{f_0} + \sqrt{p_{\boldsymbol{\theta}}})^2} \leq \frac{1}{m}\|f_0 - p_{\boldsymbol{\theta}}\|_\infty^2 \lesssim J^{-2\alpha} + J^{2K_0}\|\boldsymbol{\theta_0} - \boldsymbol{\theta}\|_2^2. \tag{3.40}$$

Note that

$$\inf p_{\boldsymbol{\theta}} \geq \inf f_0 - \|f_0 - p_{\boldsymbol{\theta}}\|_\infty \geq \inf f_0 - J^{-\alpha} - J^{K_0}\|\boldsymbol{\theta_0} - \boldsymbol{\theta}\|_2 > 0$$

given that J is sufficiently large and $J^{K_0}\|\boldsymbol{\theta_0} - \boldsymbol{\theta}\|_2$ is sufficiently small. Using Lemma 8 of

Ghosal and van der Vaart (2007b), we have

$$\begin{aligned} K(f_0, p_{\boldsymbol{\theta}}) &\leq 2h^2(f_0, p_{\boldsymbol{\theta}}) \Big\| \frac{f_0}{p_{\boldsymbol{\theta}}} \Big\|_\infty \lesssim J^{-2\alpha} + J^{2K_0} \|\boldsymbol{\theta_0} - \boldsymbol{\theta}\|_2^2, \\ V(f_0, p_{\boldsymbol{\theta}}) &\lesssim h^2(f_0, p_{\boldsymbol{\theta}}) \Big(1 + \Big\| \frac{f_0}{p_{\boldsymbol{\theta}}} \Big\|_\infty \Big)^2 \lesssim J^{-2\alpha} + J^{2K_0} \|\boldsymbol{\theta_0} - \boldsymbol{\theta}\|_2^2. \end{aligned} \tag{3.41}$$

If $J^{-\alpha} \leq \epsilon$ and $\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|_2 \leq J^{-K_0}\epsilon$, then $K(f_0, p_{\boldsymbol{\theta_0}}) \leq \epsilon^2$ and $V(f_0, p_{\boldsymbol{\theta_0}}) \leq \epsilon^2$. Hence it suffices to control $\Pi(\|\boldsymbol{\theta_0} - \boldsymbol{\theta}\|_2 \leq J^{-K_0}\epsilon)$, which is bounded below by $\exp\{-c_2 J(\log^{t_1 \vee 1} J + \log(1/\epsilon))\}$ in view of Conditions (A2) and (A3). We apply Theorem 11 for d as the supremum distance, $\bar{J}_n = (n/\log n)^{1/(2\alpha+1)}$, $\bar{\epsilon}_n = (n/\log n)^{-\alpha/(2\alpha+1)}$, $J_n = n^{1/(2\alpha+1)}(\log n)^{2\alpha/(2\alpha+1)-t_1}$, $\boldsymbol{\eta}$ as vacuous and $M_n = n^{1/t_3}$. The result now follows.

Similarly, if we choose

$$\bar{J}_n = (n/\log n)^{1/(2\alpha+1)},\ M_n = n^{1/t_3},\ J_n = n^{1/(2\alpha+1)}(\log n)^{(2\alpha+2)/(2\alpha+1)-t_1},$$

$$\text{and } \bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}(\log n)^{(\alpha+1)/(2\alpha+1)}$$

in Theorem 11, we obtain Part (b).

Part (c) is a consequence of Part (d) of Lemma 7 and the calculations in Part (a). □

3.4.2 Density on the real line

Let $f_0 \in \mathcal{C}^\alpha(\mathbb{R})$. Consider a fixed monotonic link function $\Phi : \mathbb{R} \to (0, 1)$ to change the domain of the function to $(0, 1)$. Suppose that there is an interval $[a, b]$ such that f_0 is bounded away from 0 on it. Consider a pseudo-metric $d_{a,b}(f_1, f_2) = \int_a^b |f_1(y) - f_2(y)|dy$. By constructing a prior on f through the representation $f(y) = \Psi\{\boldsymbol{\theta}^T \boldsymbol{\xi}(\Phi(y))\}$ and arguing as in Theorem 13, we obtain the same posterior convergence rates with respect to $d_{a,b}$. The same method also

applies for half open intervals.

3.4.3 Density estimation using B-splines

Consider using B-splines as the basis. We follow the setting in Subsection 3.4.1. It is well known that (see Schumaker, 2007, Chap. 4)

$$\int_0^1 B_i(x)dx = \begin{cases} i/(q(J-q+1)), & i = 1, \ldots, q-1, \\ 1/(J-q+1), & i = q, \ldots, J-q+1, \cdot \\ (J-i+1)/(q(J-q+1)), & i = J-q+2, \ldots, J \end{cases} \tag{3.42}$$

By Part (c) of Theorem 13, we can restrict the coefficients $\boldsymbol{\theta}$ to the J-dimensional simplex Δ_J. Hence a density $f = \sum_{k=1}^{J} \theta_k B_k^*$ is formed. We put a Dirichlet prior on $\boldsymbol{\theta} \sim \mathrm{Dir}(J; a_1, a_2, \ldots, a_J)$ for any $J \in \mathbb{N}$. Finally, we assign a prior Π on J. Thus a prior on the density f is induced. Given the observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ and a fixed dimension J, the posterior density of $\boldsymbol{\theta}$ is a mixture of Dirichlet distribution:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{X}, J) &\propto \prod_{k=1}^{J} \theta_k^{a_k - 1} \prod_{i=1}^{n} \{ \sum_{k=1}^{J} \theta_k B_k^*(X_i) \} \\ &= \sum_{i_1=1}^{J} \cdots \sum_{i_n=1}^{J} \prod_{k=1}^{J} \theta_k^{a_k - 1} \prod_{s=1}^{n} \theta_{i_s} B_{i_s}^*(X_s). \end{aligned}$$

Using the fact that

$$p(J, \boldsymbol{\theta}|\boldsymbol{X}) \propto p(\boldsymbol{X}|J, \boldsymbol{\theta}) \Pi(\boldsymbol{\theta}|J) \Pi(J), \tag{3.43}$$

the posterior mean of f at a point x is given by

$$\frac{\sum_{j=1}^{\infty}\int_{\boldsymbol{\theta}} f(x)p(\boldsymbol{X}|J=j,\boldsymbol{\theta})\Pi(\boldsymbol{\theta}|J=j)\Pi(J=j)d\boldsymbol{\theta}}{\sum_{j=1}^{\infty}\int_{\boldsymbol{\theta}} p(\boldsymbol{X}|J=j,\boldsymbol{\theta})\Pi(\boldsymbol{\theta}|J=j)\Pi(J=j)d\boldsymbol{\theta}}$$
$$= \frac{\sum_{j=1}^{\infty}\Pi(j)\sum_{i_0=1}^{j}\sum_{i_1=1}^{j}\cdots\sum_{i_n=1}^{j}\int_{\boldsymbol{\theta}\in\Omega_j}\prod_{k=1}^{j}\theta_k^{a_k-1}\prod_{s=0}^{n}\theta_{i_s}B_{i_s}^*(X_s)d\boldsymbol{\theta}}{\sum_{j=1}^{\infty}\Pi(j)\sum_{i_1=1}^{j}\cdots\sum_{i_n=1}^{j}\int_{\boldsymbol{\theta}\in\Omega_j}\prod_{k=1}^{j}\theta_k^{a_k-1}\prod_{s=1}^{n}\theta_{i_s}B_{i_s}^*(X_s)d\boldsymbol{\theta}}, \quad (3.44)$$

where X_0 stands for x. Define $I_{k,j,0}^{\boldsymbol{i}} = \sum_{s=0}^{n}\mathbb{1}\{i_s = k\}$ and $I_{k,j,1}^{\boldsymbol{i}} = \sum_{s=1}^{n}\mathbb{1}\{i_s = k\}$. Then the expression in (3.44) can be simplified to

$$\frac{\sum_{j=1}^{\infty}\Pi(j)\sum_{i_0=1}^{j}\sum_{i_1=1}^{j}\cdots\sum_{i_n=1}^{j}\prod_{k=1}^{j}\Gamma(a_k+I_{k,j,0}^{\boldsymbol{i}})\prod_{s=0}^{n}B_{i_s}^*(X_s)/\Gamma\Big(\sum_{i=1}^{j}a_i+n+1\Big)}{\sum_{j=1}^{\infty}\Pi(j)\sum_{i_1=1}^{j}\cdots\sum_{i_n=1}^{j}\prod_{k=1}^{j}\Gamma(a_k+I_{k,j,1}^{\boldsymbol{i}})\prod_{s=1}^{n}B_{i_s}^*(X_s)/\Gamma\Big(\sum_{i=1}^{j}a_i+n\Big)}. \quad (3.45)$$

A basis function takes nonzero values only at q intervals, so the calculation involves a multiple of q^{n+1} steps. More details are given in Section 3.11. Similar expressions can be obtained for other posterior moments, in particular, for the posterior variance.

Remark 7. Note that the adaptive Bayesian estimator has a connection with histogram and kernel methods. If $q = 1$, the sums over indices $i_1, \ldots, i_n$ in (3.45) will be redundant as there is only one term. Hence it gives a histogram estimate whose bin length and weights are posterior averaged, which is similar to the Bayes estimator in Gasparini (1996). Although doing so brings easy computation, it cannot adapt to the smoothness level greater than 1. The B-spline random series prior can also be viewed as a kernel mixture prior, where the kernel is a B-spline function indexed by a discrete parameter and the kernel takes positive values only on a finite interval. When q and J are chosen larger, the kernel becomes more flat.

3.5 Whittle estimation of spectral density

3.5.1 Posterior convergence rates

Consider a second order stationary time series $\{X_t, t \in \mathbb{Z}\}$ with mean 0 and autocovariance function $\gamma_r = \mathrm{E}(X_t X_{t+r})$. Assume that $\sum_r |\gamma_r| < \infty$. Define the spectral density of $\{X_t\}$ by $f(\lambda) = (2\pi)^{-1} \sum_{r=-\infty}^{\infty} \gamma_r e^{-ir\pi\lambda}$ on $(0, 1)$. Let $I_n(\lambda) = (2\pi n)^{-1} |\sum_{t=1}^{n} X_t e^{-it\pi\lambda}|^2$ be the periodogram. Instead of using the true likelihood, which is complicated even for a Gaussian time series, Whittle (1957) proposed using an approximate likelihood. Let $\nu = \lfloor n/2 \rfloor$ and $\omega_j = 2j/n$, $j = 1, \ldots, \nu$. The Whittle likelihood is formed by pretending that $U_j = I_n(\omega_j)$, $j = 1, \ldots, \nu$, with means $f(\omega_j)$, are exponentially distributed. The advantage of using the Whittle likelihood is that it is an explicit function of the spectral density f, rather than being expressed in terms of infinitely many autocorrelation coefficients.

A nonparametric Bayesian method has been used by Choudhuri et al. (2004a) where the prior is constructed via Bernstein polynomials. They also gave a posterior consistency result. Posterior convergence rates can be obtained by a contiguity result as established in Choudhuri et al. (2004b). Their results state that for a Gaussian time series, the distributions of $(U_1, \ldots, U_\nu)$ given by the exact and the Whittle likelihood are contiguous. As long as asymptotic convergence results are concerned, this allows us to work under the pretending assumption that $U_1, \ldots, U_\nu$ are actually independent and exponentially distributed.

Assume that the true spectral density $f_0 \in \mathcal{C}^\alpha[0, 1]$ is positive throughout. A prior is constructed on f using a monotonic link function Ψ. If the B-splines are used to construct a basis, then we can also use a link function $\Psi : (0, \infty) \to (0, \infty)$ such as $\Psi(x) = x$ or $\Psi(x) = x^{-1}$ and construct a prior for $\boldsymbol{\theta}$ on $(0, \infty)^J$. If $\Psi(x) = e^x$, then the prior for $\boldsymbol{\theta}$ is constructed on $\mathbb{R}^J$.

Define a discretized L_2-distance as follows:

$$d_n^2(f_1, f_2) = \nu^{-1} \sum_{i=1}^{\nu} \{f_1(2i/n) - f_2(2i/n)\}^2. \tag{3.46}$$

Then we have the following theorem.

Theorem 14. *Assume that the true spectral density f_0 is strictly positive and satisfies $\Psi^{-1}(f_0) \in \mathcal{C}^\alpha[0,1]$, $\alpha \geq 1$ for some monotonic, Lipschitz continuous link function Ψ. Also the prior satisfies Conditions (A2) and (A3).*

(a) If the basis $\boldsymbol{\xi}$ satisfies Condition (B1), then the posterior distribution converges at the rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$ with respect to the d_n.

(b) If the basis $\boldsymbol{\xi}$ satisfies Condition (B2), then the posterior distribution converges at the rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{(\alpha+1)/(2\alpha+1)+(1-t_1)/2}$ with respect to the d_n.

(c) If Ψ is the identity or the reciprocal function and the B-spline basis is used, then the prior on $\boldsymbol{\theta}$ can be restricted to $(0,\infty)^J$ to obtain the convergence rate in (a).

Proof. Let f_0 take values in some compact interval $[m, M] \subset (0, \infty)$. Since $\alpha \geq 1$, observe that

$$\|f_0 - \Psi(\boldsymbol{\theta_0}^T \boldsymbol{\xi})\|_\infty \lesssim \|\Psi^{-1}(f_0) - \boldsymbol{\theta_0}^T \boldsymbol{\xi}\|_\infty \lesssim J^{-\alpha}$$

for some $\boldsymbol{\theta_0} \in \mathbb{R}^J$. Using results in Section 7.3 of Ghosal and van der Vaart (2007a),

$$\max\left\{\nu^{-1}\sum_{i=1}^{\nu} K(P_{f_0,i}, P_{f,i}), \nu^{-1}\sum_{i=1}^{\nu} V(P_{f_0,i}, P_{f,i})\right\} \lesssim d_n^2(f_0, f) \lesssim \|f_0 - f\|_\infty^2. \tag{3.47}$$

Note that the average squared Hellinger distance h_n is bounded by d_n. Applying Theorem 11 for d as the supremum distance, $\boldsymbol{\eta}$ as vacuous, $\bar{J}_n = (n/\log n)^{1/(2\alpha+1)}$, $\bar{\epsilon}_n = (n/\log n)^{-\alpha/(2\alpha+1)}$,

$J_n = n^{1/(2\alpha+1)}(\log n)^{2\alpha/(2\alpha+1)-t_1}$, $M_n = n^{1/t_3}$ and using the fact that the L_2-distance is bounded by the Hellinger distance as f_0 is bounded, we obtain Part (a).

Choosing $\bar{J}_n = (n/\log n)^{1/(2\alpha+1)}$, $M_n = n^{1/t_3}$, $J_n = n^{1/(2\alpha+1)}(\log n)^{(2\alpha+2)/(2\alpha+1)-t_1}$, d as the supremum distance, $\boldsymbol{\eta}$ as vacuous and $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}(\log n)^{(\alpha+1)/(2\alpha+1)}$, we obtain Part (b).

Part (c) is a consequence of Part (b), Lemma 7 and the calculation in Part (a). □

3.5.2 Computation using B-splines

We use B-splines as basis functions and choose the link function $\Psi(x) = x^{-1}$. Then the likelihood is given by

$$\frac{1}{\prod_{s=1}^{\nu} f(\omega_s)} \exp\left\{ -\sum_{s=1}^{\nu} U_s / f(\omega_s) \right\}. \tag{3.48}$$

We consider independent gamma prior distributions on $\boldsymbol{\theta}$: $\theta_i \overset{\text{ind}}{\sim} \text{Gamma}(a_i, b_i)$ for some positive numbers a_i and b_i. The approximation property holds by Theorem 14, Part (c). By a calculation similar to that in Subsection 3.4.3, the posterior mean of f_0^{-1} is given by a ratio, whose numerator is

$$\begin{aligned} \sum_{j=1}^{\infty} \Pi(j) \sum_{i_0=1}^{j} \cdots \sum_{i_\nu=1}^{j} \int_{\boldsymbol{\theta}>0} \prod_{k=1}^{j} \frac{\theta_k^{a_k-1} b_k^{a_k}}{\Gamma(a_k)} \prod_{s=0}^{\nu} \theta_{i_s} B_{i_s}(\omega_s) \\ \times \exp\Big\{ -\sum_{s=1}^{\nu} U_s \sum_{i=1}^{j} \theta_i B_i(\omega_s) - \sum_{i=1}^{j} b_i \theta_i \Big\} d\boldsymbol{\theta} \end{aligned} \tag{3.49}$$

and the denominator is

$$\sum_{j=1}^{\infty}\Pi(j)\sum_{i_1=1}^{j}\cdots\sum_{i_\nu=1}^{j}\int_{\boldsymbol{\theta}>0}\prod_{k=1}^{j}\frac{\theta_k^{a_k-1}b_k^{a_k}}{\Gamma(a_k)}\prod_{s=1}^{\nu}\theta_{i_s}B_{i_s}(\omega_s)$$
$$\times\exp\Big\{-\sum_{s=1}^{\nu}U_s\sum_{i=1}^{j}\theta_iB_i(\omega_s)-\sum_{i=1}^{j}b_i\theta_i\Big\}d\boldsymbol{\theta}. \tag{3.50}$$

Define $I^{\boldsymbol{i}}_{k,j,0}=\sum_{s=0}^{\nu}\mathbb{1}\{i_s=k\}$ and $I^{\boldsymbol{i}}_{k,j,1}=\sum_{s=1}^{\nu}\mathbb{1}\{i_s=k\}$. Then the posterior mean of $\{f(\omega_0)\}^{-1}$ can be simplified to

$$\frac{\displaystyle\sum_{j=1}^{\infty}\Pi(j)\sum_{i_0=1}^{j}\cdots\sum_{i_\nu=1}^{j}\prod_{s=0}^{\nu}B_{i_s}(\omega_s)\prod_{k=1}^{j}\frac{b_k^{a_k}}{\{b_k+\sum_{s=1}^{\nu}U_sB_k(\omega_s)\}^{I^{\boldsymbol{i}}_{k,j,0}+a_k}}\frac{\Gamma(I^{\boldsymbol{i}}_{k,j,0}+a_k)}{\Gamma(a_k)}}{\displaystyle\sum_{j=1}^{\infty}\Pi(j)\sum_{i_1=1}^{j}\cdots\sum_{i_\nu=1}^{j}\prod_{s=1}^{\nu}B_{i_s}(\omega_s)\prod_{k=1}^{j}\frac{b_k^{a_k}}{\{b_k+\sum_{s=1}^{\nu}U_sB_k(\omega_s)\}^{I^{\boldsymbol{i}}_{k,j,1}+a_k}}\frac{\Gamma(I^{\boldsymbol{i}}_{k,j,1}+a_k)}{\Gamma(a_k)}}. \tag{3.51}$$

3.6 Nonparametric regression with Gaussian errors

3.6.1 Posterior convergence rates

We consider a nonparametric regression model with additive error $X_i=f(Z_i)+\epsilon_i$, where $\epsilon_i\overset{\text{iid}}{\sim}\mathrm{N}(0,\sigma^2)$, and σ is an unknown parameter. The covariates $Z_1,\ldots,Z_n$ can be either fixed or random and are assumed to be in $(0,1)$. We shall use the same prior on f as in Section 3.2.2. Also, we put a prior distribution on σ that satisfies tail conditions for sufficiently small $\sigma_1,\sigma_2,\sigma_3>0$, any fixed $\sigma_0>0$, some positive numbers t_4, t_5 and t_6:

$$\Pi(\sigma<\sigma_1)\lesssim\exp\{-\exp(\sigma_1^{-t_4})\}, \tag{3.52}$$

$$\Pi(\sigma>\sigma_2^{-1})\lesssim\sigma_2^{t_5}, \tag{3.53}$$

$$\Pi(|\sigma-\sigma_0|<\sigma_3)\gtrsim\sigma_3^{t_6}. \tag{3.54}$$

Condition (3.52) is automatically satisfied if σ has a confirmed lower bound as assumed in earlier Bayesian papers. For fixed covariates $\boldsymbol{Z}$, define a empirical measure $\mathbb{P}_n^Z = n^{-1}\sum_{i=1}^n \delta_{Z_i}$, the covariance matrix $\Sigma_{n,J}$, whose (k,l)-th element is $\int \xi_k \xi_l d\mathbb{P}_n^Z$ for $k, l \leq J$ and $\|\cdot\|_{2,n}$ as the norm on $L_2(\mathbb{P}_n^Z)$. For random covariates $Z_1, \ldots, Z_n \sim G$, define $L_{2,G}$ as the L_2-distance with respect to the probability measure G. When the domain of Z_i's is $\mathbb{R}$, we consider pseudo-metrics $\|\cdot\|_{2,n}^*$ given by $\left(\|f_1 - f_2\|_{2,n}^*\right)^2 = \int_a^b |f_1 - f_2|^2 d\mathbb{P}_n^Z$ and $\|\cdot\|_{2,G}^*$ defined by $\left(\|f_1 - f_2\|_{2,G}^*\right)^2 = \int_a^b |f_1 - f_2|^2 dG$, as in Section 3.4.2. Then the posterior convergence rate is obtained below.

Theorem 15. *Suppose that the true regression function $f_0 \in \mathcal{C}^\alpha[0,1]$ and the prior satisfies Conditions (A2), (A3) and relations* (3.52)*,* (3.53) *and* (3.54)*.*

(a) If f_0 is defined on $(0,1)$ and the covariates $Z_1, \ldots, Z_n$ are fixed, then the posterior converges at the rate

(i) $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$ if the basis $\boldsymbol{\xi}$ satisfies Condition (B1),

(ii) $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{(\alpha+1)/(2\alpha+1)+(1-t_1)/2}$ if $\boldsymbol{\xi}$ satisfies Condition (B2)

relative to $\|\cdot\|_{2,n}$.

(b) If the covariates $Z_1, \ldots, Z_n \sim G$ are random, G has a bounded density g and $\boldsymbol{\xi}$ satisfies Condition (B3), then the posterior converges at $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$ with respect to $\|\cdot\|_2$ when f_0 is defined on $(0,1)$.

(c) If the support of f_0 is unbounded, e.g., the real line, then the posterior converges at the same rate as in (b) with respect to $\|\cdot\|_{2,n}^*$.

Proof. We first consider Part (a). Define $P_{f,i}$ as the normal measure with mean $f(Z_i)$ and variance σ^2. Results in Birgé (2006) imply that the likelihood ratio test for f_0 versus another f

satisfy conditions in Lemma 2 of Ghosal and van der Vaart (2007a) with respect to $\|\cdot\|_{2,n}$; see the latter paper for details. This implies that we can work on $\|\cdot\|_{2,n}$ instead of the Hellinger distance. Using the arguments in Section 7.2 of Ghosal and van der Vaart (2007a), we get

$$\max\left\{n^{-1}\sum_{i=1}^{n}K(P_{f_0,i},P_{f,i}),n^{-1}\sum_{i=1}^{n}V(P_{f_0,i},P_{f,i})\right\}\leq\|f_0-f\|_n^2/\sigma^2. \tag{3.55}$$

Using the fact that $\|f_0-f\|_{2,n}^2\lesssim\|f_0-f\|_\infty^2$, there exists $\boldsymbol{\theta_0}\in\mathbb{R}^J$, $\|\boldsymbol{\theta_0}\|_\infty<\infty$ such that

$$\max\left\{n^{-1}\sum_{i=1}^{n}K(P_{f_0,i},P_{f,i}),n^{-1}\sum_{i=1}^{n}V(P_{f_0,i},P_{f,i})\right\}\lesssim J^{-2\alpha}/\sigma_0^2. \tag{3.56}$$

In order to apply Theorem 11, we define $\boldsymbol{\eta}$ as σ and choose $\mathcal{F}_n=((\log n)^{-1/t_4},\exp\{an\bar{\epsilon}_n^2/t_5\})$ for some sufficiently large $a>0$. Note that

$$\Pi(\mathcal{F}_n^c)\lesssim\exp\{-\exp(\log n)\}+\exp\{-an\bar{\epsilon}_n^2\}\lesssim\exp\{-an\bar{\epsilon}_n^2\},$$

$$\log D(\epsilon_n,\mathcal{F}_n,\|\cdot\|_2)\leq\log\{e^{n\bar{\epsilon}_n^2/t_5}/\epsilon_n\}\lesssim n\bar{\epsilon}_n^2/t_5+\log n\leq n\epsilon_n^2$$

given n is sufficiently large. Then (3.21), (3.25) and (3.26) hold. Choosing

$$\bar{J}_n=(n/\log n)^{1/(2\alpha+1)},\bar{\epsilon}_n=(n/\log n)^{-\alpha/(2\alpha+1)},J_n=n^{1/(2\alpha+1)}(\log n)^{2\alpha/(2\alpha+1)-t_1},$$

d as the supremum distance and $M_n=n^{1/t_3}$, we obtain the posterior convergence results for h_n. Note that $\|\cdot\|_{2,n}$ satisfies conditions in Lemma 2 of Ghosal and van der Vaart (2007a). Hence h_n can be replaced by $\|\cdot\|_{2,n}$.

For Part (b), note that $\|f_0-f\|_{2,G}^2\lesssim\|f_0-f\|_2^2$ since g is bounded. Hence by applying

Theorem 11 for d as the L_2-distance,

$$\bar{J}_n = (n/\log n)^{1/(2\alpha+1)},\ M_n = n^{1/t_3},$$

$$J_n = n^{1/(2\alpha+1)}(\log n)^{(2\alpha+2)/(2\alpha+1)-t_1},\ \bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}(\log n)^{(\alpha+1)/(2\alpha+1)},$$

the result follows.

Part (c) proceeds as in Section 3.4.2. □

3.6.2 MCMC-free Computation

Define an $n \times J$ matrix $\boldsymbol{D}_J$ whose (s,t)-entry is $\xi_t(Z_s)$. Then the model can be written as $\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\theta}, J, \sigma \sim \mathrm{N}(\boldsymbol{D}_J\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$. We consider the Zellner's g-prior (Zellner, 1986) on $\boldsymbol{\theta}$. Namely $\boldsymbol{\theta}|\sigma^2, \boldsymbol{D}_J, J \sim \mathrm{N}(\boldsymbol{\mu_0}, g\sigma^2(\boldsymbol{D}_J^T\boldsymbol{D}_J)^{-1})$ for some $g > 0$. The posterior of $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Z}, \sigma, J \sim N\left(\frac{(\boldsymbol{D}_J^T\boldsymbol{D}_J)^{-1}}{g+1}(g\boldsymbol{D}_J^T\boldsymbol{X} + \boldsymbol{D}_J^T\boldsymbol{D}_J\boldsymbol{\mu_0}), \frac{g\sigma^2}{g+1}(\boldsymbol{D}_J^T\boldsymbol{D}_J)^{-1}\right). \tag{3.57}$$

Assume $\boldsymbol{\mu_0} = \mathbf{0}$, by calculations similar to those in Subsection 3.4.3, the posterior mean of f_0 at point Z_0 is a ratio between

$$\sum_{j=1}^{\infty} \Pi(j)(2\pi)^{-(j+n)/2} \det(\boldsymbol{D}_j^T\boldsymbol{D}_j)^{1/2}\boldsymbol{B}_{\mathbf{0}}^T$$
$$\times \int_{\sigma>0}\int_{\boldsymbol{\theta}\in\mathbb{R}^j} \sigma^{-(n+j)}\Pi(\sigma)\boldsymbol{\theta}\exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{X}^T\boldsymbol{X} - 2\boldsymbol{X}^T\boldsymbol{D}_j\boldsymbol{\theta} + \frac{g+1}{g}\boldsymbol{\theta}^T\boldsymbol{D}_j^T\boldsymbol{D}_j\boldsymbol{\theta}\right)\right\} d\boldsymbol{\theta}d\sigma$$

and

$$\sum_{j=1}^{\infty} \Pi(j)(2\pi)^{-(j+n)/2} \det(\boldsymbol{D}_j^T \boldsymbol{D}_j)^{1/2}$$
$$\times \int_{\sigma>0} \int_{\boldsymbol{\theta}\in\mathbb{R}^j} \sigma^{-(n+j)} \Pi(\sigma) \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{X}^T\boldsymbol{X} - 2\boldsymbol{X}^T\boldsymbol{D}_j\boldsymbol{\theta} + \frac{g+1}{g}\boldsymbol{\theta}^T\boldsymbol{D}_j^T\boldsymbol{D}_j\boldsymbol{\theta}\right)\right\} d\boldsymbol{\theta} d\sigma,$$

where $\boldsymbol{B_0} = \boldsymbol{B}(Z_0)$. By integration, the ratio can be simplified to

$$\frac{g\sum_{j=1}^{\infty} \Pi(j)\boldsymbol{B}_{\boldsymbol{0}}^T(\boldsymbol{D}_j^T\boldsymbol{D}_j)^{-1}\boldsymbol{D}_j^T\boldsymbol{X}\int_{\sigma>0}\Pi(\sigma)\sigma^{-n}\exp\left\{-\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P}_j)\boldsymbol{X}/2\sigma^2\right\}d\sigma}{(g+1)\sum_{j=1}^{\infty}\Pi(j)\int_{\sigma>0}\Pi(\sigma)\sigma^{-n}\exp\left\{-\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P}_j)\boldsymbol{X}/2\sigma^2\right\}d\sigma}, \quad (3.58)$$

where $\boldsymbol{P}_j = \frac{g}{g+1}\boldsymbol{D}_j(\boldsymbol{D}_j^T\boldsymbol{D}_j)^{-1}\boldsymbol{D}_j^T$.

3.7 Nonparametric binary regression

3.7.1 Posterior convergence rates

A prior on the response probability function in a binary regression model is commonly induced from a Gaussian process. A posterior consistency result was obtained by Ghosal and Roy (2006) while posterior convergence rates were obtained by van der Vaart and van Zanten (2008, 2009). Useful computational methods were developed by Choudhuri et al. (2007) using the technique of Albert and Chib (1993) and by Rasmussen and Williams (2006).

Assume that we have n independent observations $(Z_1, X_1), \ldots, (Z_n, X_n)$ from a binary regression model $\mathrm{P}(X = 1|Z = z) = 1 - \mathrm{P}(X = 0|Z = z) = f_0(z)$, where X's take values in $\{0, 1\}$ and Z's are either fixed or random covariates in some domain $\mathcal{Z}$. Given a link function $\Psi : \mathcal{Z} \to (0, 1)$, we can construct a random series prior on the regression function f_0 using a basis $\boldsymbol{\xi}$ as $f_{\boldsymbol{\theta}}(z) = \Psi\{\boldsymbol{\theta}^T\boldsymbol{\xi}(z)\}$. Commonly, a cumulative distribution function on $\mathbb{R}$ such as

the logit or probit function is chosen as the link function and the coefficient vector $\boldsymbol{\theta}$ can be chosen any values in $\mathbb{R}^J$. Then any basis with approximation property for the Hölder class may be used.

For random covariates Z, the likelihood function for (Z, X) can be written as

$$L_{\boldsymbol{\theta}}(z, x) = f_{\boldsymbol{\theta}}(z)^x (1 - f_{\boldsymbol{\theta}}(z))^{1-x} g(z), \tag{3.59}$$

where g is the marginal density of Z. When integrating out the posterior distribution, g will be canceled out. Therefore, in the following study, we can absorb g into the dominating measure, denoted it by G and remove g from the model (3.59). Define the empirical measure $\mathbb{P}_n^Z = n^{-1} \sum_{i=1}^n \delta_{Z_i}$, the covariance matrix $\Sigma_{n,J}$, whose (k, l)-th element is $\int \xi_k \xi_l d\mathbb{P}_n^Z$ for $k, l \leq J$ and $\|\cdot\|_{2,n}$ as the norm on $L_2(\mathbb{P}_n^Z)$ for fixed covariates Z. For random covariates $Z_1, \ldots, Z_n \sim G$, define $L_{2,G}$ as the L_2-distance with respect to the probability measure G.

Theorem 16. *Suppose that $\Psi^{-1}(f_0) \in \mathcal{C}^\alpha[0, 1]$ is bounded away from 0 and 1, and the link function Ψ has a bounded derivative ψ such that $\psi/(\Psi(1 - \Psi))$ is bounded. Assume that the prior satisfies Conditions (A2) and (A3).*

(a) If the covariates $Z_1, \ldots, Z_n$ are fixed, then the posterior converges at the rate

(i) $\epsilon_n = n^{-\alpha/(2\alpha+1)} (\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$ if the basis satisfies Condition (B1),

(ii) $\epsilon_n = n^{-\alpha/(2\alpha+1)} (\log n)^{(\alpha+1)/(2\alpha+1)+(1-t_1)/2}$ if the basis satisfies Condition (B2)

relative to $\|\cdot\|_{2,n}$.

(b) If the covariates $Z_1, \ldots, Z_n \sim G$ are random, G has a bounded density g and $\boldsymbol{\xi}$ satisfies Condition (B3), then the posterior convergence rate is $n^{-\alpha/(2\alpha+1)} (\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$ with respect to $\|\cdot\|_2$.

(c) If Ψ is the identity and the B-spline basis is used, then $\boldsymbol{\theta}$ can be restricted in $(0,1)^J$ and maintain the rate in (a)(i) and (b).

Proof. We first consider the case of random covariates. Define $p_{\boldsymbol{\theta}} = \Psi(\boldsymbol{\theta}^T\boldsymbol{\xi})^x(1-\Psi(\boldsymbol{\theta}^T\boldsymbol{\xi}))^{1-x}$ and $p_{f_0} = f_0^x(1-f_0)^{1-x}$, By a similar calculation in Lemma 3.2 of van der Vaart and van Zanten (2008), for any $\boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \mathbb{R}^J$, $J \geq 1$ and $r > 1$, we have the following:

$$\|p_{\boldsymbol{\theta_1}} - p_{\boldsymbol{\theta_2}}\|_r \lesssim \|\boldsymbol{\theta_1} - \boldsymbol{\theta_2}\|_{r,G} \tag{3.60}$$

$$\max\left\{K(p_{\boldsymbol{\theta_1}}, p_{f_0}), V(p_{\boldsymbol{\theta_1}}, p_{f_0})\right\} \lesssim \|\Psi(\boldsymbol{\theta_1}^T\boldsymbol{\xi}) - f_0\|_{2,G}^2. \tag{3.61}$$

Since $\psi/(\Psi(1-\Psi))$ is bounded, $p_{\boldsymbol{\theta}}$ is uniformly bounded below and above. Hence it suffices to work with $\|\cdot\|_{2,G}$ on $f_{\boldsymbol{\theta}}$ instead of the Hellinger distance on $p_{\boldsymbol{\theta}}$. Choose $r = 2$ in (3.60), (3.61) and apply Theorem 11 for $\bar{J}_n = (n/\log n)^{1/(2\alpha+1)}$, $M_n = n^{1/t_3}$, $J_n = n^{1/(2\alpha+1)}(\log n)^{(2\alpha+2)/(2\alpha+1)-t_1}$, $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}(\log n)^{(\alpha+1)/(2\alpha+1)}$ and $\boldsymbol{\eta}$ as the vacuous parameter, the posterior convergence results follows for h_n. Note that $\|\cdot\|_2$ satisfies conditions in Lemma 2 of Ghosal and van der Vaart (2007a). As a result, h_n can be replaced by $\|\cdot\|_2$ and the posterior convergence rate holds for $\|\cdot\|_2$ as well.

The proof of Part (a) proceeds as in Part (a) of Theorem 15. Part (c) is a consequence of Part (c) of Lemma 10. □

3.7.2 Computation using B-splines

When specifically using B-splines as the basis, some simpler choices of link function Ψ are possible in view of Parts (b) and (c) of Lemma 10. For example, if we choose Ψ as the identity link, then a prior for $\boldsymbol{\theta}$ can be constructed on $(0,1)^J$. If we choose $\Psi(z) = z/(1+z)$, then the prior for $\boldsymbol{\theta}$ can be constructed on $(0,\infty)^J$.

As in Section 3.4.3, the likelihood function simplifies significantly when we choose the identity link function and use beta priors $\theta_i \overset{\text{ind}}{\sim} \text{Beta}(a_i, b_i)$ for some positive numbers a_i and b_i. This expresses the posterior mean of $f_0(z)$ as a ratio whose numerator is

$$\sum_{j=1}^{\infty} \Pi(j) \prod_{i=1}^{j} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \sum_{i_0=1}^{j} \cdots \sum_{i_n=1}^{j} \int_0^1 \prod_{k=1}^{j} \theta_k^{a_k - 1}(1-\theta_k)^{b_k - 1} \prod_{s=0}^{n} \theta_{i_s}^{X_s}(1-\theta_{i_s})^{1-X_s} B_{i_s}(Z_s) d\boldsymbol{\theta} \tag{3.62}$$

and the denominator is

$$\sum_{j=1}^{\infty} \Pi(j) \prod_{i=1}^{j} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \sum_{i_1=1}^{j} \cdots \sum_{i_n=1}^{j} \int_0^1 \prod_{k=1}^{j} \theta_k^{a_k - 1}(1-\theta_k)^{b_k - 1} \prod_{s=1}^{n} \theta_{i_s}^{X_s}(1-\theta_{i_s})^{1-X_s} B_{i_s}(Z_s) d\boldsymbol{\theta}, \tag{3.63}$$

where $Z_0 = z$. Define $I^{\boldsymbol{i}}_{k,j,0} = \sum_{s=0}^{n} \mathbb{1}\{i_s = k\}$ and $I^{\boldsymbol{i}}_{k,j,1} = \sum_{s=1}^{n} \mathbb{1}\{i_s = k\}$. The expression in (3.62) simplifies to

$$\sum_{j=1}^{\infty} \Pi(j) \prod_{i=1}^{j} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \sum_{i_0=1}^{j} \cdots \sum_{i_n=1}^{j} \prod_{s=0}^{n} B_{i_s}(Z_s) \prod_{k=1}^{j} \frac{\Gamma\Big(a_k + \sum\limits_{\substack{t=0,\ldots,j \\ i_t = k}} X_t\Big)\Gamma\Big(b_k + I^{\boldsymbol{i}}_{k,j,0} - \sum\limits_{\substack{t=0,\ldots,j \\ i_t = k}} X_t\Big)}{\Gamma(a_k + b_k + I^{\boldsymbol{i}}_{k,j,0})} \tag{3.64}$$

while the expression in (3.63) simplifies to

$$\sum_{j=1}^{\infty} \Pi(j) \prod_{i=1}^{j} \frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)} \sum_{i_1=1}^{j} \cdots \sum_{i_n=1}^{j} \prod_{s=1}^{n} B_{i_s}(Z_s) \prod_{k=1}^{j} \frac{\Gamma\Big(a_k + \sum\limits_{\substack{t=1,\ldots,j \\ i_t=k}} X_t\Big)\Gamma\Big(b_k + I^{\boldsymbol{i}}_{k,j,1} - \sum\limits_{\substack{t=1,\ldots,j \\ i_t=k}} X_t\Big)}{\Gamma(a_k+b_k+I^{\boldsymbol{i}}_{k,j,1})}. \tag{3.65}$$

3.8 Nonparametric Poisson regression

3.8.1 Posterior convergence rates

Consider a Poisson regression model $X_i \overset{\text{ind}}{\sim} \text{Poi}(f(Z_i))$, where f is an unknown function and Z_i's are covariates. For convenience, we assume that Z_i's are in $(0,1)$. Using a random series expansion, f can be modeled through a link function $f(z) = \Psi(\boldsymbol{\theta}^T \boldsymbol{\xi})(z)$.

The adaptation results can be obtained as in Section 3.6. Define the empirical measure $\mathbb{P}_n^Z = n^{-1}\sum_{i=1}^n \delta_{Z_i}$, the covariance matrix $\Sigma_{n,j}$, whose (k,l)-th element equals to $\int \xi_k \xi_l d\mathbb{P}_n^Z$ and $\|\cdot\|_{2,n}$ as the norm on $L_2(\mathbb{P}_n^Z)$. For random covariates $Z_1, \ldots, Z_n \sim G$, define $L_{2,G}$ as the L_2-distance with respect to the probability measure G. The posterior rates theorem is stated as follows.

Theorem 17. *Suppose that the function $\Psi^{-1}(f_0) \in \mathcal{C}^{\alpha}[0,1]$ and f_0 is strictly positive. Assume that the prior satisfies (A2) and (A3) and the link function Ψ is monotonic and Lipschitz continuous.*

(a) If the covariates $Z_1, \ldots, Z_n$ are fixed, then the posterior converges at the rate

(i) $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$ if the basis satisfies Condition (B1),

(ii) $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{(\alpha+1)/(2\alpha+1)+(1-t_1)/2}$ if the basis satisfies Condition (B2)

relative to $\|\cdot\|_{2,n}$.

(b) If the covariates $Z_1, \ldots, Z_n \sim G$ *are random,* G *has a bounded density* g *and* $\boldsymbol{\xi}$ *satisfies Condition (B3), then the posterior convergence rate is* $n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$ *with respect to* $\|\cdot\|_2$.

(c) If Ψ *is the identity link function and the B-spline basis is used, then* $\boldsymbol{\theta}$ *can be restricted to* $(0,\infty)^J$ *to maintain the convergence rate in (a)(i).*

Proof. By applying the arguments in Section 7.1.1 of Ghosal and van der Vaart (2007a), we only need approximation results for L_2-distance on f when Z's are random and L_∞-approximation results on f for fixed Z's. For a basis satisfying Condition $(B1)$, we apply Theorem 11 with d as the L_∞-distance, $\boldsymbol{\eta}$ as vacuous, $\bar{J}_n = (n/\log n)^{1/(2\alpha+1)}$, $M_n = n^{1/t_3}$, $J_n = n^{1/(2\alpha+1)}(\log n)^{(2\alpha+2)/(2\alpha+1)-t_1}$ and $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}(\log n)^{(\alpha+1)/(2\alpha+1)}$; for a basis satisfying $(B3)$, apply the same set of sequences and choose d as the L_2-distance, for a basis satisfying $(B2)$, apply $\bar{J}_n = (n/\log n)^{1/(2\alpha+1)}$, $M_n = n^{1/t_3}$, $J_n = n^{1/(2\alpha+1)}(\log n)^{(2\alpha+2)/(2\alpha+1)-t_1}$ and $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}(\log n)^{(\alpha+1)/(2\alpha+1)}$. Then Parts (a) and (b) are obtained.

Part (c) is a consequence of Part (b) of Theorem 10 and the arguments above. □

3.8.2 Computation using B-splines

If we use B-splines to form the basis, in view of Part (c) of Lemma 10, we are allowed to restrict θ_j's to positive values. By choosing the identity link, then it is possible to carry out MCMC-free computation by letting $\theta_i \overset{\text{ind}}{\sim} \text{Gamma}(a_i, b_i)$ for some positive numbers a_i and b_i.

The posterior mean of $f(z)$ is a ratio whose numerator is

$$\sum_{j=1}^{\infty}\Pi(j)\int_0^{\infty}\sum_{s=1}^{j}\theta_s B_s(z)\prod_{k=1}^{j}\frac{b_k^{a_k}\theta_k^{a_k-1}}{\Gamma(a_k)} \times \exp\Big\{-\sum_{i=1}^{j}\theta_i\Big(\sum_{k=1}^{n}B_i(Z_k)+b_i\Big)\Big\}\prod_{k=1}^{n}\Big\{\sum_{i=1}^{j}\theta_i B_i(Z_k)\Big\}^{X_k}d\boldsymbol{\theta} \tag{3.66}$$

and the denominator is

$$\sum_{j=1}^{\infty}\Pi(j)\int_0^{\infty}\prod_{k=1}^{j}\frac{b_k^{a_k}\theta_k^{a_k-1}}{\Gamma(a_k)} \times \exp\Big\{-\sum_{i=1}^{j}\theta_i\Big(\sum_{k=1}^{n}B_i(Z_k)+b_i\Big)\Big\}\prod_{k=1}^{n}\Big\{\sum_{i=1}^{j}\theta_i B_i(Z_k)\Big\}^{X_k}d\boldsymbol{\theta}. \tag{3.67}$$

Denote $Z_0 = z$, $X_0 = 1$. Define $\boldsymbol{b} = (b_1, \ldots, b_J)^T$, $\boldsymbol{s}(k)$ as the k-th element of vector $\boldsymbol{s}$ and $\boldsymbol{B}(z)^{\boldsymbol{s}} = \prod_{i=1}^{j} B_i(z)^{s_i}$ for some index vector $\boldsymbol{s}$, the above quantity simplifies to

$$\frac{\displaystyle\sum_{j=1}^{\infty}\Pi(j)\sum_{\substack{\boldsymbol{s_0},\ldots,\boldsymbol{s_n}\in\mathbb{N}_0^j\\ |\boldsymbol{s_0}|=X_0,\ldots,|\boldsymbol{s_n}|=X_n\\ \boldsymbol{s}=\boldsymbol{s_0}+\cdots+\boldsymbol{s_n}}}\prod_{i=0}^{n}\frac{\{\boldsymbol{B}(Z_i)\}^{\boldsymbol{s_i}}}{\boldsymbol{s_i}!}\prod_{k=1}^{j}\frac{\Gamma(a_k+\boldsymbol{s}(k))b_k^{a_k}}{\Gamma(a_k)(b_k+\sum_{t=0}^{n}B_k(Z_t))^{a_k+\boldsymbol{s}(k)}}}{\displaystyle\sum_{j=1}^{\infty}\Pi(j)\sum_{\substack{\boldsymbol{s_1},\ldots,\boldsymbol{s_n}\in\mathbb{N}_0^j\\ |\boldsymbol{s_1}|=X_1,\ldots,|\boldsymbol{s_n}|=X_n\\ \boldsymbol{s}=\boldsymbol{s_1}+\cdots+\boldsymbol{s_n}}}\prod_{i=1}^{n}\frac{\{\boldsymbol{B}(Z_i)\}^{\boldsymbol{s_i}}}{\boldsymbol{s_i}!}\prod_{k=1}^{j}\frac{\Gamma(a_k+\boldsymbol{s}(k))b_k^{a_k}}{\Gamma(a_k)(b_k+\sum_{t=0}^{n}B_k(Z_t))^{a_k+\boldsymbol{s}(k)}}}. \tag{3.68}$$

3.9 Functional regression model

Spline functions are widely used to model functional data; see Cardot et al. (2003) for example. A rate of convergence result was obtained in Hall and Horowitz (2007). A Bayesian method based on splines is given by Goldsmith et al. (2011). However, to the best of our knowledge, no results on posterior convergence rates for this model are yet available. We consider two types

of functional regression model. The first one assumes only the covariates $Z(t)$ and the effects $\beta(t)$ depend on time t. The second one allows functional observations $X(t)$. We can use any basis with general approximation properties for Hölder classes under the L_2-distance.

3.9.1 Functional covariates

Suppose that we observe i.i.d pairs $(Z_1, X_1), \ldots, (Z_n, X_n)$, where Z_i's are square integrable random functions defined on $(0, 1)$ and X's are scalers. A functional regression model can be formulated as follows:

$$X_i = \int_0^1 Z_i(t)\beta(t)dt + \epsilon_i, \tag{3.69}$$

where $\beta(t)$ is the coefficient function we want to estimate, $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ and $Z_i \overset{\text{iid}}{\sim} G$. The posterior convergence rate theorem is as follows.

Theorem 18. *Suppose that the true regression function $\beta \in \mathcal{C}^\alpha[0, 1]$ and $\mathrm{E}Z^2(t)$ is uniformly bounded away from 0 and ∞ for every $t \in [0, 1]$. If the prior satisfies Conditions (A2) and (A3), relations* (3.52)*,* (3.53) *and* (3.54) *and the basis satisfies Condition (B3), then the posterior converges at the rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$ relative to the L_2-distance.*

Proof. We consider a basis expansion $\beta(t) = \sum_{k=1}^J \theta_k \xi_k(t)$. Denote $W_{ik} = \int_0^1 Z_i(t)\xi_k(t)dt$, then the model can be written as

$$X_i = \sum_{k=1}^J \theta_k W_{ik} + \epsilon_i. \tag{3.70}$$

Define $P_\beta(\cdot|Z)$ as the normal measure with mean $\int_0^1 \beta(t)dt$ and variance σ^2. This allows us to

bound the KL divergences between P_{β_0} and P_β using Cauchy-Schwarz inequality:

$$\begin{aligned}\max\left\{K(P_{\beta_0},P_\beta),V(P_{\beta_0},P_\beta)\right\} &\lesssim \frac{1}{\sigma^2}\mathrm{E}_Z\Big(\int_0^1 Z(t)\{\beta(t)-\beta_0(t)\}dt\Big)^2\\ &\leq \frac{1}{\sigma^2}\|\beta-\beta_0\|_2^2\sup_t \mathrm{E}Z^2(t)\\ &\lesssim \frac{1}{\sigma^2}\|\beta-\beta_0\|_2^2. \end{aligned} \tag{3.71}$$

Hence we can apply Theorem 11 with $\boldsymbol{\eta}$ as σ, $\mathcal{F}_n = ((\log n)^{-1/t_4}, \exp\{an\bar{\epsilon}_n^2/t_5\})$ for sufficiently large constant $a > 0$, $\bar{J}_n = (n/\log n)^{1/(2\alpha+1)}$, $\bar{\epsilon}_n = (n/\log n)^{-\alpha/(2\alpha+1)}$, $J_n = n^{1/(2\alpha+1)}(\log n)^{2\alpha/(2\alpha+1)-t_1}$, d as the L_2-distance and $M_n = n^{1/t_3}$. Then the posterior converges at the rate ϵ_n relative to the h_n. Note that $\|\cdot\|_2$ satisfies conditions in Lemma 2 of Ghosal and van der Vaart (2007a). Thus h_n can be replaced by $\|\cdot\|_2$ and the posterior convergence results hold for $\|\cdot\|_2$ as well. □

3.9.2 Functional responses

Next, we consider a longitudinal type of functional model:

$$X_i(T_i) = Z_i(T_i)\beta(T_i) + \epsilon(T_i). \tag{3.72}$$

For each object i, we observe its response X_i at a random time $T_i \in (0,1)$ with some random covariate Z_i. We assume $(Z_i, T_i) \overset{\text{iid}}{\sim} (G, T)$ and $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0,\sigma^2)$, where σ^2 is unknown.

Theorem 19. *Suppose that the true regression function $\beta \in \mathcal{C}^\alpha[0,1]$ and $\mathrm{E}Z^2(t)$ and the observation time T are uniformly bounded away from 0 and ∞ for every $t \in (0,1)$. If the prior satisfies Conditions (A2) and (A3), relations* (3.52) (3.53) *and* (3.54) *and the basis satisfies Condition (B3), then the posterior converges at the rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$*

relative to the L_2-distance.

Proof. We have

$$\begin{aligned}\max\left\{K(P_{\beta_0},P_\beta),V(P_{\beta_0},P_\beta)\right\} &\lesssim \frac{1}{\sigma^2}\mathrm{E}\int_0^1 Z^2(t)(\beta(t)-\beta_0(t))^2 g(t)dt \\ &\leq \frac{1}{\sigma^2}\|\beta-\beta_0\|_2^2 \sup_t \mathrm{E}Z^2(t) \\ &\lesssim \frac{1}{\sigma^2}\|\beta-\beta_0\|_2^2 \end{aligned} \tag{3.73}$$

Using the same arguments in Theorem 18, we have the convergence theorem. This rate coincides with the optimal rate obtained in Cai and Yuan (2011) within a logarithm factor. □

3.9.3 MCMC-free computation

Consider a basis expansion $\beta(t)=\sum_{k=1}^J \theta_k\xi_k(t)$. Denote $D_{ij}=\int_0^1 Z_i(t)\xi_j(t)dt$ and $\boldsymbol{D}_J$ as an n by J matrix whose (i,j)-entry is D_{ij}, then the model can be written as $\boldsymbol{X}=\boldsymbol{D}_J\boldsymbol{\theta}+\boldsymbol{\epsilon}$. We use the Zellner's g-prior (Zellner, 1986) $\boldsymbol{\theta}|\sigma^2,\boldsymbol{D}_J,J\sim \mathrm{N}(\mathbf{0},g\sigma^2(\boldsymbol{D}_J^T\boldsymbol{D}_J)^{-1})$ for some $g>0$. By a calculation similar to Subsection 3.6.2, we obtain the posterior mean of β_0 at point Z_0 as

$$\frac{g\sum_{j=1}^\infty \Pi(j)\boldsymbol{B}_\mathbf{0}^T(\boldsymbol{D}_j^T\boldsymbol{D}_j)^{-1}\boldsymbol{D}_j^T\boldsymbol{X}\int_{\sigma>0}\Pi(\sigma)\sigma^{-n}\exp\left\{-\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P}_j)\boldsymbol{X}/2\sigma^2\right\}d\sigma}{(g+1)\sum_{j=1}^\infty \Pi(j)\int_{\sigma>0}\Pi(\sigma)\sigma^{-n}\exp\left\{-\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P}_j)\boldsymbol{X}/2\sigma^2\right\}d\sigma}, \tag{3.74}$$

where $\boldsymbol{B}_\mathbf{0}=\boldsymbol{B}(Z_0)$ and $\boldsymbol{P}_j=\frac{g}{g+1}\boldsymbol{D}_j(\boldsymbol{D}_j^T\boldsymbol{D}_j)^{-1}\boldsymbol{D}_j^T$.

3.10 Extension to other function spaces

So far, our discussions are restricted to functions that belong to Hölder classes. Recently, there is a lot of interest in studying more general space of functions such as Sobolev and Besov

spaces. We first give their definitions, then present the approximation results of B-splines in these function spaces.

Definition 5. *The Sobolev space $W_p^\alpha(a,b)$ on a bounded interval (a,b) for $1 \le p \le \infty$ and a positive integer α is defined by*

$$W_p^\alpha(a,b) = \{f : f^{(j)} \in L_p(a,b), \;\; j = 1, \ldots, \alpha\} \tag{3.75}$$

with an associated norm $\|f\|_{W_p^\alpha(a,b)} = \sum_{j=0}^{\alpha} \|f^{(j)}\|_p$.

Definition 6. *The Besov space $B_{p,q}^\alpha(a,b)$ on (a,b) for $1 \le p, q \le \infty$ and $\alpha > 0$ is the collection of all functions f such that*

$$|f|_{B_{p,q}^\alpha} = \Big[\int_0^\infty \{t^{-\alpha} \omega_r(f,t)_p\}^q \frac{dt}{t} \Big]^{1/q} < \infty, \tag{3.76}$$

where $r > \alpha$ is an integer and $\omega_r(f,t) = \sup_{|h| \le t} \|\Delta_h^r(f,\cdot)\|_p$ is called the modulus of smoothness of order r.

If $q = \infty$, then $|f|_{B_{p,q}^\alpha}$ is defined as $\sup_{t>0} t^{-\alpha} \omega_r(f,t)_p$ instead. Further define an associated norm on $B_{p,q}^\alpha(a,b)$ by

$$\|f\|_{B_{p,q}^\alpha} = \|f\|_p + |f|_{B_{p,q}^\alpha}. \tag{3.77}$$

The following two lemmas are taken from Chapter 6 of Schumaker (2007). They describe approximation abilities of B-splines in these function spaces, which directly determine the posterior convergence rates of the corresponding Bayesian models.

Lemma 8. *Suppose $1 \le p \le q \le \infty$, for any $0 \le r \le \alpha - 1$ and $f \in W_p^\alpha(a,b)$, there exists a*

constant C and $\boldsymbol{\theta} \in \mathbb{R}^J$ such that

$$\|f - \boldsymbol{\theta}^T \boldsymbol{B}\|_{W_q^\tau(a,b)} \leq C J^{r-\alpha-1/q+1/p} \|f\|_{W_p^\alpha(a,b)}.$$

In particular, if we choose $r = 0$ and $p = q$, the above inequality simplifies into

$$\|f - \boldsymbol{\theta}^T \boldsymbol{B}\|_p \leq C J^{-\alpha} \|f\|_p.$$

Lemma 9. *For any function $f \in B_{p,q}^\alpha(a,b)$, $1 \leq p \leq p' \leq \infty$ and $1 \leq q, q' \leq \infty$, $0 < \tau < \alpha$, there exists a constant C and $\boldsymbol{\theta} \in \mathbb{R}^J$ such that*

$$\|f - \boldsymbol{\theta}^T \boldsymbol{B}\|_{B_{p',q'}^\tau(a,b)} \leq C J^{\tau-\alpha-1/p'+1/p} \|f\|_{B_{p,q}^\alpha(a,b)}.$$

If we can bound the KL-divergence by an appropriate norm in the new function spaces, then the posterior rates calculation will be very similar to what we have done in the previous discussion. For example, to extend Theorem 14, we take $p = q = 2, r = 0$ in Lemma 8, use a B-spline basis and an exponential link function and bound the KL-divergence for $f_0 \in W_2^\alpha(a,b)$ by

$$K(f_0, f_{\boldsymbol{\theta}}) \lesssim \|\log f_0 - \boldsymbol{\theta}^T \boldsymbol{B}\|_2, \; V(f_0, f_{\boldsymbol{\theta}}) \lesssim \|\log f_0 - \boldsymbol{\theta}^T \boldsymbol{B}\|_2^2 \tag{3.78}$$

where $f_{\boldsymbol{\theta}} = \exp\{\boldsymbol{\theta}^T \boldsymbol{B}\}$ is the approximating function. Also, for any $\boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \mathbb{R}^J$, we have $h^2(f_{\boldsymbol{\theta_1}}, f_{\boldsymbol{\theta_2}}) \leq K(f_{\boldsymbol{\theta_1}}, f_{\boldsymbol{\theta_2}})$. Then (1.9), (1.10) together imply $\bar{J}_n \log n \asymp n\bar{\epsilon}_n^2$ and (1.11) gives $\bar{J}_n^{-\alpha} = \bar{\epsilon}_n^2$. Hence we choose $\bar{J}_n = (n/\log n)^{1/(\alpha+1)}$, d_2 as the L_2-distance, $\bar{\epsilon}_n = (n/\log n)^{-\alpha/(2\alpha+2)}$, the rate is $\epsilon_n = n^{-\alpha/(2\alpha+2)}$ multiplied by some logarithm factor.

For Sobolev spaces, the restriction is that Lemma 8 only considers the case when the

smoothness parameter α is an integer. It is not clear whether such approximation result remains valid for non-integer values of α. For Besov spaces, notice that there is an extra power J^τ in Lemma 9, where τ is strictly greater than 0. This suggests a rate $\epsilon_n = n^{-\alpha/(2\alpha+2)+\tau}(\log n)^c$, where $\tau, c > 0$, can be achieved by using the B-spline basis and an exponential link function; here τ can be chosen arbitrarily close to 0. This is because we only need to bound the approximation error under the L_2-distance, and hence extra approximation results on derivatives are not necessary.

3.11 Numerical results

3.11.1 Density estimation

We illustrate the use of conjugate prior structure as described in Section 3.4.3 on density estimation problems. Following Lenk (1991), we consider two types of the true density: Beta densities, and a mixture density of exponential and a normal distribution:

$$f_0(x) \propto \frac{3}{4} 3e^{-3x} + \frac{1}{4}\frac{\sqrt{32}}{\sqrt{\pi}} e^{-32(x-0.75)^2}. \tag{3.79}$$

For each density, we generate $n = 20, 50, 100$ and 300 samples and then implement the random series prior for $q = 1$ and $q = 3$. When $q = 1$, the exact value of the posterior mean can be calculated. When $q = 3$, instead of evaluating all possible terms to get (3.45), we randomly sample $N = 3000$ of them and take the associated average values. We choose a geometric prior for J restricted between 5 and 25. The lower truncation ensures a minimum number of terms in the series expansion while an upper truncation is necessary to carry out the actual computation using a computer. For $\boldsymbol{\theta}$, we use a Dirichlet distribution with parameters $a_1 = \cdots = a_J = 1$. We obtain density estimates at 100 grid points in the unit interval.

We compare our results with that using the Gaussian process (GP) prior in Tokdar (2007), Dirichlet mixture (DM) of normal kernels (Escobar and West, 1995), mixture of Polya Trees (PT) prior (Lavine, 1992, 1994) and frequentist kernel and spline estimates. Mean absolute errors (l_1), mean squared errors (l_2) and computing time t (in seconds) are summarized in Table 3.1. Standard errors (s.e.) are calculated based on 100 Monte-Carlo replications. The results show that random series prior has a comparable performance in accuracy and computation time with Gaussian process prior. During the computation, we also found that RJMCMC will take more than 30 minutes for running a single chain with 5000 iterations. Hence our method is faster. All programs are implemented on six DELL Dual Processor Xeon Twelve Core 3.6 GHz machines with 80GB RAM running 64Bit CentOS Linux 5.0.

3.11.2 Binary regression

For binary regression problems, we apply MCMC-free calculation described in Section 3.7.2. We consider the following three examples of the true classification function.

$$P_1(X=1|Z) = Z^2,\ 0.1 \le Z \le 0.9 \tag{3.80}$$

$$P_2(X=1|Z) = 3.8(Z-0.5)^2,\ 0.1 \le Z \le 0.9 \tag{3.81}$$

$$P_3(X=1|Z) = 0.5\sin(3\pi Z)+0.5,\ 0.1 \le Z \le 0.9. \tag{3.82}$$

In the prior, we use independent uniform distributions on the coefficients. For each example, we generate $n = 100$, 400 and 1000 samples, randomly select 75% of them as the training data and treat the rest of the them as the test data. We implement the MCMC-free calculation (rsp) for $q = 1$ and $q = 3$ and compare the misclassification rate (mis.) and the L_1-distance error on the test data with linear discriminant analysis (LDA), logistic regression (logistic) and linear

Table 3.1: Density estimation examples

Mixture density (3.79)												
	$n = 20$			$n = 50$			$n = 100$			$n = 300$		
	l_2	l_1	t	l_2	l_1	t	l_2	l_1	t	l_2	l_1	t
rsp($q = 1$)	.27	.40	.44	.20	.33	.58	.18	.30	.69	.17	.29	1.24
rsp($q = 3$)	.16	.31	255	.11	.25	317	.10	.24	320	.09	.22	425
GP	.11	.23	53.9	.06	.17	58.8	.04	.14	61.5	.02	.10	66.9
DM	.46	.59	9.6	.28	.44	22.0	.17	.34	33.6	.11	.28	99.5
PT	.18	.33	14.0	.10	.24	17.6	.10	.22	23.8	.07	.19	51.0
Kernel	.13	.26	.00	.05	.17	.00	.04	.14	.00	.02	.09	.00
Spline	.15	.30	.45	.06	.19	.71	.06	.19	.93	.03	.12	1.91
max s.e.	.02	.01	-	.02	.01	-	.01	.01	-	.00	.01	-
Beta$(.5, .5)$ density												
rsp($q = 1$)	.35	.45	.45	.31	.42	.57	.27	.39	.67	.25	.37	1.23
rsp($q = 3$)	.16	.27	267	.15	.27	314	.14	.25	324	.11	.22	428
GP	.34	.39	55.7	.27	.34	61.6	.24	.31	60.9	.19	.26	74.9
DM	.31	.38	7.93	.32	.36	18.5	.27	.29	48.2	.25	.29	116
PT	.61	.74	14.8	.61	.74	17.6	.60	.73	23.9	.60	.73	49.7
Kernel	.37	.33	.00	.32	.29	.00	.28	.25	.00	.24	.22	.00
Spline	.11	.22	.42	.13	.20	.63	.13	.18	1.03	.14	.19	2.44
max s.e.	.01	.01	-	.01	.01	-	.01	.01	-	.00	.00	-
Beta$(2, 4)$ density												
rsp($q = 1$)	.27	.40	.44	.20	.33	.58	.18	.30	.69	.17	.29	1.24
rsp($q = 3$)	.16	.31	255	.11	.25	317	.10	.24	320	.09	.22	425
GP	.11	.23	53.9	.06	.17	58.8	.04	.14	61.5	.02	.10	66.9
DM	.46	.59	9.6	.28	.44	22.0	.17	.34	33.6	.11	.28	99.5
PT	.18	.33	14.0	.10	.24	17.6	.10	.22	23.8	.07	.19	51.0
Kernel	.13	.26	.00	.05	.17	.00	.04	.14	.00	.02	.09	.00
Spline	.15	.30	.45	.06	.19	.71	.06	.19	.93	.03	.12	1.91
max s.e.	.02	.01	-	.02	.01	-	.01	.01	-	.00	.01	-

Table 3.2: Binary regression examples (3.80)–(3.82).

Example (3.80)									
	rsp($q=1$)		rsp($q=3$)		LDA		logistic		SVM
	mis.	l_1	mis.	l_1	mis.	l_1	mis.	l_1	mis.
$n=100$	.408	.240	.394	.234	.240	.053	.238	.051	.274
$n=400$	.300	.208	.300	.208	.215	.032	.211	.030	.252
$n=1000$	.312	.207	.312	.207	.247	.023	.246	.020	.254
max s.e.	.026	.012	.025	.012	.015	.005	.016	.005	.018
Example (3.81)									
$n=100$	.236	.171	.220	.158	.222	.162	.222	.162	.430
$n=400$	.206	.155	.206	.157	.206	.156	.206	.156	.355
$n=1000$	.205	.155	.205	.159	.205	.157	.205	.157	.256
max s.e.	.022	.010	.021	.008	.021	.005	.021	.005	.035
Example (3.82)									
$n=100$	.470	.335	.438	.328	.472	.334	.474	.334	.418
$n=400$	.464	.347	.451	.334	.445	.333	.445	.333	.431
$n=1000$	.435	.341	.430	.339	.423	.335	.423	.335	.425
max s.e.	.034	.010	.037	.010	.036	.008	.037	.008	.032

support vector machine (SVM). Results are summarized in Table 3.2 based on 20 replications. Note that our MCMC-free calculation has a comparable performance with other methods when the true classification function is not monotonic (examples (3.81) and (3.82)).

3.12 Conclusion

In this chapter, we established posterior convergence rates and MCMC-free computation algorithms for a few univariate function estimation problems. Relevant results are summarized in Table 3.3 for Gaussian white noise problem (Gauss. sign.), density estimation (den.), spectral density estimation (spec. den.), nonparametric regression with Gaussian errors (Gauss. reg.), binary regression (bin. reg.), Poisson regression (Poi. reg.) and functional regression (fun. reg.). For each estimation problem, we specify the smoothness class (Smooth.), the posterior convergence rate up to a logarithmic factor (Rate), relevant distance, type of covariates (fixed or random) for regression problems, and transformation function/prior distribution on the coefficients for MCMC-free calculation.

Table 3.3: Summary of convergence rates and MCMC-free computation

Function f	Smooth.	Rate	Metric	Cov.	MCMC-free	Note
Gauss. sign.	Sobolev α	$n^{-\alpha/(2\alpha+1)}$	ℓ_2	-	iden./gauss.	-
Den.	Hölder α	$n^{-\alpha/(2\alpha+1)}$	d_H	-	iden./Dir.	$f \in (0,1)$
	Sobolev α	$n^{-\alpha/(2\alpha+2)}$	d_H	-	-	α integer
	Besov α	$n^{-\alpha/(2\alpha+2)+\tau}$	d_H	-	-	$\tau > 0$
Spec. den.	Hölder α	$n^{-\alpha/(2\alpha+1)}$	dis. L_2	-	inv./gamma	$\alpha \geq 1$
Gauss. reg.	Hölder α	$n^{-\alpha/(2\alpha+1)}$	emp. L_2 L_2	fixed random	iden./gauss.	-
Bin. reg.	Hölder α	$n^{-\alpha/(2\alpha+1)}$	emp. L_2 L_2	fixed random	iden./beta	$f \in (0,1)$
Poi. reg.	Hölder α	$n^{-\alpha/(2\alpha+1)}$	emp. L_2 L_2	fixed random	iden./gamma	$f > 0$
Func. reg.	Hölder α	$n^{-\alpha/(2\alpha+1)}$	L_2	func.	iden./gauss.	-

Chapter 4

Density regression for high-dimensional data

4.1 Introduction

We consider Bayesian estimation of the conditional distribution of the response Y given a large number of predictors $\boldsymbol{X} = (X_1, \ldots, X_p)$, where p is possibly larger than the sample size n. This problem is referred to as a density regression problem and has received attention in many scientific application areas such as genome association studies. Density regression has been widely studied in both frequentist and Bayesian literature. Many non-Bayesian methods focus on the kernel approach (Fan et al., 1996; Hall et al., 1999), which requires estimating the bandwidth using cross-validation (Fan and Yim, 2004), bootstrap (Hall et al., 1999) and other methods. Meanwhile, in the Bayesian literature, there are two common approaches. One is to model the joint density and obtain the conditional density as a by-product. Another is to directly model the conditional density while leaving the marginal distribution of $\boldsymbol{X}$ unspecified. In this paper, we focus on the latter approach in a nonparametric manner. Many of the existing

methods are based on assigning priors on the space of densities through countable mixtures of location-scale densities (Müller et al., 1996) or through generalizing stick-breaking representations (Griffin and Steel, 2006; Dunson et al., 2007; Dunson and Park, 2008; Chung and Dunson, 2009). Also, there are some recent papers on construction of priors based on stochastic processes such as transforming a Gaussian process (Tokdar et al., 2010; Jara and Hanson, 2011), a multivariate generalization of a beta process (Trippa et al., 2011). Ma (2012) proposed a generalized Pólya tree, which possesses nice posterior conjugacy properties that brings fast computation.

In modern data analysis, it is very common for the data to be high dimensional. Statistical analysis in such a setting is possible only under some sparsity assumption and only if a variable selection procedure is required. Many variable selection techniques have been introduced in the frequentist literature, such as discrete subset selection and penalization methods. Popular methods include the least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996) and the sure independence screening (SIS, Fan and Lv, 2008).

Bayesian variable selection methods have also gained popularity. For example, stochastic search variable selection (SSVS) adopts an efficient sampling-based method to avoid comparing all possible sub-models (George and McCulloch, 1993, 1997; Carlin and Chib, 1995). Bayesian model averaging methods incorporate model uncertainty into estimation and predictions (Brown et al., 1998, 2002). For density regression problems, variable selection is commonly accomplished by assigning a Bernoulli distribution prior on covariates (Chung and Dunson, 2009), whereas an efficient SSVS algorithm is implemented to search the model space and combine the posterior estimation results from different models. Tokdar et al. (2010) extended the variable selection work to dimension reduction by allowing the true sets of covariates depend on $\boldsymbol{X}$ only through a projection $\mathbb{P}_{\boldsymbol{S}}\boldsymbol{X}$ of $\boldsymbol{X}$ on a linear subspace $\boldsymbol{S}$. While these proposed methods show promising numerical results, rates of convergence are largely unknown. More-

over, modern applications also routinely require that we treat the dimension p of the regression to be also large, possibly much larger than n. So far such results are completely absent from the Bayesian nonparametric literature.

In the recent years, the literature on Bayesian asymptotics has flourished with many fundamental breakthroughs. General results have been established by Ghosal et al. (2000), Shen and Wasserman (2001), Ghosal and van der Vaart (2007a,b) and van der Vaart and van Zanten (2008) for posterior convergence rates. For density regression models, a consistency result is obtained by Tokdar et al. (2010) for the use of logistic Gaussian process prior, by Norets and Pelenis (2012) for kernel stick-breaking process priors and by Pati et al. (2013) for probit stick-breaking processes. For high dimensional Bayesian models, there are very few convergence rates results available. Parametric models have been studied by Jiang (2007) for generalized linear model and by Castillo and van der Vaart (2012) for Gaussian white noise model. A classification model with categorical predictors is recently considered by Yang and Dunson (2012), who constructed priors using tensor factorizations and obtained a posterior convergence rate allowing p grow exponentially with n.

In this chapter, we consider the Bayesian density regression problem and use a finite linear combination of tensor product B-splines to construct the prior distributions. We show that the posterior convergence rate is adaptive in the sense that it achieves optimality (up to a logarithm factor) for all smoothness levels of the true conditional density function given that the true model only contains a fixed number of covariates. There are two main steps in the proof of Bayesian adaptation results. The first is to construct a sieve that contains the true underlying model while its size is well controlled. The size is usually quantified by the existence of certain tests or in terms of entropy bounds (Ghosal et al., 2000). The other step is to construct an approximation of the true function while its approximation accuracy increases appropriately with the increasing level of smoothness. We choose B-spline function because its approximation

abilities and non-negativity properties. Then the estimation of the function becomes equivalent to the estimation of the coefficients in the B-spline basis expansion (Truong et al., 2005).

Traditional Bayesian variable selection method are very sensitive to the choice of priors on hyper-parameters and a posterior threshold for variable selection. A reversible jump Markov Chain Monte Carlo (RJMCMC) introduced by Green (1995) is commonly applied. For high dimensional data, these methods may be computationally difficult and not stable. We propose a group-wise Dirichlet distribution prior on the coefficients of B-spline functions that leads to conjugacy structure and possible MCMC-free algorithms. Similarly with the univariate density estimation example in Chapter 3, our method presents closed forms for posterior moments including the mean and the variance.

4.2 Bayesian density regression

4.2.1 Prior

Given p covariates $x_1, \ldots, x_p$, we model the conditional density $f(y|x_1, \ldots, x_p)$ by

$$f(y|x_1, \ldots, x_p) = \sum_{j_1=1}^{J_1} \cdots \sum_{j_p=1}^{J_p} \theta_{\boldsymbol{j}}(y) B_{\boldsymbol{j}}(\boldsymbol{x}), \tag{4.1}$$

where $\boldsymbol{j} = (j_1, \ldots, j_p)$ is a p-dimensional index, $\theta_{\boldsymbol{j}}$'s are density functions for every $\boldsymbol{j}$ and $B_{\boldsymbol{j}}$ are tensor-product B-splines of $\boldsymbol{x}$ as described in Appendix A.2. Now expand $\theta_{\boldsymbol{j}}$ by re-scaled B-splines $\theta_{\boldsymbol{j}} = \sum_{j_0=1}^{J_0} \eta_{j_0,\boldsymbol{j}} B^*_{j_0}$, in which $(\eta_{1,\boldsymbol{j}}, \ldots, \eta_{J_0,\boldsymbol{j}})^T$ lies in a J_0-dimensional simplex. Thus the function can be written as

$$f(y, \boldsymbol{x}|\boldsymbol{J}, \boldsymbol{\eta}) = \sum_{j_0=1}^{J_0} \cdots \sum_{j_p=1}^{J_p} \eta_{\boldsymbol{j}} B^*_{j_0}(y) \prod_{k=1}^{p} B_{i_k}(x_k), \tag{4.2}$$

where $\boldsymbol{j} = (j_0, \ldots, j_p)$ is a $(p+1)$-dimensional index. Note that if $\boldsymbol{\eta}_{j_0,\ldots,j_p}$ does not change for $j_k = 1, \ldots, J_k$, then it is equivalent to removing the k-th variable from the model. We define variable inclusion indicators $\gamma_k = \mathbb{1}(\text{the } k\text{-th variable is in the model})$. By assigning prior distributions on indicator variables γ_k, number of terms J_k and the corresponding coefficients $\boldsymbol{\eta}$, we obtain an induced prior on f:

(A1) Prior on the model size: We restrict the total number of variables $r = \sum_{i=1}^{p} \gamma_i$ withnin a pre-specified positive integer $\bar{r}$, which possibly changes with n. Also we assume that the prior Π_r on r satisfies $\Pi_r(x) \geq c_1 \bar{r}^{-t_0}$ $(x = 1, \ldots, \bar{r})$ for some constants $t_0 \geq 1$ and $c_1 > 0$.

(A2) Prior on the inclusion variables Π_V: Given a fixed r, we assume that the probability of each set of variables of size r is lower bounded by $c_2 / \binom{p}{r}$ for some positive constant c_2.

(A3) Prior on the number of expansion terms Π_J: For these directions that satisfy $\gamma_k = 1$, we consider independent prior distributions taking values in $\mathbb{N}$ on J_k such that $\Pi_J(J_k > j) \leq \exp\{-c_3 j \log^{t_1} j\}$ and $\Pi_J(j \leq J_k \leq c_4 j) \geq \exp\{-c_5 j \log^{t_2} j\}$ hold for sufficiently large j, some positive constants c_3, c_4, c_5 and $t_1 \leq t_2 \leq 1$.

(A4) Prior on the coefficients $\Pi_{\boldsymbol{\eta}}$: We assign a J_0-dimensional Dirichlet prior $\mathrm{Dir}(a, \ldots, a)$ on $(\eta_{1,\boldsymbol{j}}, \ldots, \eta_{J_0,\boldsymbol{j}})$ for every $\boldsymbol{j}$ and $a > 0$.

Our variable selection prior (A1) is more general than Jiang (2007) and Yang and Dunson (2012). Condition (A3) is satisfied for the geometric, Poisson and negative binomial distributions on J respectively with $t_1 = t_2 = 0$, $t_1 = t_2 = 1$ and $t_1 = t_2 = 0$.

4.3 Posterior convergence rates

In this section, we establish posterior convergence rate results for our model. We add a subscript n to the total number of covariates p, the expected number of covariates r and the maximum number of covariates $\bar{r}$ to denote they are diverging with respect to the sample size n.

Let Π_n be the prior as defined in (A1)–(A4) and denote the posterior distribution based on n pairs of observations $(Y_1, \boldsymbol{X_1}), \ldots, (Y_n, \boldsymbol{X_n})$ by $\Pi_n(\cdot \mid \boldsymbol{X}^n, \boldsymbol{Y}^n)$. Let $\{\epsilon_n\}_{n\geq 1}$ be a sequence of positive numbers with $\lim_{n\to\infty} \epsilon_n = 0$. Also let ρ denote a suitable metric on the space of probability densities on $\mathbb{R}^d$, such as the L_1-metric or the Hellinger metric. Then for fixed conditional density function f_0 on $[0,1]^{d+1}$, we say the posterior of the density regression model based on Π_n converges around f_0 at a rate ϵ_n in the metric ρ if for any $M < \infty$,

$$\lim_{n\to 0} \Pi_n \left[\{f : \rho(f_0, f) > M\epsilon_n\} | \boldsymbol{X}^n, \boldsymbol{Y}^n\right] = 0 \text{ almost surely.} \tag{4.3}$$

Although (4.3) only establishes ϵ_n as a bound on the convergence rate at f_0, it serves as a useful calibration when checked against the optimal rate for the smoothness class to which f_0 belongs. It is known that for a fixed d, the minimax rate associated with a $(d+1)$-dimensional β-Hölder class is $n^{-\beta/(2\beta+d+1)}$. We establish (4.3) for this class with ϵ_n as $n^{-\beta/(2\beta+d+1)}$ up to a factor a power of $\log n$. A formal result requires some additional conditions on f_0, as summarized below.

(B1) Assume that f_0 is only related to a fixed number of predictors $x_1, \ldots, x_d$. Also $f_0 \in C^{\beta}[0,1]^{d+1}$.

(B2) Assume that $p_n \leq \exp(Cn^{\alpha})$ for $0 < \alpha < 1$. When $\alpha = 0$, define $p_n \leq n^K$ for some constant $K > 0$.

(B3) Assume that the true conditional density f_0 is bounded below by a positive constant m.

Theorem 20. *Suppose we have i.i.d observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from an unknown probability distribution G on $[0,1]^{p_n}$. Assume that the true conditional density satisfies Conditions (B1)–(B3). Given the prior constructed in (A1)–(A4) and the maximal model size $\bar{r}_n = O(\log n)$, then the posterior distribution of f converges at the rate $\epsilon_n = n^{-s_1}(\log n)^{s_2}$ with respect to d_C, where $s_1 = \min\{\beta/(2\beta+d+1), (1-\alpha)/2\}$, $s_2 = \beta/(2\beta+d+1) + 1 - t_1/2$ and $d_C(f, f_0) = \int |f(y|x_1, \ldots, x_p) - f_0(y|x_1, \ldots, x_p)| G(dx_1, \ldots, dx_p)$.*

In order to prove the theorem, we need the following lemma, which describes the approximation ability of tensor product B-splines.

Lemma 10. *(a) For any function $f \in \mathcal{C}^\beta[0,1]^{d_0}$, $0 < \beta \leq q$, there exists $\boldsymbol{\theta} \in \mathbb{R}^{J_0}$ and a constant $C > 0$ such that*

$$\left\| f - \sum_{j_1=1}^{J_0} \cdots \sum_{j_d=1}^{J_0} \theta_{\boldsymbol{j}} \boldsymbol{B}_{\boldsymbol{j}}(\boldsymbol{x}) \right\|_\infty \leq C J_0^{-\beta} \|f^{(\beta)}\|_\infty, \tag{4.4}$$

where $\boldsymbol{j} = (j_1, \ldots, j_d)$.

(b) Further, if $f > 0$, then for sufficiently large J_0, we can choose every element of $\boldsymbol{\theta}$ to be positive.

(c) Assume that $f(y|x_1, \ldots, x_d)$ is a density function in y lying in $\mathcal{C}^\beta[0,1]^{d+1}$. Then for sufficiently large J, there exists $\boldsymbol{\eta} \in (0,1)^J$ satisfying $\sum_{i_0=1}^{J_0} \eta_{i_0,\ldots,i_d} = 1$ for every fixed $(i_1, \ldots, i_d) \in \{1, \ldots, J_0\}^d$, $J = J_0^{d+1}$ and a constant $C_2 > 0$ such that

$$\left\| f(y|x_1, \ldots, x_d) - \sum_{j_0=1}^{J_0} \cdots \sum_{j_d=1}^{J_0} \eta_{\boldsymbol{j}} B^*_{j_0}(y) \prod_{k=1}^{d} B_{j_k}(x_k) \right\|_\infty \leq C_2 J^{-\beta/(d+1)}. \tag{4.5}$$

Proof. Part (a) is a well-known approximation result for tensor product splines, see Lemma 2.1 of de Jonge and van Zanten (2012) for example. Part (b) is a direct multivariate generalization of Lemma 7, Part (b).

For Part (c), note that by Part (b), we have $\boldsymbol{\theta} \in (0,1)^J$ such that

$$\left\| f(y|x_1,\ldots,x_d) - \sum_{j_0=1}^{J_0} \cdots \sum_{j_d=1}^{J_0} \theta_{\boldsymbol{j}} B_{j_0}(y) \prod_{k=1}^{d} B_{j_k}(x_k) \right\|_\infty \leq C_1 J^{-\beta/(d+1)} \tag{4.6}$$

for constant $C_1 = C\|f^{(\beta)}\|_\infty$. Define $\boldsymbol{\xi}$ by $\xi_{j_0,\ldots,j_d} = \theta_{j_0,\ldots,j_d} \int_0^1 B_{j_0}(y)dy$ and $\boldsymbol{B}^{**}$ as the column vector of $B^{**}_{\boldsymbol{j}}(y,\boldsymbol{x}) = B^*_{j_0}(y) \prod_{k=1}^d B_{j_k}(x_k)$, then

$$\left\| f(y|x_1,\ldots,x_d) - \boldsymbol{\xi}^T \boldsymbol{B}^{**} \right\|_\infty \leq C_1 J^{-\beta/(d+1)}. \tag{4.7}$$

By integration, it follows that

$$\left\| 1 - \sum_{j_0=1}^{J_0} \cdots \sum_{j_d=1}^{J_0} \xi_{\boldsymbol{j}} \prod_{k=1}^{d} B_{j_k}(x_k) \right\|_\infty \leq C_1 J^{-\beta/(d+1)}$$

for any $\boldsymbol{x} \in [0,1]^d$. Therefore $\left|1 - \sum_{j_0=1}^{J_0} \xi_{j_0,\ldots,j_d}\right| \leq C_1 J^{-\beta/(d+1)}$ for any $(j_1,\ldots,j_d) \in \{1,\ldots,J_0\}^d$. Define $\boldsymbol{\eta}$ as $\eta_{j_0,\ldots,j_d} = \xi_{j_0,\ldots,j_d} / \sum_{j_0=1}^{J_0} \xi_{j_0,\ldots,j_d}$, then $\|\boldsymbol{\eta} - \boldsymbol{\eta}\|_1 \leq C_1 J^{-\beta/(d+1)}$. Since $\|\boldsymbol{\theta}^T \boldsymbol{B}^{**} - \boldsymbol{\eta}^T \boldsymbol{B}^{**}\|_\infty \leq \|\boldsymbol{\theta} - \boldsymbol{\eta}\|_1 \|\boldsymbol{B}^{**}\|_\infty$ and every B^{**}_i is uniformly bounded above, we have $\|f - \boldsymbol{\eta}^T \boldsymbol{B}^{**}\|_\infty \leq \|f - \boldsymbol{\theta}^T \boldsymbol{B}^{**}\|_\infty + \|\boldsymbol{\theta} - \boldsymbol{\eta}\|_\infty \|\boldsymbol{B}^{**}\|_\infty \leq C_2 J^{-\beta/(d+1)}$ for some positive constant C_2. □

Proof of Theorem 20. Note that the posterior convergence results on the conditional density of y given $\boldsymbol{x}$ is equivalent to the results on the joint density of $(\boldsymbol{x}, y)$ since the dominating measure $(d\boldsymbol{x}, dy)$ is the product of the measures $d\boldsymbol{x}$ and dy.

In order to obtain the posterior convergence rate, we apply Theorem 2. We choose a sieve

$\mathcal{F}_n$ as

$$\left\{\gamma_i \in \{0,1\}^{p_n}, i = 1, \ldots, p_n; \sum_{i=1}^{p_n} \gamma_i \leq \bar{r}_n\right\} \times \left\{\boldsymbol{\eta} \in (-n,n)^{j_0 + \sum_{i=1}^{p_n} j_i \gamma_i}\right\} \times \{1 \leq j_i \leq J_n; i = 0, \ldots, p_n\}, \tag{4.8}$$

and define

$$\begin{aligned} \epsilon_n &= n^{-\min\{\beta/(2\beta+d+1),(1-\alpha)/2\}} (\log n)^{1-t_1/2+\beta/(2\beta+d+1)}, \\ \bar{\epsilon}_n &= n^{-\min\{\beta/(2\beta+d+1),(1-\alpha)/2\}} (\log n)^{\beta/(2\beta+d+1)}, \\ J_n &= \lfloor n^{\max\{(d+1)/(2\beta+d+1),\alpha\}} (\log n)^{(2\beta)/(2\beta+d+1)-t_1} \rfloor + 1. \end{aligned} \tag{4.9}$$

For (1.9), use the fact that $\|\boldsymbol{\theta}_1^T \boldsymbol{B} - \boldsymbol{\theta}_2^T \boldsymbol{B}\|_2 \leq \sqrt{J} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$ for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^J$, we have

$$\begin{aligned} D(\epsilon_n, \mathcal{F}_n, \|\cdot\|_2) &\leq \bar{r}_n \binom{p_n}{\bar{r}_n} J_n D\left(\epsilon_n, \left\{\boldsymbol{\theta} \in (-n,n)^{(\bar{r}_n+1)J_n}\right\}, \|\cdot\|_2\right) \\ &\leq \bar{r}_n p_n^{\bar{r}_n} J_n \left\{\frac{2n\sqrt{J_n}}{\epsilon_n}\right\}^{(\bar{r}_n+1)J_n} \\ &\lesssim \exp\left\{\bar{r}_n \log p_n + (\bar{r}_n+1) J_n \log(2n\sqrt{J_n}/\epsilon_n)\right\}. \end{aligned}$$

Taking the logarithm on both sides, (3.28) holds when $\bar{r}_n \log p_n \leq n\epsilon_n^2$ and $\bar{r}_n J_n \log n \leq n\epsilon_n^2$.

Now, for (1.10), note that since $r \leq \bar{r}_n$, $\Pi(\mathcal{F}_n^c)$ is bounded by

$$\begin{aligned} &\Pi(j_k > J_n, \text{ for at least one } k) + \sum_{j=1}^{J_n} \Pi(\boldsymbol{\eta} \notin (-n,n)^{(\bar{r}_n+1)j}) \\ &\quad \leq \bar{r}_n \exp(-J_n \log^{t_1} J_n) + (\bar{r}_n+1) J_n \exp(-Cn^2) \end{aligned} \tag{4.10}$$

for some constant $C > 0$. Thus we need $J_n \log^{t_1} n \geq n\bar{\epsilon}_n^2$, which is satisfied by the definition in (4.9).

For (1.11), we assume that the true regression depends on $\boldsymbol{x}$ only through $x_1, \ldots, x_d$, otherwise we simply relabel. Define $\bar{J}_n = \lfloor (n/\log n)^{(d+1)/(2\beta+d+1)} \rfloor + 1$. By Lemma 10 and Condition (B1), there exists positive constants $c_0, \ldots, c_d$, $j_i = \lfloor c_i J_n^{1/(d+1)} \rfloor + 1$, $i = 0, \ldots, d$, $\boldsymbol{J_0} = (j_0, \ldots, j_d, 1, \ldots, 1)$ and $\boldsymbol{\eta_0} \in (-n, n)^{\sum_{i=0}^d j_i}$ such that the estimator proposed in (4.2) using $\boldsymbol{J_0}$ and $\boldsymbol{\eta_0}$ satisfies

$$|f_0(y|x_1, \ldots, x_d) - f(y, \boldsymbol{x}|\boldsymbol{J_0}, \boldsymbol{\eta_0})| \leq \bar{J}_n^{-\beta/d} \leq \bar{\epsilon}_n.$$

Also, for these coefficients vector $\boldsymbol{\eta}$ that satisfies $\|\boldsymbol{\eta} - \boldsymbol{\eta_0}\|_2 \leq \bar{\epsilon}_n \bar{J}_n^{-(d+1)/2}$, we have

$$|f(y, \boldsymbol{x}|\boldsymbol{J_0}, \boldsymbol{\eta}) - f(y, \boldsymbol{x}|\boldsymbol{J_0}, \boldsymbol{\eta_0})| \leq \bar{\epsilon}_n$$

and hence $|f_0(y|x_1, \ldots, x_d) - f(y, \boldsymbol{x}|\boldsymbol{J_0}, \boldsymbol{\eta})| \leq 2\bar{\epsilon}_n$. Using Condition (B3), it follows that

$$h^2(f_0, f) = \int \frac{|f_0 - f|^2}{(f_0^{1/2} + f^{1/2})^{1/2}} \leq \frac{1}{m} \|f_0 - f\|_\infty^2 \lesssim \bar{\epsilon}_n^2. \tag{4.11}$$

Note that $\inf f \geq \inf f_0 - \|f_0 - f\|_\infty \geq m/2$ given that $\bar{\epsilon}_n$ is sufficiently small. Using Lemma 8 of Ghosal and van der Vaart (2007b), we have

$$\begin{aligned} K(f_0, f) &\leq 2h^2(f_0, f) \left\| \frac{f_0}{f} \right\|_\infty \lesssim \bar{\epsilon}_n^2, \\ V(f_0, f) &\lesssim h^2(f_0, f) \Big(1 + \left\| \frac{f_0}{f} \right\|_\infty \Big)^2 \lesssim \bar{\epsilon}_n^2. \end{aligned} \tag{4.12}$$

Thus, by (A1), (A2) and the small ball probability estimates of a Dirichlet distribution in

Lemma 6.1 of Ghosal et al. (2000), we obtain

$$
\begin{aligned}
&\Pi(\mathcal{K}(f_0, \bar{\epsilon}_n)) \\
&\quad \geq \quad \Pi_r(r = d)\Pi_V(x_1, \ldots, x_d)\Pi_J(j_k = \lambda_k \bar{J}_n, k = 0, \ldots, d)\Pi(\|\boldsymbol{\eta} - \boldsymbol{\eta_0}\|_2 \leq \bar{\epsilon}_n \bar{J}_n^{-(d+1)/2}) \\
&\quad \gtrsim \quad \exp\left\{-t_0 \log(\log n) - d \log p_n - C_1 d \bar{J}_n \log^{t_2} \bar{J}_n - C_2(d+1)\bar{J}_n \log\left(\bar{J}_n^{(d+1)/2}\right)\right\} \\
&\quad \gtrsim \quad \exp\{-n\bar{\epsilon}_n^2\} \qquad (4.13)
\end{aligned}
$$

for some positive constants $\lambda_0, \ldots, \lambda_d$ and C_1, C_2 since $\log p_n \leq n\bar{\epsilon}_n^2$ and $\bar{J}_n \log n \leq n\bar{\epsilon}_n^2$. $\square$

4.3.1 MCMC-free Computation

First, we assume that we know the true model only contains a fixed subset of covariates $x_1, \ldots, x_d$. Then given the observations $(\boldsymbol{X_i}, Y_i) = (X_{i1}, \ldots, X_{id}; Y_i)$, $i = 1, \ldots, n$, the conditional likelihood $L(\boldsymbol{\eta}, \boldsymbol{J}|\boldsymbol{x}, y)$ is

$$
\begin{aligned}
&\prod_{i=1}^{n} \left\{ \sum_{j_0=1}^{J_0} \cdots \sum_{j_d=1}^{J_d} \eta_{\boldsymbol{j}} B^*_{j_0}(Y_i) \prod_{k=1}^{d} B_{j_k}(X_{i,j_k}) \right\} \\
&\quad = \sum_{\boldsymbol{s_1} \in \{1,\ldots,J_0\} \times \cdots \times \{1,\ldots,J_d\}} \cdots \sum_{\boldsymbol{s_n} \in \{1,\ldots,J_0\} \times \cdots \times \{1,\ldots,J_d\}} \prod_{i=1}^{n} \eta_{\boldsymbol{s_i}} B^{**}_{\boldsymbol{s_i}}(Y_i, \boldsymbol{X_i}),
\end{aligned}
$$

where $B^{**}_{\boldsymbol{s}}(\boldsymbol{x})$ is defined as $B^*_{s_1}(x_1) \prod_{k=2}^{J} B_{s_k}(x_k)$ for every J-dimensional index $\boldsymbol{s}$ and $J \geq 2$. Since $\mathrm{p}(\boldsymbol{J}, \boldsymbol{\eta}|\boldsymbol{x}, y) \propto \mathrm{p}(\boldsymbol{x}, y|\boldsymbol{J}, \boldsymbol{\eta})\Pi(\boldsymbol{\eta}|\boldsymbol{J})\Pi(\boldsymbol{J})$, the posterior mean of $f(y|\boldsymbol{x})$ at point $(\boldsymbol{X_0}, Y_0)$ is given by

$$
\frac{\sum_{j_0=1}^{\infty} \cdots \sum_{j_d=1}^{\infty} \prod_{k=0}^{d} \Pi(j_k) \int_{\boldsymbol{\eta}} f(y|\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{j}) \Pi(\boldsymbol{\eta}) L(\boldsymbol{\theta}, \boldsymbol{j}|\boldsymbol{X}, Y) d\boldsymbol{\eta}}{\sum_{j_0=1}^{\infty} \cdots \sum_{j_d=1}^{\infty} \prod_{k=0}^{d} \Pi(j_k) \int_{\boldsymbol{\eta}} \Pi(\boldsymbol{\eta}) L(\boldsymbol{\theta}, \boldsymbol{j}|\boldsymbol{X}, Y) d\boldsymbol{\eta}}. \qquad (4.14)
$$

In both numerator and denominator, it involves computing an integral $I(\boldsymbol{s_1}, \ldots, \boldsymbol{s_n}) =$

$\int_{\boldsymbol{\eta}} \Pi(\boldsymbol{\eta}) \prod_{k=1}^{n} \eta_{\boldsymbol{s_k}} d\boldsymbol{\eta}$. For a multivariate index $\boldsymbol{i}$, define $\boldsymbol{i}^{(-1)}$ as the index $\boldsymbol{i}$ without the first component. Then observe that the prior on $\boldsymbol{\eta}$ are block-wise independent for $\eta_{\boldsymbol{i}}$ with different $\boldsymbol{i}^{(-1)}$s. Hence this integral can be easily computed. Define $\Lambda_{\boldsymbol{j}} = \{1, \ldots, j_0\} \times \cdots \times \{1, \ldots, j_d\}$, then (4.14) simplifies into

$$\frac{\sum_{j_0=1}^{\infty} \cdots \sum_{j_d=1}^{\infty} \prod_{k=0}^{d} \Pi(j_k) \sum_{\boldsymbol{s_0} \in \Lambda_{\boldsymbol{j}}} \cdots \sum_{\boldsymbol{s_n} \in \Lambda_{\boldsymbol{j}}} I(\boldsymbol{s_0}, \ldots, \boldsymbol{s_n}) \prod_{i=0}^{n} B^{**}_{\boldsymbol{s_i}}(Y_i, \boldsymbol{X_i})}{\sum_{j_0=1}^{\infty} \cdots \sum_{j_d=1}^{\infty} \prod_{k=0}^{d} \Pi(j_k) \sum_{\boldsymbol{s_1} \in \Lambda_{\boldsymbol{j}}} \cdots \sum_{\boldsymbol{s_n} \in \Lambda_{\boldsymbol{j}}} I(\boldsymbol{s_1}, \ldots, \boldsymbol{s_n}) \prod_{i=1}^{n} B^{**}_{\boldsymbol{s_i}}(Y_i, \boldsymbol{X_i})}. \tag{4.15}$$

Now, we take the variable selection into consideration. Suppose that the true model contains d covariates $X_{m_1}, \ldots, X_{m_d}$. Define $m_0 = 0$ and

$$\begin{aligned}
&W^0(m_1, \ldots, m_d) \\
&= \sum_{j_{m_0}=1}^{\infty} \cdots \sum_{j_{m_d}=1}^{\infty} \Pi_J(j_{m_0}) \cdots \Pi_J(j_{m_d}) \sum_{\boldsymbol{s_0} \in \Lambda_{\boldsymbol{j}}} \cdots \sum_{\boldsymbol{s_n} \in \Lambda_{\boldsymbol{j}}} I(\boldsymbol{s_0}, \ldots, \boldsymbol{s_n}) \prod_{i=0}^{n} B^{**}_{\boldsymbol{s_i}}(Y_i, \boldsymbol{X_i}), \\
&W^1(m_1, \ldots, m_d) \\
&= \sum_{j_{m_0}=1}^{\infty} \cdots \sum_{j_{m_d}=1}^{\infty} \Pi_J(j_{m_0}) \cdots \Pi_J(j_{m_d}) \sum_{\boldsymbol{s_1} \in \Lambda_{\boldsymbol{j}}} \cdots \sum_{\boldsymbol{s_n} \in \Lambda_{\boldsymbol{j}}} I(\boldsymbol{s_1}, \ldots, \boldsymbol{s_n}) \prod_{i=1}^{n} B^{**}_{\boldsymbol{s_i}}(Y_i, \boldsymbol{X_i}).
\end{aligned}$$

Then the posterior mean of f at point $(\boldsymbol{x_0}, y_0)$ is

$$\frac{\sum_{d=1}^{\bar{r}} \Pi_r(d) \sum_{1 \le m_1 < \cdots < m_d \le p} \Pi_V(x_{m_1}, \ldots, x_{m_d}) W^0(m_1, \ldots, m_d)}{\sum_{d=1}^{\bar{r}} \Pi_r(d) \sum_{1 \le m_1 < \cdots < m_d \le p} \Pi_V(x_{m_1}, \ldots, x_{m_d}) W^1(m_1, \ldots, m_d)}. \tag{4.16}$$

Similar expressions can be obtained for other posterior moments, in particular, for the posterior variance. This estimate can be viewed as a kernel mixture estimator whose kernel is determined jointly by selected covariates and associated tensor product B-splines. Since a B-spline basis

function takes nonzero values only at q intervals, the calculation of W^0 involves a multiple of $(j_{\max} - j_{\min} + 1)^d q^{d(n+1)}$ terms if we restrict each j_i taking values between $j_{\min}$ and $j_{\max}$. Then there will be a multiple of $\sum_{d=1}^{\bar{r}} \binom{p}{d} (j_{\max} - j_{\min} + 1)^d q^{d(n+1)}$ terms in total. Instead of evaluating all terms, we randomly sample a number of terms in both numerator and denominator and take the associated average values. If we pick $q = 1$, then this number will reduce to $\sum_{d=1}^{\bar{r}} \binom{p}{d} (j_{\max} - j_{\min} + 1)^d$. See more details in Section 4.4.

4.4 Numerical results

4.4.1 Computing algorithm

In the following, we provide more details in the prior construction of our model.

(C0) We choose $q = 1$, in other words, we consider indicator functions as the base.

(C1) We assign a uniform prior on the model size ranging from 2 to $\bar{r}$. Then Condition (A1) is satisfied for $t_0 = 1$.

(C2) The prior probability of γ_i follows a Bernoulli distribution with parameter w_i for $0 \leq w_i \leq 1$ and $i = 1, \ldots, p$. The values of w_i can depend on the marginal correlation between x_i and y.

(C3) We choose Π_J as a geometric distribution with mean 10 and restrict it between 5 and 15.

(C4) Given $J = J_0$, we assign a J_0-dimensional uniform prior on $(\eta_{1,\boldsymbol{j}}, \ldots, \eta_{J_0,\boldsymbol{j}})$ for every $\boldsymbol{j}$. Then Condition (A4) is satisfied for $a = 1$.

4.4.2 Simulation studies

We apply the MCMC-free calculation on the following two examples,

$$Y|\boldsymbol{X} \sim \text{Beta}(4X_1 + 3X_2^2, 10X_2), \tag{4.17}$$

$$Y|\boldsymbol{X} \sim \text{Beta}(5X_2 \exp(2X_1), 5X_3^2 + 3X_4). \tag{4.18}$$

For each example, we generate p covariates $X_1, \ldots, X_p$ uniformly from $[.05, .95]$. We compare our method (rsp) with least-squares kernel conditional density estimation (ls) developed by Sugiyama et al. (2010), where they use L_1-regularization to select variables. Prediction errors under L_1- and L_2- losses and their maximum standard errors associated with 10 Monte-Carlo replications are summarized in Table 4.1. When the true density function has a simple form as in Example (4.17), our result is more accurate when p is larger than n. When the true density becomes more complicated, our method gives uniformly better results than least-squares for all values of p.

Table 4.1: Density regression examples (4.17) and (4.18).

Example (4.17)								
	$n = 100$				$n = 500$			
	rsp (L_1)	rsp (L_2)	ls (L_1)	ls (L_2)	rsp (L_1)	rsp (L_2)	ls (L_1)	ls (L_2)
$p = 5$	.75	.79	.45	.35	.84	.93	.41	.31
$p = 10$	.83	.95	.59	.60	.83	.95	.65	.79
$p = 50$	.72	.71	.69	.74	.73	.77	.67	.77
$p = 100$	.68	.68	.68	.76	.75	.82	.71	.84
$p = 500$	.73	.74	.86	1.24	.76	.79	.85	1.26
$p = 1000$	.73	.73	1.00	1.63	.75	.78	.91	1.51
max s.e.	.05	.09	.05	.11	.05	.09	.10	.20
Example (4.18)								
$p = 5$	.65	.61	.73	.85	.70	.67	.70	.77
$p = 10$	.66	.59	.78	.92	.74	.76	.81	1.14
$p = 50$	.67	.63	.65	.66	.74	.74	.73	.84
$p = 100$	.70	.68	.70	.78	.66	.62	.65	.69
$p = 500$	.58	.50	.69	.80	.77	.83	.78	1.16
$p = 1000$	.74	.73	.75	1.14	.66	.61	.81	1.17
max s.e.	.04	.08	.06	.12	.05	.09	.08	.18

Chapter 5

Adaptive derivative estimation

5.1 Introduction

Estimations of derivatives of density functions have been well studied in the literature and have many applications in estimating score functions, Fisher information, regression curves and hypothesis testing problems etc. (Singh, 1977a). Rosenblatt (1956) introduced a method of kernel type estimation for derivatives of a univariate density. Parzen (1962) and Bartlett (1963) studied the asymptotic behaviors of such estimators and came up with an improved estimator by truncation techniques that reduces estimation bias. The convergence rates results were first obtained by Bhattacharya (1967) and Schuster (1969). However, neither of them achieved optimality (Farrell, 1972). An improved result was shown by Singh (1977b). Besides kernel methods, wavelet is another popular tool in derivative estimation. See Prakasa Rao (1996) and Hosseinioun et al. (2011, 2012) for an example.

To the best of our knowledge, so far there has been no discussion on using Bayesian methods for derivative estimations. In this chapter, we focus on estimating derivatives of a univariate density function using both kernel and spline methods. We establish posterior convergence rate

under L_2-metric. The main challenge lies in the need of working directly with Euclidean distances instead of the most commonly used Hellinger distances and extending the original Le Cam's testing theory under more general situations. We employ the technique developed by Giné and Nickl (2011), in which they obtain posterior convergence rates under L_p-norm by constructing an appropriate test that differentiate the target density of interest and their estimates and then control probabilities of both type I and type II errors using Talagrands inequality (Talagrand, 1996). In our work, we establish a test that differentiate the true derivative $f_0^{(r)}$ and its estimate $f^{(r)}$:

$$H_0 : f^{(r)} = f_0^{(r)} \quad \text{against } H_A : \|f^{(r)} - f_0^{(r)}\|_p \geq \epsilon_n. \tag{5.1}$$

We show that under certain boundary conditions on f_0, the distance on the space of densities and the distance on the space of their derivatives are closely related in both directions. Hence these two types of tests are equivalent except that the value of ϵ_n may be different. Such results allow us to directly apply results in Giné and Nickl (2011) to obtain convergence rates under L_p-norm for general values of p.

5.2 Prior

5.2.1 Dirichlet mixture prior

We consider estimating derivatives of a density function $f \in \mathcal{C}^\alpha[0,1]$. For $\alpha > 1$, we are able to estimate derivatives up to order $r < \lfloor \alpha \rfloor$. We first describe prior distributions constructed based on kernel estimates. Define the r-th order derivative of a standard normal density function by

$$\phi^{(r)}(x) = \frac{1}{\sqrt{2\pi}} \frac{d^r}{dx^r} e^{-x^2/2},$$

which can be written in terms of Hermite polynomials in $\phi(x)$. Similarly with Section 2.2.1, we consider a Dirichlet mixture prior as follows

$$\int_{-\infty}^{\infty} \frac{1}{\sigma^{r+1}} \phi^{(r)}\left(\frac{x-\mu}{\sigma}\right) dF(\mu),\ x \in [0,1],\ \ F \sim \mathrm{DP}(\alpha),\ \ \sigma \sim G, \tag{5.2}$$

where α is some finite positive measure defined on $\mathbb{R}$ and G is a prior distribution on σ satisfying the following conditions for some positive constants $a_1, a_2, a_3, a_4, b_1, b_2, b_3, C_1, C_2, C_3$:

$$\begin{aligned} G(\sigma > x) &\leq b_1 x^{-a_1} \text{ for all sufficiently large } x > 0, & (5.3)\\ G(\sigma \leq x) &\leq b_2 \exp(-C_1 x^{a_2}) \text{ for all sufficiently small } x > 0, & (5.4)\\ G(s < \sigma < (1+C_2)s) &\geq b_3 s^{a_3} C_2^{a_4} \exp(-C_3 s^{1/2}) \text{ for any } s > 0. & (5.5) \end{aligned}$$

Also, define $\bar{\alpha} = \alpha/\alpha(\mathbb{R})$ and assume that for some constant $C_4, a_5 > 0$,

$$1 - \bar{\alpha}([-x,x]) \leq b_4 \exp(-C_4 x^{a_5}) \text{ for all sufficiently large } x > 0. \tag{5.6}$$

Conditions (5.3)–(5.5) on σ do not hold for the commonly used inverse gamma distribution. However, the square of an inverse gamma random variable satisfies these conditions. Condition (5.6) is satisfied for a lot of probability distributions with exponential tails such as Gaussian, exponential etc.

5.2.2 Finite mixture prior

As an alternative to Dirichlet mixture, we can also use a finite mixture prior as follows:

$$\sum_{h=1}^{H} \omega_h \phi_\sigma^{(r)}(x - \mu_h),\ x \in [0,1],\ \ \sum_{h=1}^{H} \omega_h = 1. \tag{5.7}$$

By assigning prior distributions on $H, \boldsymbol{\omega} = (\omega_1, \ldots, \omega_H)$ and $\mu_1, \ldots, \mu_H$, we obtain an induced prior on the $K_{\sigma,r} f$. We assume that the prior on σ satisfies (5.3)–(5.5) and the prior on H satisfies

$$b_5 \exp\{-C_5 x (\log x)^{\tau_1}\} \leq \Pi(H \geq x) \leq b_6 \exp\{-C_6 x (\log x)^{\tau_1}\} \tag{5.8}$$

for sufficiently large $x > 0$ and some positive constants b_5, b_6, C_6, τ_1. Also, for every fixed $H = h$ and some positive constants a_6, b_7, b_8, C_8,

$$\Pi(\mu_i \notin [-x, x]) \leq b_7 \exp(-C_7 x^{a_6}), \text{ for sufficiently large } x > 0, \ \ i = 1, \ldots, h, \tag{5.9}$$

$$\Pi(\|\boldsymbol{\omega} - \boldsymbol{\omega_0}\| \leq \epsilon) \geq b_8 \exp\{-C_8 h \log(1/\epsilon)\}, \text{ for all } 0 < \epsilon < 1/h \text{ and all } \boldsymbol{\omega_0} \in \Delta_h. \tag{5.10}$$

5.2.3 Random series prior

Here we consider random series prior discussed in Chapter 3. In particular, we choose B-spline function as the basis because of its nice approximation ability. For a density f, we consider

$$f_{\boldsymbol{\theta}}(x) = \sum_{j=1}^{J} \theta_j B_j^*(x), \ x \in [0, 1], \tag{5.11}$$

where B_j^*'s are re-scaled B-spline basis functions defined in Section 3.2.3. Then by taking the derivative of B^* and assigning priors on J and coefficients $\theta_1, \ldots, \theta_J$, we obtain an induced prior on $f^{(r)}$. We assume the following conditions hold:

$$\Pi(J > x) \ \leq \ \exp\{-C_9 x \log^{t_1} x\} \text{ for sufficiently large } x > 0, \tag{5.12}$$

$$\Pi(x \leq J < (1 + C_{10})x) \ \geq \ \exp\{-C_{11} x \log^{t_2} x\} \text{ for sufficiently large } x > 0, \tag{5.13}$$

and when the value of J is fixed at j, the following hold:

$$\Pi(\|\boldsymbol{\theta}-\boldsymbol{\theta_0}\|_2 \leq \epsilon) \;\geq\; \exp\{-C_{12} j \log(1/\epsilon)\} \text{ for all } \|\boldsymbol{\theta_0}\|_\infty \leq H, \tag{5.14}$$

$$\Pi(\boldsymbol{\theta} \notin [-M, M]^j) \;\leq\; j \exp\{-C_{13} M^{t_3}\} \text{ for sufficiently large } M > 0 \tag{5.15}$$

for some positive constants $C_9, C_{10}, C_{11}, C_{12}, C_{13}$, t_1, t_2, t_3 and large $H > 0$.

Note that B_j^* is not differentiable at boundaries of its support. Hence we define all relevant derivatives equal to zero at the boundaries.

As discussed in Remarks 3 and 1, Conditions (5.12) and (5.13) hold for Geometric, Poisson and negative binomial distributions. Conditions (5.14) and (5.15) hold for multivariate normal, Dirichlet and independent gamma, exponential distributions.

5.3 Posterior convergence rate

5.3.1 Main results

Theorem 21. *Suppose that we have i.i.d observations $X_1, \ldots, X_n$ from a density function $f_0 \in \mathcal{C}^\alpha[0,1]$. Assume that f_0 satisfies*

$$f_0^{(k)}(0) = c_k, \; k = 0, \ldots, r-1, \tag{5.16}$$

where r is a positive integer with $r < \lfloor \alpha \rfloor$, $\alpha > 1$ and $c_0, \ldots, c_{r-1}$ are some fixed constants. Also, there exists positive numbers W_1 and W_2 such that $W_1 \leq f_0 \leq W_2$. Let the prior be constructed as in Section 5.2, $1 \leq p \leq \infty$ and $\bar{p} = \max(2, p)$, then

$$\Pi\{\|f^{(r)} - f_0^{(r)}\|_p \geq M_n n^{-s_1} (\log n)^{s_2} | X_1, \ldots, X_n\} \to 0 \text{ for any } M_n \to \infty. \tag{5.17}$$

The numbers s_1 and s_2 are determined in the following way:

(i) When using Dirichlet mixture prior in Section 5.2.1,

$$s_1 = \frac{(\alpha - 1/2 + 1/\bar{p})(\alpha - r)}{\alpha(2\alpha + 1)}, \;\; s_2 = \frac{2(\bar{p} - 1)(\alpha - r)}{\bar{p}\alpha};$$

(ii) When using finite mixture prior in Section 5.2.2,

$$s_1 = \frac{(\alpha - 1/2 + 1/\bar{p})(\alpha - r)}{\alpha(2\alpha + 1)}, \;\; s_2 = \{2 + \max(0, 1 - \tau_1)\}\frac{(\bar{p} - 1)(\alpha - r)}{\bar{p}\alpha};$$

(iii) When using random series prior in Section 5.2.3,

$$s_1 = \frac{(\alpha - 1/2 + 1/\bar{p})(\alpha - r)}{\alpha(2\alpha + 1)}, \;\; s_2 = \left(\frac{2\alpha}{2\alpha + 1} + 1 - t_1\right)\frac{(\bar{p} - 1)(\alpha - r)}{\bar{p}\alpha}.$$

When $p \leq 2$, we have $\bar{p} = 2$, hence $s_1 = (\alpha - r)/(2\alpha + 1)$, which leads to the minimax rate for estimating an r-th order derivative of a α-Hölder function. When $p > 2$, there will be a deterioration factor caused by large value of p.

Condition (5.16) is commonly assumed when $r = 1$, i.e., when we are estimating the first order derivative, it is natural to assume that the true density has zero values at the boundary. Then the posterior of Bayesian estimate f' concentrates around the true f_0' at a rate $n^{-(\alpha-1)/(2\alpha+1)}$ multiplied by some logarithmic factor under L_2-metric.

Here, we briefly explain the idea of the proof. We first establish posterior convergence rates for density estimation problem under L_p-metric using Theorem 3 of Giné and Nickl (2011). That theorem states that for two sequences of positive numbers $\epsilon_n, \delta_n \to 0$, one needs to verify the original sets of conditions involving lower bounding the prior probability on the sieve (1.10) and the prior concentration probability (1.11) and replace the entropy condition

(1.9) by obtaining more accurate bounds on the approximation errors in terms of δ_n. More precisely,

$$\|f_n^{(r)} - f_0^{(r)}\|_p = O(\delta_n^{(\alpha-r)/\alpha}), \tag{5.18}$$

where $f_n^{(r)}$ stands for our posterior estimate of r-th order derivative based on n observations. The other conditions in that theorem are easily satisfied because the true function f_0 and all of its derivatives are bounded by our assumption. Therefore, we may choose ϵ_n as the convergence rate in our original theorems with respect to the Hellinger distance, then (1.10) and (1.11) hold for the same sieve used when proving the original theorem. Also, (5.18) is guaranteed by approximation results stated in Lemmas 12 and 13. According to Theorem 3 of Giné and Nickl (2011), define $\delta_n = \epsilon_n(n\epsilon_n^2)^{1/2-1/\bar{p}}$, we obtain $\delta_n^{(\alpha-r)/\alpha}$ as the convergence rate under L_p-metric by Lemma 11.

The following lemma establishes the relationship between the distances on the space of densities and the distance on the space of derivatives.

Lemma 11. *Suppose that $f, g \in \mathcal{C}^\alpha[0,1]$ are two density functions with $\alpha > 1$. Given a sequence of positive number $\epsilon_n \to 0$ as $n \to 0$, every positive integer $r < \lfloor\alpha\rfloor$ and $1 \le p \le \infty$, suppose that $|f^{(k)}(0) - g^{(k)}(0)| \le \epsilon_n$ for $k = 0, 1, \ldots, \lfloor\alpha\rfloor$, we have*

(i) If $\|f^{(r)} - g^{(r)}\|_p \le \epsilon_n$, then $\|f^{(r-1)} - g^{(r-1)}\|_p \lesssim \epsilon_n$;

(ii) If $\|f^{(r-1)} - g^{(r-1)}\|_p \le \epsilon_n$, then $\|f^{(r)} - g^{(r)}\|_p \le C\epsilon_n^{(\alpha-r)/(\alpha-r+1)}$ for some constant $C > 0$.

(iii) If $\|f^{(r)} - g^{(r)}\|_p \le \epsilon_n$, then $\|f - g\|_p \le \epsilon_n$;

(iv) If $\|f - g\|_p \le \epsilon_n$, then $\|f^{(r)} - g^{(r)}\|_p \le C_0\epsilon_n^{(\alpha-r)/\alpha}$ for some constant $C_0 > 0$.

Proof. For Part (i), by integration,

$$\begin{aligned} \|f^{(r-1)} - g^{(r-1)}\|_p &= \left\| f^{(r-1)}(0) + \int_0^x f^{(r)}(\xi)d\xi - g^{(r-1)}(0) - \int_0^x g^{(r)}(\xi)d\xi \right\|_p \\ &\leq \epsilon_n + \int_0^1 \|f^{(r)}(\xi) - g^{(r)}(\xi)\|_p d\xi \\ &\lesssim \epsilon_n. \end{aligned} \tag{5.19}$$

For Part (ii), note that for any $x \in (0,1)$, there exists $\xi \geq x$ such that

$$f^{(r)}(x) - g^{(r)}(x) = \frac{f^{(r-1)}(\xi) - g^{(r-1)}(\xi)}{\xi}. \tag{5.20}$$

Let $\delta_n \to 0$ be a sequence of positive numbers. If $\xi \leq \delta_n$, then by the fact that both $f^{(r-1)}$ and $g^{(r-1)}$ belong to $\mathcal{C}^{\alpha-r+1}$ and using the local smoothness condition, we obtain that both functions are bounded by a constant multiple of $\xi^{\alpha-r+1}$. If $\xi > \delta_n$, then the L_p-norm of (5.20) can be bounded by ϵ_n/ξ_n. Therefore

$$\begin{aligned} \|f^{(r)}(x) - g^{(r)}\|_p &\lesssim \frac{\|f^{(r-1)} - g^{(r-1)}\|_p}{\delta_n} + \delta_n^{\alpha-r} \\ &\lesssim \frac{\epsilon_n}{\delta_n} + \delta_n^{\alpha-r}. \end{aligned}$$

Choose $\delta_n = \epsilon_n^{1/(\alpha-r+1)}$, then the result follows.

Parts (iii) and (iv) are immediate consequences of (i) and (ii). □

The following proposition is an analog of Theorems 2 and 3 of Giné and Nickl (2011). It forms the basis of posterior convergence rate calculations for derivative estimation.

Proposition 3. *Suppose we have i.i.d. observations $X_1, \ldots, X_n$ following the law of some true density function $f_0 \in \mathcal{C}^\alpha[0,1]$. Let $1 \leq p \leq \infty$, r be a positive integer strictly smaller than*

$\lfloor\alpha\rfloor$ and $\epsilon_n \to 0$ be a sequence of positive numbers such that $n\epsilon_n^2 \to \infty$. Let $\mathcal{F}$ be a set of probability density functions f defined on $\mathcal{C}^\alpha[0,1]$ satisfying

$$|f^{(k)}(0) - f_0^{(k)}(0)| \leq \epsilon_n, \text{ for every nonnegative integer } k < \lfloor\alpha\rfloor \tag{5.21}$$

and Π_n be priors defined on some σ-algebra of $\mathcal{F}$. Also, there exist positive numbers W_1 and W_2 such that $W_1 \leq f_0 \leq W_2$. Define

$$\delta_n = \epsilon_n(n\epsilon_n^2)^{1/2-1/\bar{p}}, \ \xi_n = \delta_n^{(\alpha-r)/\alpha}, \ \ \bar{p} = \max(2,p). \tag{5.22}$$

Let $\mathcal{F}_n$ be a sequence of subsets of

$$\{f \in \mathcal{F} : \|T_{n,r}(f) - f^{(r)}\|_p \leq C_n\xi_n\}, \tag{5.23}$$

where C_n is a sequence of positive constants and $T_{n,r}(f)$ is an admissible approximation sequence of $f^{(r)}$. Assume there exists $C > 0$ such that for sufficiently large n,

$$\Pi_n(\mathcal{F}_n^c) \leq e^{-(C+4)n\epsilon_n^2}, \tag{5.24}$$

$$\Pi_n\{\mathcal{K}(f_0,\epsilon_n)\} \geq e^{-Cn\epsilon_n^2}. \tag{5.25}$$

If

$$\|T_{n,r}(f_0) - f_0^{(r)}\|_p = O(\xi_n) \tag{5.26}$$

and $\delta_n \to 0$ *as* $n \to 0$*, then for every* $M_n \to \infty$*,*

$$\Pi_n\{f \in \mathcal{F} : \|f^{(r)} - f_0^{(r)}\|_p \geq M_n \xi_n | X_1, \ldots, X_n\} \to 0, \quad \textit{in probability}. \tag{5.27}$$

Proof. The proof is analogous to proofs of Theorems 2 and 3 of Giné and Nickl (2011). In their proofs, they show that there exist test functions $\phi_n = \phi_n(X_1, \ldots, X_n; f_0)$ such that for some constants $L_1, L_2 > 0$,

$$P_0^n \phi_n \leq e^{-L_1 n \epsilon_n^2}, \tag{5.28}$$

$$\sup_{f \in \mathcal{F}_n : \|f - f_0\|_p \geq M\delta_n} P^n(1 - \phi_n) \leq e^{-L_2 n \epsilon_n^2}, \tag{5.29}$$

where P_0^n and P^n are probability measures of f_0 and f. Using Lemma 11, we obtain

$$P_0^n \phi_n \leq e^{-L_1 n \epsilon_n^2}, \tag{5.30}$$

$$\sup_{f \in \mathcal{F}_n : \|f^{(r)} - f_0^{(r)}\|_p \geq M\xi_n} P^n(1 - \phi_n) \leq e^{-L_2 n \epsilon_n^2} \tag{5.31}$$

for a different set of constants. Then the conclusion follows, see arguments in the proof of Lemma 2.1 of Ghosal et al. (2000). □

5.3.2 Approximation results

We first discuss the approximation ability of kernel estimators.

Lemma 12. *For* $\alpha > 1, \sigma > 0$ *and a density function* $f \in \mathcal{C}^\alpha[0, 1]$*, define*

$$K_{\sigma,r} f(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma^{r+1}} \phi^{(r)}\left(\frac{x - \mu}{\sigma}\right) f(z) dz.$$

Given σ sufficiently small and $f(x) \geq M$ for some constant $M > 0$, for any nonnegative integer $r < \lfloor \alpha \rfloor$, there exists a probability density function g such that

$$\|K_{\sigma,r} g - f^{(r)}\|_p \lesssim \sigma^{\alpha - r} \tag{5.32}$$

for $1 \leq p \leq \infty$.

Proof. We only need to prove the result for L_∞-metric. We start from the approximation construction as in de Jonge and van Zanten (2010) and Rousseau (2010). Define $m_j = \int_{-\infty}^{\infty} y^j \phi(y) dy$ as the moment of normal kernel. For every $n \in \mathbb{N}$, we define two sequences of numbers as follows:

$$c_n = \sum_{n=l+k, l \geq 1, k \geq 1} \frac{(-1)^{k+1}}{k!} m_k d_l, \;\; d_n = \frac{(-1)^n m_n}{n!} + c_n. \tag{5.33}$$

Define α_0 as the largest integer strictly smaller than α and

$$T = f - \sum_{j=1}^{\alpha_0 - r} d_j \sigma^j f^{(j)}. \tag{5.34}$$

Using Lemma 3.4 in de Jonge and van Zanten (2010), we obtain

$$\|K_\sigma T^{(r)} - f^{(r)}\|_\infty \lesssim \sigma^{\alpha - r}. \tag{5.35}$$

Using the similar arguments as in Bhattacharya (1967), we have the following:

$$\begin{aligned} & \frac{1}{\sigma^{r+1}} \phi^{(r)}\big(\frac{x-\mu}{\sigma}\big) T(\mu) d\mu \\ = \;\; & \frac{1}{\sigma} \phi\big(\frac{x-\mu}{\sigma}\big) T^{(r)}(\mu) - \frac{d}{d\mu} \sum_{s=0}^{r-1} \phi^{(s)}\big(\frac{x-\mu}{\sigma}\big) T^{(r-s-1)}(\mu). \end{aligned} \tag{5.36}$$

Upon integration, we obtain

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{1}{\sigma^{r+1}} \phi^{(r)}\left(\frac{x-\mu}{\sigma}\right) T(\mu) d\mu \\
&= \frac{1}{\sigma} \int_{-\infty}^{\infty} \phi\left(\frac{x-\mu}{\sigma}\right) T^{(r)}(\mu) d\mu - \lim_{\mu\to\infty} \sum_{s=0}^{r-1} \phi^{(s)}\left(\frac{x-\mu}{\sigma}\right) T^{(r-s-1)}(\mu) \\
&\quad + \lim_{\mu\to-\infty} \sum_{s=0}^{r-1} \phi^{(s)}\left(\frac{x-\mu}{\sigma}\right) T^{(r-s-1)}(\mu) \\
&= \frac{1}{\sigma} \int_{-\infty}^{\infty} \phi\left(\frac{x-\mu}{\sigma}\right) T^{(r)}(\mu) d\mu \qquad (5.37)
\end{aligned}$$

Combing this result with (5.35), we have $\|K_{\sigma,r}T - f^{(r)}\|_\infty \lesssim \sigma^{\alpha-r}$.

Observe that the absolute values of d_j and $f^{(j)}$ are bounded. Given σ sufficiently small, we have $T(x) > f(x)/2$. Define a re-scaled version of T as $g(x) = T(x)/\int_0^1 T(u)du$. Taking $r = 0$, we have $\|K_\sigma T - f\|_\infty \lesssim \sigma^\alpha$. Hence $|1 - \int_0^1 T(u)du| \lesssim \sigma^\alpha$. As a result, $\|g - T_0\|_\infty \lesssim \sigma^\alpha$. The conclusion then follows. □

For B-spline models, we have the following lemma.

Lemma 13. *For $\alpha > 1$ and a density function $f_0 \in \mathcal{C}^\alpha[0,1]$, given the smoothness level of the B-spline basis function $q > \alpha$, for sufficiently large J, there exists a $\boldsymbol{\theta_0} \in \Delta_J$ such that*

$$\left\|\boldsymbol{\theta}_\mathbf{0}^T \boldsymbol{B}^{*(r)}(x)\} - f^{(r)}\right\|_p \lesssim J^{r-\alpha} \qquad (5.38)$$

for any nonnegative integer r satisfying $0 \le r < \lfloor\alpha\rfloor$ and $1 \le p \le \infty$.

Proof. It is well known that (see Chapter 6, Schumaker (2007) for example) their exists a $\boldsymbol{\theta_1} \in (0,\infty)^J$ such that

$$\left\|\boldsymbol{\theta}_\mathbf{1}^T \boldsymbol{B}^{(r)}(x)\} - f^{(r)}\right\|_\infty \lesssim J^{r-\alpha}$$

Using the same arguments in Part (d) of Lemma 7, we can find a $\boldsymbol{\theta_0} \in \Delta_J$ that enjoys the same approximation order. □

5.3.3 Proof of Theorem 21

Proof. We apply proposition 3 with ϵ_n as the posterior convergence rate obtained for density estimation problem under the Hellinger distance. Note that (5.26) holds because of the approximation results established in Lemmas 12 and 13.

For Dirichlet mixture prior, according to Theorem 3, we take $d^* = 1$ and remove the tail parameter τ on f_0, then the rate there is $\epsilon_n = n^{-\alpha/(2\alpha+1)} \log n$. Hence the rate is

$$\left\{\epsilon_n (n\epsilon_n^2)^{1/2-1/\bar{p}}\right\}^{(\alpha-r)/\alpha} = n^{-\frac{(\alpha-r)(\alpha-1/2+1/\bar{p})}{\alpha(2\alpha+1)}} (\log n)^{2(\alpha-r)(\bar{p}-1)/(\bar{p}\alpha)}$$

For finite mixture prior, according to Theorem 4, we can take

$$\epsilon_n = n^{-\alpha/(2\alpha+1)} (\log n)^{1+\max\{0,(1-\tau_1)/2\}}.$$

For random series prior, according to Thereom 13, part (a), we take

$$\epsilon_n = n^{-\alpha/(2\alpha+1)} (\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}.$$

Then the similar calculation gives the conclusion. □

REFERENCES

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679.

Babenko, A. and Belitser, E. (2010). Oracle convergence rate of posterior under projection prior and Bayesian model selection. *Mathematical Methods of Statistics*, 19:219–245.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 70:825–848.

Barron, A., Birgè, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413.

Bartlett, M. S. (1963). Statistical estimation of density function. *Sankhyā*, 25:245–254.

Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *The Annals of Statistics*, 31:536–559.

Bhattacharya, P. K. (1967). Estimation of a probability density function and its derivatives. *Sankhyā Series A*, 29:373–382.

Birgé, L. (2006). Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques*, 42:273–325.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distribution via Polya urn scheme. *The Annals of Statistics*, 1:353–355.

Brown, P. J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society Series B*, 60:627–641.

Brown, P. J., Vannucci, M., and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, Series B*, 64:519–536.

Cai, T. and Yuan, M. (2011). Optimal estimation of the mean function based on discretely sampled funcitonal data: phase transition. *The Annals of Statistics*, 39:2330–2355.

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13:571–591.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society Series B*, 57:473–484.

Castillo, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics*, 2:1281–1299.

Castillo, I. (2012). A semi-parametric Bernstein-von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields*, 152:53–99.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40:2069–2101.

Choi, T. and Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98:1969–1987.

Choudhuri, N., Ghosal, S., and Roy, A. (2004a). Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association*, 99:1050–1059.

Choudhuri, N., Ghosal, S., and Roy, A. (2004b). Contiguity of the Whittle measure for a Gaussian time series. *Biometrika*, 91:211–218.

Choudhuri, N., Ghosal, S., and Roy, A. (2007). Nonparametric binary regression using a Gaussian process prior. *Statistical Methodology*, 4:227–243.

Chung, Y. and Dunson, D. B. (2009). Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104:1646–1660.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer.

de Boor, C. and Daniel, J. W. (1974). Splines with nonnegative B-spline coefficients. *Mathematics of Computation*, 28:565–568.

de Jonge, R. and van Zanten, 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, H. (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electronic Journal of Statistics*, 6:1984–2001.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society. Series B*, 60:333–350.

Dunson, D. B. (2010). Nonparametric Bayes applications to biostatistics. In Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., editors, *Bayesian Nonparametrics*, pages 223–273. Cambridge University Press.

Dunson, D. B. and Park, J. H. (2008). Kernel stick-breaking processes. *Biometrika*, 95:307–323.

Dunson, D. B., Pillai, N., and Park, J. H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society Series B*, 69:163–183.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B*, 70:849–911.

Fan, J., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83:189–206.

Fan, J. and Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika*, 91:819–834.

Farrell, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *The Annals of Mathematical Statistics*, 43:170–180.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230.

Ferguson, T. S. (1974). Prior distribtion on the spaces of probability measures. *The Annals of Statistics*, 2:615–629.

Ferguson, T. S. (1983). *Bayesian density estimation by mixtures of normal distributions*, pages 287–302. Academic Press.

Freedman, D. (1963). On the asymptotic distribution of bayes estimates in the discrete case I. *The Annals of Mathematical Statistics*, 34:1386–1403.

Freedman, D. (1965). On the asymptotic distribution of bayes estimates in the discrete case II. *The Annals of Mathematical Statistics*, 36:454–456.

Freedman, D. (1999). On the bernstein-von mises theorem with infinite dimensional parameters. *The Annals of Statistics*, 27:1119–1140.

Gasparini, M. (1996). Bayesian density estimation via Dirichlet density processes. *Journal of Nonparametric Statistics*, 6:355–366.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373.

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *The Annals of Statistics*, 29(5):1264–1280.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531.

Ghosal, S., Lember, J., and van der Vaart, A. (2003). On Bayesian adaptation. *In Proceedings of the Eighth Vilnius Conference on Probability Theory and Mathematical Statistics, Part II (2002)*, 79:165–175.

Ghosal, S., Lember, J., and van der Vaart, A. (2008). Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89.

Ghosal, S. and Roy, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, 34:2413–2429.

Ghosal, S. and van der Vaart, A. (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. (2007a). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35:192–223.

Ghosal, S. and van der Vaart, A. (2007b). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35:697–723.

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*, 39:2883–2911.

Goldsmith, J., Wand, M. P., and Crainiceanu, C. (2011). Functional regression via variational Bayes. *Electronic Journal of Statistics*, 5:572–602.

Green, P, J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Griffin, J. E. and Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101:179–194.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35:70–91.

Hall, P., Wolff, R. C., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94:154–163.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, 35:403–410.

Hanson, T. and Johnson, W. (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association*, 97:1020–1033.

Hasminskii, R. Z. (1978). A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory of Probability and Its Applications*, 23:794–796.

Hesthaven, J. S., Gottlieb, S., and Gottlieb, D. (2007). *Spectral Methods for Time-Dependent Problems*. Cambridge University Press.

Hoffmann, M. and Lepski, O. (2002). Random rates in anisotropic regression. *The Annals of Statistics*, 30:325–358.

Hosseinioun, N., Doosti, H., and Niroumand, H. A. (2011). Nonparametric estimation of a multivariate probablity density for mixing sequences by the method of wavelets. *Italian Journal of Pure and Applied Mathematics*, 28:31–40.

Hosseinioun, N., Doosti, H., and Nirumand, H. A. (2012). Nonparametric estimation of the derivatives of a density by the method of wavelet for mixing sequences. *Statistical Papers*, 53:195–203.

Huang, T.-M. (2004). Convergence rates for posterior distributions and adaptive estimation. *The Annals of Statistics*, 32:1556–1593.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173.

Jackson, D. (1930). *The Theory of Approximation*. AMS Colloquium Publication Volume XI, New York.

Jara, A. and Hanson, T. E. (2011). A class of mixture of dependent tail-free processes. *Biometrika*, 98:553–566.

Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *The Annals of Statistics*, 35:1487–1511.

Kemperman, J. H. B. (1969). On the optimum rate of transmitting information. *The Annals of Mathematical Statistics*, 40:2156–2177.

Kim, Y. and Lee, J. (2001). On posterior consistency of survival models. *The Annals of Statistics*, 29:666–686.

Kruijer, W., Rousseau, J., and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257.

Kruijer, W. and van der Vaart, A. (2008). Posterior convergence rates for dirichlet mixtures of beta densities. *Journal of Statistical Planning and Inference*, 138:1981–1992.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *The Annals of Statistics*, 14:1222–1235.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modeling. *The Annals of Statistics*, 22:1161–1176.

Le Cam, L. (1986). *Asymptotic Methods in statistical Decision Theory*. Springer-Verlag, New York.

Lenk, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *Journal of the American Statistical Association*, 83:509–516.

Lenk, P. J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, 78:531–543.

Leonard, T. (1978). Density estimaiton, stochastic processes and prior information (with discussion). *Journal of the Royal Statistical Society, Series B*, 40:113–146.

Lian, H. (2011). On posterior distribution of Bayesian wavelet thresholding. *Journal of Statistical Planning and Inference*, 141:318–324.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357.

Lorenz, G. G. (1953). *Bernstein Polynomials*. Univ. Toronto Press.

Ma, L. (2012). Recursive partitioning and Bayesian inference on conditional distributions. Technical report, Duke University.

MacEachern, S. N. and Müller, P. (1998). Estimating mixtures of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238.

Mallat, S, G. (1989). Multiresolution approximations and wavelet orthonormal bases of $L_2(R)$. *Transactions of the American Mathematical Society*, 315:69–87.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83:67–79.

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science*, 19:95–111.

Neal, R. M. (1998). Regression and classification using Gaussian process priors. *Bayesian Statistics*, 6:475501.

Norets, A. and Pelenis, J. (2012). Posterior consistency in conditional density estimation by covariate dependent mixtures. Technical report.

Papaspiliopoulos, O. and Robers, G. O. (2008). Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, 95:169–186.

Parzen, E. (1962). On the estimation of a probability density and mode. *The Annals of Mathematical Statistics*, 33:1065–1076.

Pati, D., Dunson, D., and Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis*, 116:456–472.

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics*, 27:105–126.

Petrone, S. (1999b). Random Bernstein polynomials. *Scandinavian Journal of Statistics*, 26:373–393.

Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian white noise. *Problems of Information Transmission*, 16:120–133.

Prakasa Rao, B. (1996). Nonparametric estimation of the derivatives of a density by the method of wavelets. *Bulletin of Informatics and Cybernetics*, 28:91–100.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Rivoirard, V. and Rousseau, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Analysis*, 7:311–334.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27:832–837.

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*, 38:146–180.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, 71:319–392.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press.

Schuster, E. F. (1969). Estimation of a probability density function and its derivatives. *The Annals of Mathematical Statistics*, 40:1187–1195.

Schwartz, L. (1965). On Bayes procedures. *Probability Theory and Related Fields*, 4:10–26.

Scott, D. W. (1992). *Multivariate Density Estimation. Thoery, Practice and Visualization*. Wiley, New York.

Scricciolo, C. (2006). Convergence rates for Bayesian density estimation on infinite-dimensional exponential families. *The Annals of Statistics*, 34:2897–2920.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *The Annals of Statistics*, 29:687–714.

Singh, R. S. (1977a). Applications of estimators of a density and its derivatives to certain statistical problems. *Journal of the Royal Statistical Society, Series B*, 39:357–363.

Singh, R. S. (1977b). Improvement on some known nonparametric uniformly consistent estimators of derivatives of a density. *The Annals of Statistics*, 5:394–399.

Stone, C. J. (1990). Large-sample inference for log-spline models. *The Annals of Statistics*, 18:717–741.

Sugiyama, M., Takeuchi, I., Kanamori, T., Suzuki, T., Hachiya, H., and Okanohara, D. (2010). Least-squares conditional density estimation. *IEICE Transactions on Information and Systems*, E93-D:583–594.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Tokdar, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā*, 68:90–110.

Tokdar, S. T. (2007). Towards a faster implementation of density estimation with logistic gaussian process priors. *Journal of Computational and Graphical Statistics*, 16:633–655.

Tokdar, S. T. and Ghosh, J. K. (2007). Posterior consistency of logistic gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 137:34–42.

Tokdar, S. T., Zhu, Y. M., and Ghosh, J. K. (2010). Bayesian density regression with logistic gaussian process and subspace projection. *Bayesian Analysis*, 5:319–344.

Trippa, L., Müller, P., and Johnson, W. (2011). The multivariate beta process and an extension of the polya tree model. *Biometrika*, 98:17–34.

Truong, Y., Kooperberg, C., and Stone, C. (2005). *Statistical Modeling with Spline Functions: Methodology and Theory*. Springer.

van der Vaart, A. and van Zanten, H. (2007). Bayesian inference with rescaled Gaussian process priors. *Electronic Journal of Statistics*, 1:433–448.

van der Vaart, A. and van Zanten, H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36:1435–1463.

van der Vaart, A. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Walker, S. G. (2007). Sampling the dirichlet mixture model with slices. *Communications in Statistics*, 36:45–54.

Wasserman, L. (2006). *All of Nonparametric Statistics: A Concise Course in Nonparametric Statistical Inference*. New York: Springer-Verlag.

Whittle, P. (1957). Curve and periodogram smoothing (with discussion). *Journal of the Royal Statistical Society, Series B*, 19:38–63.

Wu, Y. and Ghosal, S. (2010). The L_1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *Journal of Multivariate Analysis*, 101:2411–2419.

Yang, Y. and Dunson, D. (2012). Bayesian conditional tensor factorizations for high-dimensional classification. Technical report, arXiv:1301.4950.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Goel, P. K. and Zellner, A., editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. North-Holland/Elsevier.

Zhao, L. H. (2000). Bayesian aspects of some nonparametric problems. *The Annals of Statistics*, 28:532–552.

APPENDIX

Appendix A

Introduction to B-splines

A.1 Univariate B-splines

Here we provide a brief introduction to B-splines. Following de Boor (2001), we divide the unit interval $[0, 1]$ into K equally spaced subintervals and consider spline functions, which are continuous, piecewise polynomials of degree at most q and are $(q - 2)$ times continuously differentiable. Then it can be shown that these spline functions form a J-dimensional linear space, where J is defined as $J = q + K - 1$. A convenient basis for this space is given by B-splines. The B-splines satisfy the following properties:

(i) $B_j(x) \geq 0$.

(ii) $\sum_{j=1}^{J} B_j(x) = 1$, for every $x \in (0, 1)$.

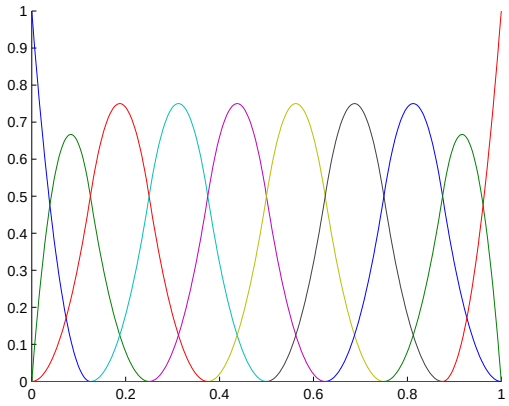
(iii) For every x, $B_j(x) > 0$ for at most q indices.

(iv) Each B_j is supported only on an interval of length q/K.

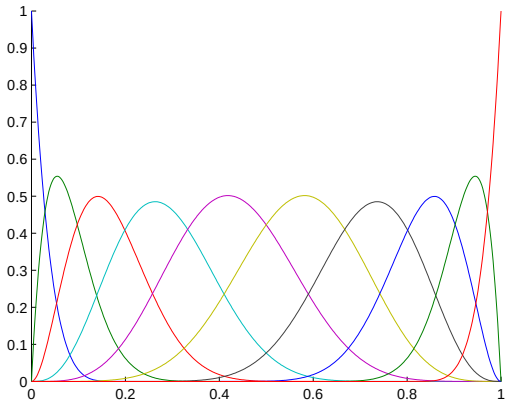
(v) B_j's are uniformly bounded.

Figure A.1: B-spline basis functions

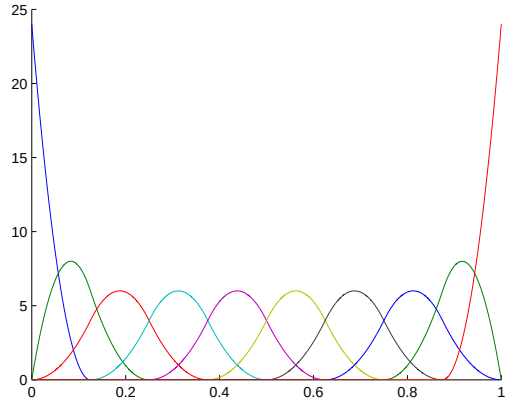
(a) B-spline, $q = 3$



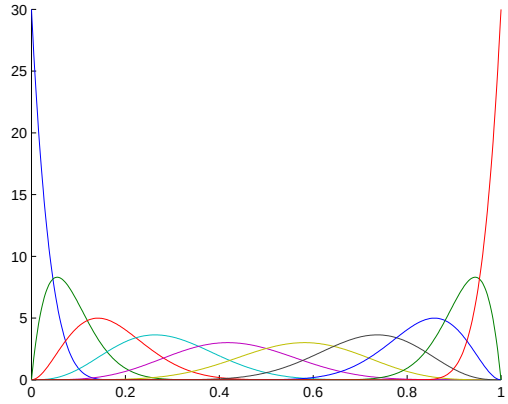
(b) B-spline, $q = 6$



(c) Normalized B-spline, $q = 3$



(d) Normalized B-spline, $q = 6$



A.2 Tensor-product B-splines

The tensor-product B-spline basis function is defined as the product of univariate basis functions of each direction. For $[0,1]^d$ and $d \in \mathbb{N}$, we split $[0,1]$ into J_k equal intervals at k-th direction. Then we have

$$B_{\boldsymbol{j}}(x_1,\ldots,x_d) = \prod_{k=1}^{d} B_{j_k}(x_k), j_k = 1,\ldots,J_k,\ k = 1,\ldots,d. \tag{A.1}$$

Hence there are $\prod_{i=1}^{d} J_k$ equal cubes in total. For simplicity, we use $\boldsymbol{B}$ to denote a column vector of all basis functions and define the total number of basis functions as $J = \prod_{k=1}^{d} J_k$.

Tensor-product B-splines maintain a lot of nice properties that the univariate B-spline enjoys. In the following, we list a few of them that will be used in our modeling:

(i) $B_{\boldsymbol{j}} \geq 0$.

(ii) $\sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} B_{\boldsymbol{j}}(\boldsymbol{x}) = 1$, for every $\boldsymbol{x} \in (0,1)^d$.

(iii) For every $\boldsymbol{x}$, $B_{\boldsymbol{j}}(\boldsymbol{x}) > 0$ for at most q^d indices.

(iv) $B_{\boldsymbol{j}}$'s are uniformly bounded.