Journal of Multivariate Analysis 116 (2013) 456–472

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Posterior consistency in conditional distribution estimation

Debdeep Pati [a,*], David B. Dunson [b], Surya T. Tokdar [b]

[a] Department of Statistics, Florida State University, United States
[b] Department of Statistical Science, Duke University, United States

ARTICLE INFO

Article history:
Received 23 February 2012
Available online 31 January 2013

AMS 2000 subject classifications:
62G07
62G20
60K35

Keywords:
Asymptotics
Bayesian nonparametrics
Density regression
Dependent Dirichlet process
Large support
Probit stick-breaking process

ABSTRACT

A wide variety of priors have been proposed for nonparametric Bayesian estimation of conditional distributions, and there is a clear need for theorems providing conditions on the prior for large support, as well as posterior consistency. Estimation of an uncountable collection of conditional distributions across different regions of the predictor space is a challenging problem, which differs in some important ways from density and mean regression estimation problems. Defining various topologies on the space of conditional distributions, we provide sufficient conditions for posterior consistency focusing on a broad class of priors formulated as predictor-dependent mixtures of Gaussian kernels. This theory is illustrated by showing that the conditions are satisfied for a class of generalized stick-breaking process mixtures in which the stick-breaking lengths are monotone, differentiable functions of a continuous stochastic process. We also provide a set of sufficient conditions for the case where stick-breaking lengths are predictor independent, such as those arising from a fixed Dirichlet process prior.

1. Introduction

One of the most common problems in data analysis is the need to characterize the dependence of a response on predictors in a flexible manner. We want to avoid parametric assumptions on the response density and how features, such as the mean, variance, skewness, shape and even modality, change with predictors. Nonparametric estimates of the conditional distribution [11,38] are appealing in this context, but in most applications one requires not just a point estimate but also a characterization of uncertainty. For this reason, and because of excellent practical performance in a rich variety of application areas, Bayesian approaches for conditional distribution estimation have become popular in recent years. The most common class of models is infinite mixture models due in part to the rich literature on algorithms for posterior computation using Markov chain Monte Carlo (MCMC) [22,49,33] and fast approximations [28]. Such MCMC algorithms are straightforward to implement, and the output can be used to estimate exact posterior densities for functionals of interest.

The ever increasing literature on new nonparametric Bayes models and exciting new applications in areas ranging from finance to biostatistics to machine learning has generated considerable enthusiasm. However, there is a clear lack of frequentist asymptotic theory supporting these models. The emphasis of this article is on substantially closing this gap focusing on a new class of generalized stick-breaking process (gSB) priors, which encompasses a number of the most widely applied priors as special cases.

In the absence of predictors, there is a rich theory and methods literature on nonparametric Bayes methods for estimating a density f using mixture models of the form

$$y_i \sim f, \quad f \sim \Pi, \tag{1.1}$$

* Corresponding author.
E-mail addresses: debdeep.isi@gmail.com, sfujimoto@sigmath.es.osaka-u.ac.jp (D. Pati).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2013.01.011

where Π is a mixture prior of the form $\sum_{h=1}^{\infty} \pi_h k(y;\theta_h)$ for suitably chosen kernel k, atoms and weights $\{(\theta_h, \pi_h), h = 1, \ldots, \infty\}$ with $\sum_{h=1}^{\infty} \pi_h = 1$ almost surely. The most common choice of Π is the Dirichlet process mixture of normals, first introduced by [26]. Original works on Dirichlet process can be found in [12,13]. Support of Π in (1.1) and asymptotic properties of the posterior are now well-understood [3,16,43,17,18,4].

Recent literature has focused on generalizing model (1.1) to the density regression setting in which the entire conditional distribution of y given $\mathbf{x}$ changes flexibly with predictors. Bayesian density regression views the entire conditional density $f(y \mid \mathbf{x})$ as a function valued parameter and allows its center, spread, skewness, modality and other such features to vary with $\mathbf{x}$. For data $\{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ let

$$y_i \mid \mathbf{x}_i \sim f(\cdot \mid \mathbf{x}_i), \qquad \{f(\cdot \mid \mathbf{x}), \mathbf{x} \in X\} \sim \Pi_{\mathcal{X}}, \tag{1.2}$$

where $\mathcal{X}$ is the predictor space and $\Pi_{\mathcal{X}}$ is a prior for the class of conditional densities $\{f_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ indexed by the predictors. Refer, for example, to [29,19,20,10,9,6,46] among others.

The primary focus of this recent development has been infinite mixture models of the form

$$f(y \mid \mathbf{x}) = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \phi \left\{ \frac{y - \mu_h(\mathbf{x})}{\sigma_h} \right\}, \tag{1.3}$$

where ϕ is the standard normal density, $\{\pi_h(\mathbf{x}), h = 1, 2, \ldots\}$ are predictor-dependent probability weights that sum to one almost surely for each $\mathbf{x} \in \mathcal{X}$, and $(\mu_h, \sigma_h) \sim G_0$ independently, with G_0 a base probability measure on $\mathcal{F}_{\mathcal{X}} \times \Re^+$, $\mathcal{F}_{\mathcal{X}} \subset \mathcal{X}^{\Re}$, the space of all $\mathcal{X} \to \Re$ functions. A single finite mixture of Gaussians is inadequate to represent the shape of the density $f(y \mid \mathbf{x})$ for different levels of the predictor $\mathbf{x}$ unless the number of components is huge. By using an infinite mixture we inherently allow for uncertainty in the number of components needed to characterize the data and bypass the difficult issue of selecting the number of components.

(1.1) is similar in spirit to kernel mixtures used in nonparametric smoothing approaches. However, a major advantage of using a Bayesian paradigm is that we do not need to deal with optimizing tuning parameters, which becomes difficult in higher dimensions. The new adaptation results [25,39] reveal that even a single prior specification can adapt to the unknown correct smoothness level of the true density and optimizes estimation in an asymptotic minimax sense. For conditional densities, smoothing needs to be done over the response space as well as the predictor space, making the choice of optimal smoothing even more difficult, especially when the predictors have varying degrees of influence on the response. A Bayesian approach offers an easier practical solution in this case.

To our knowledge, only [2] have considered formalizing the notions of support for dependent stick-breaking processes. We focus on a novel class of gSB processes, which express the probability weights $\pi_h(\mathbf{x})$ in stick-breaking form, with the stick lengths constructed through mapping continuous stochastic processes to the unit interval using a monotone differentiable link function. This class includes dependent Dirichlet processes [27] as a special case.

Only a few papers have considered asymptotic properties of the posterior in conditional density estimation. [46] considers posterior consistency in estimating conditional distributions focusing exclusively on logistic Gaussian process priors [45]. Such priors lack the computational simplicity of the countable mixture priors in (1.3). [52] considers posterior consistency in conditional distribution estimation through a limited information approach by approximating the likelihood by the quantiles of the true distribution. [41,42] provide sufficient conditions for showing posterior consistency in estimating an autoregressive conditional density and a transition density rather than regression with respect to another covariate.

In this article, focusing on model (1.3), we initially provide sufficient conditions on the prior and true data-generating model under which the prior leads to weak and various types of strong posterior consistency. In this context, we first define notions of weak and L_1-integrated neighborhoods. We then show that the sufficient conditions are satisfied for gSB priors. The theory is illustrated through application to a model relying on probit transformations of Gaussian processes, an approach related to the probit stick-breaking process of [6,37]. We also considered Gaussian mixtures of fixed-π dependent processes [27,8].

[31] showed posterior consistency in conditional density estimation using kernel stick breaking process mixtures of Gaussians in a very recent unpublished article. They approximated a conditional density by a smooth mixture of linear regressions as in [30] to demonstrate the KL property. In this paper, we have shown KL support using a more direct approach of approximating the true density by a kernel mixture of a compactly supported conditional measure.

The fundamental contribution of this article is formalizing the notion of support of the gSB process mixture of Gaussians on the space of conditional densities and formulating sufficient conditions to ensure that it leads to a consistent posterior. In doing so, a key technical contribution is the development of a novel method of constructing a sieve for the proposed class of priors. It has been noted by [51] that the usual method of constructing a sieve by controlling prior probabilities is unable to lead to a consistency theorem in the multivariate case. This is because of the explosion of the L_1-metric entropy with increasing dimension. They developed a technique specific to the Dirichlet process in the multivariate case for showing weak and strong posterior consistency. The proposed sieve[1] avoids the pitfall mentioned by [51] in showing consistency using multivariate mixtures. Our sieve construction has been recently used for studying convergence rates in multivariate density estimation [40,44].

[1] A similar sieve appears in [31] with a citation to an earlier draft of our paper.

2. Notations

Throughout the paper, Lebesgue measure on $\Re$ or $\Re^p$ is denoted by λ and the set of natural numbers by $\mathbb{N}$. The supremum and the L_1-norms are denoted by $\|\cdot\|_\infty$ and $\|\cdot\|_1$ respectively. The indicator function of a set B is denoted by 1_B. Let $L_p(\nu, M)$ denote the space of real valued measurable functions defined on M with ν-integrable pth absolute power. For two density functions f, g, the Kullback–Leibler divergence is given by $K(f, g) = \int \log(f/g) f d\lambda$. A ball of radius r with centre x_0 relative to the metric d is defined as $B(x_0, r; d)$. The diameter of a bounded metric space M relative to a metric d is defined to be $\sup\{d(x, y) : x, y \in M\}$. The ϵ-covering number $N(\epsilon, M, d)$ of a semi-metric space M relative to the semi-metric d is the minimal number of balls of radius ϵ needed to cover M. The logarithm of the covering number is referred to as the entropy. "$\lesssim$" stands for inequality up to a constant multiple or if the constant multiple is irrelevant to the given situation. δ_0 stands for a distribution degenerate at 0 and $\text{supp}(\nu)$ for the support of a measure ν.

3. Conditional density estimation

In this section, we will define the space of conditional densities and construct a prior on this space. It is first necessary to generalize the topologies to allow appropriate neighborhoods to be constructed around an uncountable collection of conditional densities indexed by predictors. With such neighborhoods in place, we then state our main theorems providing sufficient conditions under which various modes of posterior consistency hold for a broad class of predictor-dependent mixtures of Gaussian kernels.

Let $\mathcal{Y} = \Re$ be the response space and $\mathcal{X}$ be the covariate space which is a compact subset of $\Re^p$. Unless otherwise stated, we will assume $\mathcal{X} = [0, 1]^p$ without loss of generality. Let $\mathcal{F}$ denote the space of densities on $\mathcal{X} \times \mathcal{Y}$ w.r.t. the Lebesgue measure and $\mathcal{F}_d$ denote a subset of the space of conditional densities satisfying,

$$\mathcal{F}_d = \left\{ g : \mathcal{X} \times \mathcal{Y} \to (0, \infty), \int_{\mathcal{Y}} g(\mathbf{x}, y) dy = 1 \, \forall \, \mathbf{x} \in \mathcal{X}, \mathbf{x} \mapsto g(\mathbf{x}, \cdot) \right.$$
$$\left. \text{continuous as a function from } \mathcal{X} \to L_1(\lambda, \mathcal{Y}) \right\}.$$

Suppose y_i is observed independently given the covariates $\boldsymbol{x}_i$, $i = 1, 2, \ldots$ which are drawn independently from a probability distribution Q on $\mathcal{X}$. Assume that Q admits a density q with respect to the Lebesgue measure.

If we define $h(\mathbf{x}, y) = q(\mathbf{x}) f(y \mid \mathbf{x})$ and $h_0(\mathbf{x}, y) = q(\mathbf{x}) f_0(y \mid \mathbf{x})$ then $h, h_0 \in \mathcal{F}$. Throughout the paper, h_0 is assumed to be a fixed density in $\mathcal{F}$ which we alternatively refer to as the *true data generating density* and $\{f_0(\cdot \mid \mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ is referred to as the true conditional density. The density $q(\mathbf{x})$ will be needed only for theoretical investigation. In practice, we do not need to know it or learn it from the data.

We propose to induce a prior $\Pi_{\mathcal{X}}$ on the space of conditional densities through a prior $\mathcal{P}_{\mathcal{X}}$ for a collection of mixing measures $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ using the following predictor-dependent mixture of kernels

$$f(y \mid \mathbf{x}) = \int \frac{1}{\sigma} \phi\left(\frac{y - \mu}{\sigma}\right) dG_{\mathbf{x}}(\psi), \tag{3.1}$$

where $\psi = (\mu, \sigma)$, and

$$G_{\mathbf{x}} = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \delta_{\{\mu_h(\mathbf{x}), \sigma_h\}}, \qquad (\mu_h, \sigma_h) \sim G_0, \tag{3.2}$$

where $\pi_h(\mathbf{x}) \geq 0$ are random functions of $\mathbf{x}$ such that $\sum_{h=1}^{\infty} \pi_h(\mathbf{x}) = 1$ a.s. for each fixed $\mathbf{x} \in \mathcal{X}$. $\{\mu_h(\mathbf{x}), \mathbf{x} \in X\}_{h=1}^{\infty}$ are i.i.d. realizations of a real valued stochastic process, i.e., G_0 is a probability distribution over $\mathcal{F}_{\mathcal{X}} \times \Re^+$, where $\mathcal{F}_{\mathcal{X}} \subset \mathcal{X}^{\Re}$, $\mathcal{X}^{\Re}$ being the space of functions from $\mathcal{X}$ to $\Re$. Hence for each $\mathbf{x} \in \mathcal{X}$, $G_{\mathbf{x}}$ is a random probability measure over the measurable Polish space $(\Re \times \Re^+, \mathcal{B}(\Re \times \Re^+))$. We are interested the following two important special cases.

3.1. Predictor dependent countable mixtures of Gaussian linear regressions

We define the predictor dependent countable mixtures of Gaussian linear regressions ($\text{MGLR}_{\mathbf{x}}$) as

$$f(y \mid \mathbf{x}) = \int \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma),$$

and

$$G_{\mathbf{x}} = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \delta_{(\boldsymbol{\beta}_h, \sigma_h)}, \quad (\boldsymbol{\beta}_h, \sigma_h) \sim G_0 \tag{3.3}$$

where $\pi_h(\mathbf{x}) \geq 0$ are random functions of $\mathbf{x}$ such that $\sum_{h=1}^{\infty} \pi_h(\mathbf{x}) = 1$ a.s. for each fixed $\mathbf{x} \in \mathcal{X}$ and $G_0 = G_{0,\boldsymbol{\beta}} \times G_{0,\sigma}$ is a probability distribution on $\Re^p \times \Re^+$ where $G_{0,\boldsymbol{\beta}}$ and $G_{0,\sigma}$ are probability distributions on $\Re^p$ and $\Re^+$ respectively. For a particular choice of $\pi_h(\mathbf{x})$'s, we obtain the probit stick-breaking mixtures of Gaussians which have been previously applied

to real data applications by [6,37,35]. The latter two articles considered probit transformations of Gaussian processes in constructing the stick-breaking weights.

3.2. Gaussian mixtures of fixed-π dependent processes

In (3.1), set $G_{\mathbf{x}}$ as in (3.2) with $\pi_h(\mathbf{x}) \equiv \pi_h$ for all $\mathbf{x} \in \mathcal{X}$ where $\pi_h \geq 0$ are random probability weights $\sum_{h=1}^{\infty} \pi_h = 1$ a.s. and $\{\mu_h(\mathbf{x}), \mathbf{x} \in X\}_{h=1}^{\infty}$ are as in (3.2). Examples include fixed-π dependent Dirichlet process mixtures of Gaussians [27]. Versions of the fixed π-DDP have been applied to ANOVA [8], survival analysis [7,23], spatial modeling [15], and many more.

A Gaussian process is a common choice for constructing stochastic processes $\pi_h(\mathbf{x})$'s and $\mu_h(\mathbf{x})$'s. Recall that a Gaussian process $\{\alpha(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ is defined as a stochastic process for which any finite dimensional representation $\{\alpha(\mathbf{x}_1), \ldots, \alpha(\mathbf{x}_p)\}$, $p \geq 1$ has a joint Gaussian distribution. We denote by GP(μ, c) a Gaussian process with mean function $\mu : \mathcal{X} \to \mathbb{R}$ and $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

4. Notions of posterior consistency for conditional densities

We recall the definition of posterior consistency through $\mathbf{y}^n = (y_1, \ldots, y_n)$ and $\mathbf{x}^n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

Definition 4.1. The posterior $\Pi_{\mathcal{X}}\left(\cdot \mid \mathbf{y}^n, \mathbf{x}^n\right)$ is consistent at $\{f_0(\cdot \mid \mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ with respect to a given topology if $\Pi_{\mathcal{X}}\left(U^c \mid \mathbf{y}^n, \mathbf{x}^n\right) \to 0$ a.s. for an arbitrary neighborhood U of $\{f_0(\cdot \mid \mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ in that topology.

Here a.s. consistency at $\{f_0(\cdot \mid \mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ means that the posterior distribution concentrates around a neighborhood of $\{f_0(\cdot \mid \mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ for almost every sequence $\{y_i, \mathbf{x}_i\}_{i=1}^{\infty}$ generated by i.i.d. sampling from the joint density $q(\mathbf{x})f_0(y \mid \mathbf{x})$.

We define the weak and ν-integrated L_1 neighborhoods of a collection of conditional densities $\{f_0(\cdot \mid \mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ in the following. A sub-base of a weak neighborhood is defined as

$$W_{\epsilon,g}(f_0) = \left\{f : f \in \mathcal{F}_d, \left|\int_{\mathcal{X}\times\mathcal{Y}} gh - \int_{\mathcal{X}\times\mathcal{Y}} gh_0\right| < \epsilon\right\}, \tag{4.1}$$

for a bounded continuous function $g : \mathcal{Y} \times \mathcal{X} \to \Re$. A weak neighborhood base is formed by finite intersections of neighborhoods of the type (4.1). Define a ν-integrated L_1 neighborhood

$$S_{\epsilon}(f_0; \nu) = \left\{f : f \in \mathcal{F}_d, \int \|f(\cdot \mid \mathbf{x}) - f_0(\cdot \mid \mathbf{x})\|_1 \, \nu(\mathbf{x})d\mathbf{x} < \epsilon\right\} \tag{4.2}$$

for any measure ν with $\text{supp}(\nu) \subset \mathcal{X}$. Observe that under the topology in (4.2), $\mathcal{F}_d$ can be identified to a closed subset of $L_1(\lambda \times \nu, \mathcal{Y} \times \text{supp}(\nu))$ making it a complete separable metric space. Thus measurability issues will not arise with these topologies.

In the following, we define the Kullback–Leibler (KL) property of $\Pi_{\mathcal{X}}$ at a given $f_0 \in \mathcal{F}_d$. Note that we define a KL-type neighborhood around the collection of conditional densities f_0 through defining a KL neighborhood around the joint density h_0, while keeping Q fixed at its true unknown value.

Definition 4.2. For any $f_0 \in \mathcal{F}_d$, such that $h_0(\mathbf{x}, y) = q(\mathbf{x})f_0(y \mid \mathbf{x})$ is the true joint data-generating density, we define an ϵ-sized KL neighborhood around f_0 as

$$K_{\epsilon}(f_0) = \{f : f \in \mathcal{F}_d, \text{KL}(h_0, h) < \epsilon, h(\mathbf{x}, y) = q(\mathbf{x})f(y \mid \mathbf{x}) \; \forall y \in \mathcal{Y}, \mathbf{x} \in \mathcal{X}\},$$

where $\text{KL}(h_0, h) = \int h_0 \log(h_0/h)$. Then, $\Pi_{\mathcal{X}}$ is said to have KL property at $f_0 \in \mathcal{F}_d$, denoted $f_0 \in \text{KL}(\Pi_{\mathcal{X}})$, if $\Pi_{\mathcal{X}}\{K_{\epsilon}(f_0)\} > 0$ for any $\epsilon > 0$.

Another definition we would require for showing the KL support is the notion of weak neighborhood of a collection of mixing measures $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ where $G_{\mathbf{x}}$ is a probability measure on $S \times \Re^+$ for each $\mathbf{x} \in \mathcal{X}$. Here $S = \Re^p$ or $\Re$ depending on the cases considered above. We formulate the notion of a sub-base of the weak neighborhood of $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ below.

Definition 4.3. For a bounded continuous function $g : S \times \Re^+ \times \mathcal{X} \to \Re$ and $\epsilon > 0$, a sub-base of the weak neighborhood of a conditional probability measure $\{F_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ is defined as

$$\left\{\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} : \left|\int_{S\times\Re^+\times\mathcal{X}} g(s, \sigma, \mathbf{x})dG_{\mathbf{x}}(s, \sigma)q(\mathbf{x})d\mathbf{x} - g(s, \sigma, \mathbf{x})dF_{\mathbf{x}}(s, \sigma)q(\mathbf{x})d\mathbf{x}\right| < \epsilon\right\}. \tag{4.3}$$

A conditional probability measure $\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ lies in the weak support of $\mathcal{P}_{\mathcal{X}}$ if $\mathcal{P}_{\mathcal{X}}$ assigns positive probability to every basic neighborhood generated by the sub-base of the type (4.3). In the sequel, we will also consider a neighborhood of the form

$$\left\{\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} : \sup_{\mathbf{x}\in\mathcal{X}} \left|\int_{S\times\Re^+} \{g(s, \sigma)dG_{\mathbf{x}}(s, \sigma) - g(s, \sigma)dF_{\mathbf{x}}(s, \sigma)\}\right| < \epsilon\right\} \tag{4.4}$$

for a bounded continuous function $g : S \times \Re^+ \to \Re$.

5. Posterior consistency in MGLR$_{\mathbf{x}}$ mixture of Gaussians

5.1. Kullback–Leibler property

We will work with a specific choice of $\mathcal{P}_{\mathcal{X}}$ motivated by the probit stick breaking process construction in [6]. Let

$$\pi_h(\mathbf{x}) = \Phi\{\alpha_h(\mathbf{x})\}\prod_{l<h}[1-\Phi\{\alpha_l(\mathbf{x})\}], \tag{5.1}$$

where $\alpha_h \sim \text{GP}(0, c_h)$, for $h = 1, 2, \ldots, \infty$. Assume the following holds.

S1. c_h is chosen so that $\alpha_h \sim \text{GP}(0, c_h)$ has continuous path realizations
S2. for any continuous function under the GP$(0, c_h)$ prior for $\alpha_h g : \mathcal{X} \mapsto \Re$,

$$\mathcal{P}_{\mathcal{X}}\left\{\sup_{\mathbf{x}\in\mathcal{X}}|\alpha_h(\mathbf{x}) - g(\mathbf{x})| < \epsilon\right\} > 0$$

$h = 1, \ldots, \infty$ and for any $\epsilon > 0$.
S3. G_0 is absolutely continuous with respect to $\lambda(\Re^p \times \Re^+)$.

Consider the subset $\mathcal{F}_d^* \subset \mathcal{F}_d$ satisfying the following conditions.

A1. f is nowhere zero and bounded by $M < \infty$.
A2. $|\int_{\mathcal{X}}\int_{\mathcal{Y}} f(y \mid \mathbf{x}) \log f(y \mid \mathbf{x}) dyq(\mathbf{x})d\mathbf{x}| < \infty$.
A3. $|\int_{\mathcal{X}}\int_{\mathcal{Y}} f(y \mid \mathbf{x}) \log \frac{f(y|\mathbf{x})}{\psi_{\mathbf{x}}(y)} dyq(\mathbf{x})d\mathbf{x}| < \infty$,
where $\psi_{\mathbf{x}}(y) = \inf_{t\in[y-1,y+1]} f(t \mid \mathbf{x})$.
A4. $\exists\, \eta > 0$ such that $\int_{\mathcal{X}}\int_{\mathcal{Y}} |y|^{2(1+\eta)} f(y \mid \mathbf{x}) dyq(\mathbf{x})d\mathbf{x} < \infty$.
A5. $(\mathbf{x}, y) \mapsto f(y \mid \mathbf{x})$ is jointly continuous.

Remark 5.1. A1 is usually satisfied by common densities arising in practice. A4 imposes a minor tail restriction; e.g., a mean regression model with continuous mean function and a heavy-tailed t residual density with 4 degrees of freedom satisfies A4. Conditions A2 and A3 are more subtle, but are also mild. A flexible class of models which satisfies A1–A5 is as follows. Let $y_i = \mu(x_i) + \epsilon_i$, with $\mu : \mathcal{X} \to \Re$ continuous and $\epsilon_i \sim f_{x_i}$, where $f_x(\epsilon) = \sum_{h=1}^{H} \pi_h(x)\psi(\epsilon; \mu_h, \sigma_h^2)$ for some $H \geq 1$, $\sum_{h=1}^{H}\pi_h(x) = 1$, $\pi_h : \mathcal{X} \to [0, 1]$ continuous and ψ is Gaussian or t with greater than 2 degrees of freedom.

Remark 5.2. S2 is satisfied if $c_h(\mathbf{x}, \mathbf{x}') = e^{-A_h\|\mathbf{x}-\mathbf{x}'\|^2}$ and the prior for A_h has full support on $\mathbb{R}^+$.

The following theorem characterizes the subset of $\mathcal{F}_d$ for which $\Pi_{\mathcal{X}}$ has the KL property. The proof of Theorem 5.3 is provided in Appendix C.

Theorem 5.3. *$f_0 \in \text{KL}(\Pi_{\mathcal{X}})$ for each f_0 in $\mathcal{F}_d^*$ if $\mathcal{P}_{\mathcal{X}}$ satisfies* S1–S3.

Remark 5.4. The conditions are satisfied for a class of gSB process mixtures in which the stick-breaking lengths are constructed through mapping continuous stochastic processes to the unit interval using a monotone differentiable link function.

To prove Theorem 5.3, we need several auxiliary results related to the support of the prior $\mathcal{P}_{\mathcal{X}}$ which might be of independent interest. The key idea for showing that the true f_0 satisfies $\Pi_{\mathcal{X}}\{K_\epsilon(f_0)\} > 0$ for any $\epsilon > 0$ is to impose certain tail conditions on $f_0(y \mid \mathbf{x})$ and approximate it by $\tilde{f}(y \mid \mathbf{x}) = \int \frac{1}{\sigma}\phi\left(\frac{y-\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) d\tilde{G}_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)$, where $\{\tilde{G}_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ is compactly supported. Observe that,

$$\text{KL}(h_0, h) = \int_{\mathcal{X}}\int_{\mathcal{Y}} f_0(y \mid \mathbf{x}) \log\frac{f_0(y \mid \mathbf{x})}{\tilde{f}(y \mid \mathbf{x})} dyq(\mathbf{x})d\mathbf{x} + \int_{\mathcal{X}}\int_{\mathcal{Y}} f_0(y \mid \mathbf{x}) \log\frac{\tilde{f}(y \mid \mathbf{x})}{f(y \mid \mathbf{x})} dyq(\mathbf{x})d\mathbf{x}. \tag{5.2}$$

We construct such an $\tilde{f}$ in Theorem 5.3 which makes the first term in the right hand side of (5.2) sufficiently small. The following lemma (which is similar to Lemma 3.1 in [43] and Theorem 3 in [16]) guarantees that the second term in the right hand side of (5.2) is also sufficiently small if $\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ lies inside a finite intersection of neighborhoods of $\{\tilde{G}_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ of the type (4.4).

Lemma 5.5. *Assume that $f_0 \in \mathcal{F}_d$ satisfies $\int_{\mathcal{X}} \int_{\mathcal{Y}} y^2 f_0(y \mid \mathbf{x}) dy q(\mathbf{x}) d\mathbf{x} < \infty$. Suppose $\tilde{f}(y \mid \mathbf{x}) = \int \frac{1}{\sigma}\phi\big(\frac{y-\mathbf{x}'\boldsymbol{\beta}}{\sigma}\big) d\tilde{G}_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)$, where $\exists\, a > 0$ and $0 < \underline{\sigma} < \overline{\sigma}$ such that*

$$\tilde{G}_{\mathbf{x}}\big([-a, a]^p \times (\underline{\sigma}, \overline{\sigma})\big) = 1 \quad \forall\, \mathbf{x} \in \mathcal{X}, \tag{5.3}$$

so that $\tilde{G}_{\mathbf{x}}$ has compact support for each $\mathbf{x} \in \mathcal{X}$. Then given any $\epsilon > 0$, $\exists$ a finite intersection W of neighborhoods of $\{\tilde{G}_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ of the type (4.4) such that for any conditional density $f(y \mid \mathbf{x}) = \int \frac{1}{\sigma}\phi\big(\frac{y-\mathbf{x}'\boldsymbol{\beta}}{\sigma}\big) dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)$, $\mathbf{x} \in \mathcal{X}$, with $\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} \in W$,

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} f_0(y \mid \mathbf{x}) \log \frac{\tilde{f}(y \mid \mathbf{x})}{f(y \mid \mathbf{x})} dy q(\mathbf{x}) d\mathbf{x} < \epsilon. \tag{5.4}$$

The proof is similar to Theorem 3 in [16] and is omitted here. In order to ensure that the weak support of $\Pi_{\mathcal{X}}$ is sufficiently large to contain all densities $\tilde{f}$ satisfying the assumptions of Lemma 5.5, we define a collection of fixed conditional probability measures on $(\Re^p \times \Re^+, \mathcal{B}(\Re^p \times \Re^+))$ denoted by $\mathcal{G}^*_{\mathcal{X}}$ satisfying

1. $\mathbf{x} \mapsto F_{\mathbf{x}}(B)$ is a continuous function of $\mathbf{x} \in \mathcal{X}$, $\forall\, B \in \mathcal{B}(\Re^p \times \Re^+)$.
2. For any sequence of sets $A_n \subset \Re^p \times \Re^+ \downarrow \emptyset$, $\sup_{\mathbf{x} \in \mathcal{X}} F_{\mathbf{x}}(A_n) \downarrow 0$.

Next we state the theorem characterizing the weak support of $\mathcal{P}_{\mathcal{X}}$ which will be proved in Appendix B.

Theorem 5.6. *If $\mathcal{P}_{\mathcal{X}}$ satisfies* S1–S3*, then any $\{F_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} \in \mathcal{G}^*_{\mathcal{X}}$ lies in the weak support of $\mathcal{P}_{\mathcal{X}}$.*

Corollary 5.7. *Assume* S1–S3 *hold and assume $F_{\mathbf{x}} \in \mathcal{G}^*_{\mathcal{X}}$ is compactly supported, i.e., there exists $a, \underline{\sigma}, \overline{\sigma} > 0$ such that $F_{\mathbf{x}}([-a, a]^p \times [\underline{\sigma}, \overline{\sigma}]) = 1$. Then for a bounded uniformly continuous function $g : \Re^p \times \Re^+ \to [0, 1]$ satisfying $g(\beta, \sigma) \to 0$ as $\|\boldsymbol{\beta}\| \to \infty, \sigma \to \infty$,*

$$\mathcal{P}_{\mathcal{X}}\left\{\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} : \sup_{\mathbf{x} \in \mathcal{X}} \left| \int_{\Re^p \times \Re^+} \{g(\boldsymbol{\beta}, \sigma) dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma) - g(\boldsymbol{\beta}, \sigma) dF_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)\} \right| < \epsilon \right\} > 0. \tag{5.5}$$

Proof. The proof is similar to Theorem 5.6 with the L_1 convergence in (B.1) replaced by convergence uniformly in $\mathbf{x}$. This is because under the assumptions of Corollary 5.7, the uniformly continuous sequence of functions $\sum_{k=1}^n g(\tilde{\boldsymbol{\beta}}_{k,n}, \tilde{\sigma}_{k,n}) F_{\mathbf{x}}(A_{k,n})$ on $\mathcal{X}$ monotonically decreases to $\int_C g(\boldsymbol{\beta}, \sigma) dF_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)$ as $n \to \infty$ where C is given by $[-a, a]^p \times [\underline{\sigma}, \overline{\sigma}]$. □

The proof of the following corollary is along the lines of the proof of Theorem 5.6 and is omitted here.

Corollary 5.8. *Under the assumptions of Corollary 5.7 for any $k_0 \geq 1$,*

$$\mathcal{P}_{\mathcal{X}}\left\{\cap_{j=1}^{k_0} U_j\right\} > 0, \tag{5.6}$$

where U_j's are neighborhoods of the type (5.5).

5.2. Strong consistency with the q-integrated L_1 neighborhood

To obtain strong consistency in the q-integrated L_1 topology, we need a very straightforward extension of Theorem 2 of [16] below.

Theorem 5.9. *Suppose $f_0 \in \mathrm{KL}(\Pi_{\mathcal{X}})$ and there exist subsets $\mathcal{F}_n \subset \mathcal{F}_d$ with*

1. $\log N(\epsilon, \mathcal{F}_n, \|\cdot\|_1) = o(n)$,
2. $\Pi_{\mathcal{X}}(\mathcal{F}_n^c) \leq c_2 e^{-n\beta_2}$ *for some* $c_2, \beta_2 > 0$,

then the posterior is strongly consistent with respect to the q-integrated L_1 neighborhood.

Before stating the main theorem on strong consistency, we consider a hierarchical extension of $\mathrm{MGLR}_{\mathbf{x}}$ where the bandwidths are taken to be random. We define a sequence of random inverse-bandwidths A_h of the Gaussian process α_h, $h \geq 1$ each having $\Re^+$ as its support. Since the first few atoms suffice to explain most of the dependence of y on $\mathbf{x}$, we expect that the variability due to the covariate in the stochastic process $\Phi\{\alpha_h\}$ decreases as h increases. This is achieved through a carefully chosen prior for the covariance kernel c_h of the Gaussian process α_h.

Let α_0 denote the base Gaussian process on $[0, 1]^p$ with covariance kernel $c_0(\mathbf{x}, \mathbf{x}') = \tau^2 e^{-\|\mathbf{x}-\mathbf{x}'\|^2}$. Then $\alpha_h(\mathbf{x}) = \alpha_0(A_h^{1/2}\mathbf{x})$ for each $\mathbf{x} \in \mathcal{X}$. The variability of α_h with respect to the covariate is shrunk or stretched to the rectangle $[0, A_h^{1/2}]^p$ as A_h decreases or increases. A_h's are constructed to be stochastically decreasing to δ_0 in the following manner. We assume that there exist $\eta, \eta_0 > 0$ and a sequence $\delta_n = O((\log n)^2/n^{5/2})$ such that $P(A_h > \delta_n) \leq \exp\{-n^{-\eta_0} h^{(\eta_0+2)/\eta} \log h\}$ for each

$h \geq 1$. Also assume that there exists a sequence $r_n \uparrow \infty$ such that $r_n^p n^\eta (\log n)^{p+1} = o(n)$ and $P(A_h > r_n) \leq e^{-n}$. We will discuss how to construct such a sequence of random variables in the Remark 5.12 following Theorem 5.10.

The following theorem provides sufficient conditions for strong posterior consistency in the q-integrated L_1 topology. The proof is provided in Appendix D.

Theorem 5.10. *Let π_h's satisfy (5.1) with $\alpha_h \sim GP(0, c_h)$ where $c_h(\mathbf{x}, \mathbf{x}') = \tau^2 e^{-A_h \|\mathbf{x}-\mathbf{x}'\|^2}$, $h \geq 1$, $\tau^2 > 0$ fixed.*

C1. *There exist sequences $a_n, h_n \uparrow \infty, l_n \downarrow 0$ with $\frac{a_n}{l_n} = O(n)$, $\frac{h_n}{l_n} = O(e^n)$, and constants $d_1, d_2 > 0$ such that $G_0\{B(0; a_n) \times [l_n, h_n]\}^c < d_1 e^{-d_2 n}$ for some $d_1, d_2 > 0$.*

C2. *A_h's are constructed as in the second last paragraph before Theorem 5.10.*

Then $f_0 \in \mathrm{KL}(\Pi_{\mathcal{X}})$ implies that $\Pi_{\mathcal{X}}$ achieves strong posterior consistency in q-integrated L_1 topology at f_0.

Remark 5.11. Verification of condition C1 of Theorem 5.10 is particularly simple. For example, if G_0 is a product of multivariate normals on $\boldsymbol{\beta}$ and an inverse Gamma prior on σ^2, the condition C1 is satisfied with $a_n = O(\sqrt{n})$, $h_n = e^n$, $l_n = O(\frac{1}{\sqrt{n}})$. It follows from [48] that $f_0 \in \mathrm{KL}(\Pi_{\mathcal{X}})$ is still satisfied when we have the additional assumptions C1–C2 together with S1–S3 on the prior $\Pi_{\mathcal{X}}$.

Remark 5.12. Since we need $r_n^p n^\eta (\log n)^{p+1} = o(n)$, r_n^p can be chosen to be $O(n^{\eta_1})$ for some $0 < \eta_1 < 1$. Let d be such that $d\eta_1/p \geq 1$ and set $\eta_0 = 3d$. Let $A_h = c_h B_h$, where $B_h^q \sim \mathrm{Exp}(\lambda)$ and $c_h = (h^{(3d+2)/\eta} \log h)^{-1/d}$ for any $0 < \eta < 1$. Then $P(A_h > n^{\eta_1/p}) \leq P(B_h > n^{\eta_1/p}) \leq e^{-n^{d\eta_1/p}} \leq e^{-n}$ and $P(A_h > (\log n)^2 n^{-5/2}) \leq \exp\{-n^{-3d} h^{(3d+2)/\eta} \log h\}$.

Remark 5.13. The theory of strong posterior consistency can be generalized to an arbitrary monotone differentiable link function $L : \Re \mapsto [0, 1]$ which is Lipschitz, i.e., there exists a constant $K > 0$ such that $\left|L(\mathbf{x}) - L(\mathbf{x}')\right| \leq K \left|\mathbf{x} - \mathbf{x}'\right|$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Also, as long as the $\pi_h(\mathbf{x})$'s satisfy the hypothesis of Lemma A.1 and possess the required tail probability in Lemma 5.15, general predictor dependent mixing weights can be used.

Below we will develop several auxiliary results required to prove Theorem 5.10. They are stated below as some of them might be of independent interest. Let $\phi_{\boldsymbol{\beta},\sigma}(\mathbf{x}, y) := \frac{1}{\sigma}\phi\left(\frac{y-\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)$ for $y \in \mathcal{Y}$ and $\mathbf{x} \in \mathcal{X}$. From [43], we obtain for $\sigma_2 > \sigma_1 > \frac{\sigma_2}{2}$ and for each $\mathbf{x} \in \mathcal{X}$,

$$\int_{\mathcal{Y}} \left|\phi_{\boldsymbol{\beta}_1,\sigma_1}(\mathbf{x}, y) - \phi_{\boldsymbol{\beta}_2,\sigma_2}(\mathbf{x}, y)\right| dy \leq \left(\frac{2}{\pi}\right)^{1/2} \frac{\left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right\| \sqrt{p}}{\sigma_2} + \frac{3(\sigma_2 - \sigma_1)}{\sigma_1}.$$

Construct a sieve for $(\boldsymbol{\beta}, \sigma)$ as

$$\Theta_{a,h,l} = \left\{\phi_{\boldsymbol{\beta},\sigma} : \|\boldsymbol{\beta}\| \leq a, l \leq \sigma \leq h\right\}. \tag{5.7}$$

In the following lemma, we provide an upper bound to $N(\Theta_{a,h,l}, \epsilon, d_{SS})$. The proof is omitted as it follows trivially from Lemma 4.1 in [43].

Lemma 5.14. *There exist constants $d_1, d_2 > 0$ such that $N(\Theta_{a,h,l}, \epsilon, d_{SS}) \leq d_1\left(\frac{a}{l}\right)^p + d_2 \log \frac{h}{l} + 1$.*

In the proof of Theorem 5.10, we will verify the sufficient conditions of Theorem 5.9. We calibrate $\mathcal{F}_d$ by a carefully chosen sequence of subsets $\mathcal{F}_n \subset \mathcal{F}_d$. The fundamental problem with mixture models $\int N(y; \mu, \sigma^2 I_p) dP(\mu)$ in estimating a multivariate density lies in attempting to compactify the model space by $\{\int N(y; \mu, \sigma^2 I_p) dP(\mu) : P((-a_n, a_n]^p) > 1-\delta\}$ for each σ leading to an entropy a_n^p growing exponentially with the dimension p. Here we marginalize P in $\int \mathrm{N}(y; \mu, \sigma^2 I_p) dP(\mu)$ to yield the following construction $\{\sum_{h=1}^{m_n} \pi_h N(y; \mu_h, \sigma^2 I_p) : \|\mu_h\| \leq a_n, h = 1, \ldots, m_n, \sum_{h=m_n+1}^{\infty} \pi_h < \epsilon\}$ leading to an entropy $m_n \log a_n$ where m_n is related to the tail-decay of $P(\sum_{h=m_n+1}^{\infty} \pi_h > \epsilon)$. With this idea in place, we extend the construction of $\mathcal{F}_n$ for conditional densities below.

Before constructing a sieve, we briefly review alternative definitions [47] of a Gaussian process as a Banach space valued element below. A Borel measurable random element W with values in a separable Banach space $(\mathbb{B}, \|\cdot\|)$ is called Gaussian if the random variable b^*W is normally distributed for any element $b^* \in \mathbb{B}^*$, the dual space of $\mathbb{B}$. Recall that in general, the reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ attached to a zero-mean Gaussian process W is defined as the collection of all EHW for H ranging over the closed linear span of the variables b^*W in $L_2(\nu, M)$ with inner product

$$\langle EW(\cdot)H_1; EW(\cdot)H_2\rangle_{\mathbb{H}} = EH_1H_2. \tag{5.8}$$

The RKHS can be viewed as a subset of $\mathbb{B}$ and the RKHS norm $\|\cdot\|_{\mathbb{H}}$ is stronger than the Banach space norm $\|\cdot\|$.

In particular, if W is a Borel measurable zero-mean Gaussian random element in a complete separable subspace of $\ell^\infty(T)$, the Banach space of uniformly bounded functions $g : T \to \mathbb{R}$ equipped with the uniform norm $\|g\| = \sup\{|g(t)| : t \in T\}$, then the RKHS is actually the completion of the linear space of functions $t \mapsto EW(t)H$ relative to the inner product (5.8) where H, H_1 and H_2 are finite linear combinations of the form $\sum_i a_i W(s_i)$ with $a_i \in \mathbb{R}$ and s_i in the index set of W. See Theorem 2.1 of [47] for details.

Next we turn to constructing the sieve. Assume $\epsilon > 0$ is given. Let $\mathbb{H}_1^a$ denote a unit ball in the RKHS of the covariance kernel $\tau^2 e^{-a\|\mathbf{x}-\mathbf{x}'\|^2}$ and $\mathbb{B}_1$ is a unit ball in $\mathbb{C}[0,1]^p$. For numbers M, m, r, δ, construct a sequence of subsets $\{B_h, h = 1, \ldots, m\}$ of $\mathbb{C}[0,1]^p$ as follows.

$$B_h = \begin{cases} \left(M\sqrt{r/\delta}\mathbb{H}_1^r + \dfrac{\epsilon}{m^2}\mathbb{B}_1\right) \cup \left(\cup_{a<\delta} M\mathbb{H}_1^a + \dfrac{\epsilon}{m^2}\mathbb{B}_1\right), & \text{if } h = 1, \ldots, m^\eta \\ \cup_{a<\delta_n} M_n\mathbb{H}_1^a + \dfrac{\epsilon}{m^2}\mathbb{B}_1, & \text{if } h = m^\eta + 1, \ldots, m. \end{cases}$$

The idea is to construct

$$\mathcal{F}_n = \left\{ f : f(y \mid \mathbf{x}) = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \frac{1}{\sigma_h}\phi\left(\frac{y - \mathbf{x}'\boldsymbol{\beta}_h}{\sigma_h}\right), \{\phi_{\boldsymbol{\beta}_h,\sigma_h}\}_{h=1}^{m_n} \in \Theta_{a_n,h_n,l_n}, \right.$$
$$\left. \alpha_h \in B_{h,n}, h = 1, \ldots, m_n, \sup_{\mathbf{x}\in\mathcal{X}} \sum_{h\geq m_n+1} \pi_h(\mathbf{x}) \leq \epsilon \right\} \tag{5.9}$$

for appropriate sequences $a_m, l_n, h_n, M_n, m_n, r_n, \delta_n$ to be chosen in the proof of Theorem 5.10.

The following lemma is also crucial to the proof of Theorem 5.10 which allows us to calculate the rate of decay of $P(\sup_{\mathbf{x}\in\mathcal{X}} \pi_h(\mathbf{x}) > \epsilon)$ with m_n.

Lemma 5.15. *Let π_h's satisfy (5.1) with $\alpha_h \sim GP(0, c_h)$ where $c_h(\mathbf{x}, \mathbf{x}') = \tau^2 e^{-A_h\|\mathbf{x}-\mathbf{x}'\|^2}$, $h \geq 1$, $\tau^2 > 0$ fixed. Then for some constant $C_7 > 0$,*

$$\Pi_{\mathcal{X}}\left(\left\|\sum_{h=m_n+1}^{\infty} \pi_h\right\|_\infty > \epsilon\right) \leq e^{-C_7 m_n \log m_n} + \sum_{h=m_n^\eta+1}^{m_n} P(A_h > \delta_n). \tag{5.10}$$

Proof. Let $W_h = -\log[1 - \Phi\{\alpha_h'\}]$ where $\alpha_h' = \inf_{\mathbf{x}\in\mathcal{X}} \alpha_h(\mathbf{x})$, $Z_h \sim \text{Ga}(1, \gamma_0)$. We will choose an appropriate value for γ_0 in the sequel. Let $t_0 = -\log\epsilon > 0$. Observe that

$$\Pi_{\mathcal{X}}\left(\left\|\sum_{h=m_n+1}^{\infty} \pi_h\right\|_\infty > \epsilon\right) = \Pi_{\mathcal{X}}\left(-\sum_{h=m_n^\eta+1}^{m_n} \log\{1 - \Phi(\alpha_h')\} < t_0\right).$$

Observe that $\Pi_{\mathcal{X}}\left(-\sum_{h=1}^{m_n} \log\{1 - \Phi(\alpha_h)\} < t_0\right) = \Pi_{\mathcal{X}}(\Lambda_h < t_0)$ where $\Lambda_h \sim \text{Ga}(m_n, 1)$. Then it is easy to show that $\Pi_{\mathcal{X}}(\Lambda_h < t_0) \precsim e^{-m_n \log m_n}$. However, the calculation gets complicated when α_h's are i.i.d. realizations of a zero mean Gaussian process. The proof relies on the fact that the supremum of Gaussian processes has sub-Gaussian tails.

Below we calculate the rate of decay of $\Pi_{\mathcal{X}}\left(\left\|\sum_{h=m_n+1}^{\infty} \pi_h\right\|_\infty > \epsilon\right)$ with m_n. We will show that there exists γ_0, depending on ϵ and τ but not depending on n, such that

$$\Pi_{\mathcal{X}}\left(\sum_{h=m_n^\eta+1}^{m_n} W_h < t_0\right) \leq \xi(\delta_n)^{m_n - m_n^\eta} \Pi_{\mathcal{X}}\left(\sum_{h=m_n^\eta+1}^{m_n} Z_h < t_0\right) + \sum_{h=m_n^\eta+1}^{m_n} P(A_h > \delta_n) \tag{5.11}$$

where there exists a constant $C_5 > 0$ such that $\xi(x) = C_5 x^{p/2}$ for $x > 0$. Observe that $\Pi_{\mathcal{X}}\left(\sum_{h=m_n^\eta+1}^{m_n} W_h < t_0\right) \leq \Pi_{\mathcal{X}}\left(\sum_{h=m_n^\eta+1}^{m_n} W_h < t_0, A_h \leq \delta_n, h = m_n^\eta + 1, \ldots, m_n\right) + \sum_{h=m_n^\eta+1}^{m_n} P(A_h > \delta_n)$.

Since $\Pi_{\mathcal{X}}\left(\sum_{h=m_n^\eta+1}^{m_n} W_h < t_0\right) = \Pi_{\mathcal{X}}\left(\sum_{h=m_n^\eta+1}^{m_n} (\tau'/\tau) W_h < \tau' t_0/\tau\right)$ for some $\tau' < 1$, we can re-parameterize t_0 as $\tau' t_0/\tau$ and τ as τ'. Hence without loss of generality we assume $\tau < 1$.

Define $g : [0, t_0] \to \Re$, $t \mapsto -\Phi^{-1}(1 - e^{-t})$. It holds that g is a continuous function on $(0, t_0]$. Assume $\alpha_0 \sim \text{GP}(0, c_0)$ where $c_0(\mathbf{x}, \mathbf{x}') = \tau^2 e^{-\|\mathbf{x}-\mathbf{x}'\|^2}$. For $h = m_n^\eta + 1, \ldots, m_n$,

$$P\left(\sup_{\mathbf{x}\in\mathcal{X}} \alpha_h(\mathbf{x}) \geq \lambda, A_h \leq \delta_n\right) \leq P\left(\sup_{\mathbf{x}\in\sqrt{\delta_n}\mathcal{X}} \alpha_0(\mathbf{x}) \geq \lambda\right).$$

Below we estimate $P(\sup_{\mathbf{x}\in\sqrt{\delta_n}\mathcal{X}} \alpha_0(\mathbf{x}) \geq \lambda)$ for large enough λ following Theorem 5.2 of [1]. However extra care is required to identify the role of δ_n. Since $N(\epsilon, \sqrt{\delta_n}\mathcal{X}, \|\cdot\|) \leq C_1(\sqrt{\delta_n}/\epsilon)^p$,

$$\int_0^\epsilon \{\log N(\epsilon, \sqrt{\delta_n}\mathcal{X}, \|\cdot\|)\}^{1/2} d\epsilon \leq C_2\epsilon\{1+\sqrt{\log(1/\epsilon)}\}$$

for some constant $C_2 > 0$. Hence

$$\begin{aligned} P\left(\sup_{\mathbf{x}\in r_n\mathcal{X}} \alpha_0(\mathbf{x}) \geq \lambda\right) &\leq C_3(\sqrt{\delta_n}\lambda)^p \exp[-1/2\{\lambda - C_2/\lambda(1+\sqrt{\log\lambda})\}^2/\tau^2] \\ &\leq C_3\delta_n^{p/2}\lambda^{p+2}\{1-\Phi(\lambda/\tau^2)\} \leq C_4\delta_n^{p/2}\{1-\Phi(\lambda)\} \end{aligned}$$

for constants $C_3, C_4 > 0$. The last inequality holds for all large λ because $\tau < 1$. Hence there exists $t_1 \in (0, t_0)$ sufficiently small and independent of n such that for all $t \in (0, t_1)$, $\Pi_{\mathcal{X}}\{\sup_{\mathbf{x}\in\sqrt{\delta_n}\mathcal{X}} \alpha_0(\mathbf{x}) \geq g(t)\} \leq C_4\delta_n^{p/2}\Phi\{-g(t)\}$. Observe that

$$\Pi_{\mathcal{X}}\left\{\sup_{\mathbf{x}\in\sqrt{\delta_n}\mathcal{X}} \alpha_0(\mathbf{x}) \geq g(t)\right\} \leq C_4\delta_n^{p/2}\Phi\{-g(t)\} < C_5\delta_n^{p/2}(1-e^{-\gamma_0 t}),$$

for any $\gamma_0 > 1$. Further choose γ_0 large enough such that $2(1-e^{-\gamma_0 t}) > 1\ \forall\ t \in [t_1, t_0]$. Hence $P(W_h \leq t, A_h \leq \delta_n) \leq \xi(\delta_n)P(Z_h < t)\ \forall\ t \in (0, t_0]$ where $\xi(\delta_n) = C_5\delta_n^{p/2}$, with $C_5 = \max\{2, C_4\}$. Applying Lemma E.1, we conclude (5.11) by induction. Lemma E.1 is proved in Appendix E. As $\sum_{h=1}^{m_n} Z_h \sim \text{Ga}(m_n, \gamma_0)$, $\Pi_{\mathcal{X}}\left(\sum_{h=1}^{m_n} Z_h < t_0\right) \leq e^{-C_6 m_n \log m_n}$ for some constant $C_6 > 0$. Since $\xi(\delta_n)^{m_n - m_n^\eta}\Pi_{\mathcal{X}}\left(\sum_{h=1}^{m_n} Z_h < t_0\right) \leq (e^{-C_7 m_n \log m_n})$ for some constant $C_7 > 0$, the result follows immediately. □

5.3. Prior specification and posterior computation

To illustrate the applicability of the proposed methods, we mention the prior choices and key steps for posterior computation for the $\text{MGLR}_{\mathbf{x}}$ model. Recall that

$$f(y \mid \mathbf{x}) = \int \frac{1}{\sigma}\phi\left(\frac{y-\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma), \tag{5.12}$$

$$G_{\mathbf{x}} = \sum_{h=1}^{\infty} \pi_h(\mathbf{x})\delta_{(\boldsymbol{\beta}_h, \sigma_h^2)}, \qquad (\boldsymbol{\beta}_h, \sigma_h^{-2}) \sim N(\boldsymbol{\beta}_0, \Sigma_0) \times \text{Ga}(\alpha_\sigma, \beta_\sigma), \tag{5.13}$$

where $\pi_h(\mathbf{x}) = \Phi\{\alpha_h(\mathbf{x})\}\prod_{l<h}\{1-\Phi\{\alpha_l(\mathbf{x})\}\}$. We assume $\alpha_h \sim GP(0, c_h)$, where $c_h(\mathbf{x}, \mathbf{x}') = \frac{1}{\tau_\alpha}e^{-A_h\|\mathbf{x}-\mathbf{x}'\|^2}$, $\tau_\alpha \sim \text{Ga}(\nu_\alpha/2, \nu_\alpha/2)$. See Remark 5.12 for constructing prior for A_h's. If the y_i's are standardized, we would expect that the total variance $\sum_{h=1}^{\infty}\pi_h\sigma_h^2$ should be around 1. Hence choose $a_\sigma = 1$, $b_\sigma = 10$ so that the $E(\sigma_h^{-2}) = 0.1$. We can resort to an MCMC algorithm, which is a hybrid of data augmentation, the exact block Gibbs sampler of [32] and Metropolis Hastings sampling to sample from the posterior of (5.12). [32] proposed the exact block Gibbs sampler as an efficient approach to posterior computation in infinite-dimensional Dirichlet process mixture models, modifying the block Gibbs sampler of [21] to avoid truncation approximations. The exact block Gibbs sampler combines characteristics of the retrospective sampler [34] and the slice sampler [49,24]. Introduce $\gamma_1, \ldots, \gamma_n$ such that $\pi_h(\mathbf{x}_i) = P(\gamma_i = h)$, $h = 1, 2, \ldots, \infty$. Then

$$\gamma_i \sim \sum_{h=1}^{\infty}\pi_h(\mathbf{x}_i)\delta_h = \sum_{h=1}^{\infty} 1(u_i < \pi_h(\mathbf{x}_i))\delta_h$$

where $u_i \sim U(0, 1)$.

We continue up to $h = 1, \ldots, h^* = \max\{h_1^*, \ldots, h_n^*\}$, where h_i^* is the minimum integer satisfying $\sum_{l=1}^{h_i^*}\pi_l(\mathbf{x}_i) > 1 - \min\{u_1, \ldots, u_n\}$, $i = 1, \ldots, n$. The Markov chain adaptively estimates the desired number of components h^* at each iteration of the MCMC, thus making it more efficient than a finite mixture model with a pre-specified large number of components. Here we describe the key steps for the posterior computation.

1. *Update u_i's and stick breaking random variables*: Generate

$$u_i|- \sim U(0, \pi_{\gamma_i}(\mathbf{x}_i))$$

where $\pi_h(\mathbf{x}_i) = \Phi\{\alpha_h(\mathbf{x}_i)\}\prod_{l<h}[1-\Phi\{\alpha_l(\mathbf{x}_i)\}]$. For $i = 1, \ldots, n$, introduce latent variables $Z_h(\mathbf{x}_i)$, $h = 1, 2, \ldots$ such that $Z_h(\mathbf{x}_i) \sim \text{N}(\alpha_h(\mathbf{x}_i), 1)$. Thus $\pi_h(\mathbf{x}_i) = P(Z_h(\mathbf{x}_i) > 0, Z_l(\mathbf{x}_i) < 0 \text{ for } l < h)$. Then

$$Z_h(\mathbf{x}_i)|- \sim \begin{cases} \text{N}(\alpha_h(\mathbf{x}_i), 1)I_{\mathbb{R}^+}, & h = \gamma_i \\ \text{N}(\alpha_h(\mathbf{x}_i), 1)I_{\mathbb{R}^-}, & h < \gamma_i. \end{cases}$$

Let $\mathbf{Z}_h = (Z_h(\mathbf{x}_1), \ldots, Z_h(\mathbf{x}_n))'$ and $\boldsymbol{\alpha}_h = (\alpha_h(\mathbf{x}_1), \ldots, \alpha_h(\mathbf{x}_n))'$. Letting $\left(\boldsymbol{\Sigma}_h\right)_{ij} = e^{-A_h\|\mathbf{x}_i - \mathbf{x}_j\|}$, $\mathbf{Z}_h \sim \mathrm{N}(\boldsymbol{\alpha}_h, \mathbf{I})$ and $\boldsymbol{\alpha}_h \sim \mathrm{N}(0, \frac{1}{\tau_\alpha}\boldsymbol{\Sigma}_h)$,

$$\boldsymbol{\alpha}_h|- \sim \mathrm{N}\big((\tau_\alpha \boldsymbol{\Sigma}_h^{-1} + \mathbf{I}_n)^{-1}\mathbf{Z}_h, (\tau_\alpha \boldsymbol{\Sigma}_h^{-1} + \mathbf{I}_n)^{-1}\big).$$

Continue up to $h = 1, \ldots, h^* = \max\{h_1^*, \ldots, h_n^*\}$, where h_i^* is the minimum integer satisfying $\sum_{l=1}^{h_i^*} \pi_l(\mathbf{x}_i) > 1 - \min\{u_1, \ldots, u_n\}$, $i = 1, \ldots, n$. Now

$$\tau_\alpha|- \sim \mathrm{Ga}\left(\frac{1}{2}\big(nh^* + \nu_\alpha\big), \frac{1}{2}\left(\sum_{l=1}^{h^*} \boldsymbol{\alpha}_k' \boldsymbol{\Sigma}_l^{-1} \boldsymbol{\alpha}_k + \nu_\alpha\right)\right),$$

while κ_α is updated using a Metropolis Hastings step.

2. *Update allocation to atoms*: Update $(\gamma_1, \ldots, \gamma_n)|$-as multinomial random variables with probabilities

$$P(\gamma_i = h) \propto \mathrm{N}(y_i; \mathbf{x}_i'\boldsymbol{\beta}_h, \tau_h^{-1}) I(u_i < \pi_h(\mathbf{x}_i)), \quad h = 1, \ldots, h^*.$$

3. *Update component-specific locations and precisions*: Let $n_h = \#\{i : \gamma_i = h\}$, $l = 1, 2, \ldots, h^*$. Let $Y_h = (y_i : \gamma_i = h)$ be a n_h dimensional vector and $\mathbf{X}_h$ is the corresponding $n_h \times p$ dimensional covariate matrix.

$$\boldsymbol{\beta}_h|- \sim \mathrm{N}\big((\mathbf{X}_h'\mathbf{X}_h + \Sigma_0^{-1})^{-1}(\mathbf{X}_h'Y_h + \Sigma_0^{-1}\boldsymbol{\beta}_0), (\mathbf{X}_h'\mathbf{X}_h + \Sigma_0^{-1})^{-1}\big)$$

$$\tau_h|- \sim \mathrm{Ga}\left(\frac{n_h}{2} + \alpha_\tau, \beta_\tau + \sum_{i:\gamma_i = h} (y_i - \mathbf{x}_i'\boldsymbol{\beta}_h)^2\right), \quad h = 1, 2, \ldots, h^*.$$

Update A_h's in a Metropolis Hastings step.

At each iteration of the MCMC, we obtain samples from the full conditional distributions of the parameters, which after discarding a burn-in can be used to get summary statistics of posterior distribution of the parameters or functionals of interest.

6. Posterior consistency in Gaussian mixture of fixed-π dependent processes

6.1. Kullback–Leibler property

The following theorem verifies that $\Pi_{\mathcal{X}}$ has KL property at $f_0 \in \mathcal{F}_d^*$. The proof of Theorem 6.1 is somewhat similar to that of Theorem 5.3 and can be found in Appendix F.

Theorem 6.1. *$f_0 \in \mathrm{KL}(\Pi_{\mathcal{X}})$ for each f_0 in $\mathcal{F}_d^*$ if $\mathcal{P}_{\mathcal{X}}$ satisfies*

T1. *G_0 is specified by $\mu_h \sim \mathrm{GP}(\mu, c)$, $\sigma_h \sim G_{0,\sigma}$ where c is chosen so that $\mathrm{GP}(0, c)$ has continuous path realizations and Π_σ is absolutely continuous w.r.t. Lebesgue measure on $\Re^+$.*

T2. *For every $k \geq 2$, $(\pi_1, \ldots, \pi_k)$ is absolutely continuous w.r.t. to the Lebesgue measure on S_{k-1}.*

T3. *For any continuous function $g : \mathcal{X} \mapsto \Re$,*

$$\mathcal{P}_{\mathcal{X}}\left\{\sup_{\mathbf{x}\in\mathcal{X}} |\mu_h(\mathbf{x}) - g(\mathbf{x})| < \epsilon\right\} > 0$$

$h = 1, \ldots, \infty$ and for any $\epsilon > 0$.

6.2. Strong consistency with the q-integrated L_1 neighborhood

Next we summarize the consistency theorem with respect to the q-integrated L_1 topology. The proof of Theorem 6.2 is also similar to that of Theorem 5.10 and is provided in Appendix G.

Theorem 6.2. *Let $\mu_h(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}_h + \eta_h(\mathbf{x})$, $\boldsymbol{\beta}_h \sim G_{\boldsymbol{\beta}}$ and $\eta_h \sim GP(0, c)$, $h = 1, \ldots, \infty$ where $c(\mathbf{x}, \mathbf{x}') = \tau^2 e^{-A\|\mathbf{x}-\mathbf{x}'\|^2}$, $A^{p(1+\eta_2)/\eta_2} \sim \mathrm{Ga}(a, b)$ for some $\eta_2 > 0$.*

F1. *There exist sequences $a_n, h_n \uparrow \infty, l_n \downarrow 0$ with $\frac{a_n}{l_n} = O(n)$, $\frac{h_n}{l_n} = O(e^n)$, and constants d_1, d_2, d_3 and $d_4 > 0$ such that $G_{\boldsymbol{\beta}}\{B(0; a_n)\}^c < d_1 e^{-d_2 n}$ and $G_{0,\sigma}\{[l_n, h_n]\}^c \leq d_3 e^{-d_4 n}$.*

F2. *$P(\sum_{h=n}^{\infty} \pi_h > \epsilon) \precsim O(e^{-n^{1+\eta_2}(\log n)^{(p+1)}})$.*

Then $f_0 \in \mathrm{KL}(\Pi_{\mathcal{X}})$ implies that $\Pi_{\mathcal{X}}$ achieves strong posterior consistency at f_0 with respect to the q-integrated L_1 topology.

Remark 6.3. F2 is satisfied if π_h's are made to decay more rapidly than the usual $\mathrm{Beta}(1, \alpha)$ stick-breaking random variables, e.g., if $\pi_h = \nu_h \prod_{l<h}(1 - \nu_h)$ and if $\nu_h \sim \mathrm{Beta}(1, \alpha_h)$ where $\alpha_h = h^{1+\eta_2}(\log h)^{p+1}\alpha_0$ for some $\alpha_0 > 0$, then F2 is satisfied. Large value of α_h for the higher indexed weights favors smaller number of components.

Remark 6.4. A Gaussian kernel is used here for technical simplification. One can obtain similar results using a variety of kernels e.g. t, Laplace, etc. However, the KL support conditions A1–A5 will be different for different kernels. Refer to [50] for a catalog of conditions for various kernels in a density estimation framework.

7. Discussion

We have provided sufficient conditions to show posterior consistency in estimating the conditional density via predictor dependent mixtures of Gaussians which include probit stick-breaking mixtures of Gaussians and the fixed-π dependent processes as special cases. The problem is of interest, providing a more flexible and informative alternative to the usual mean regression. For both the models, we need the same set of tail conditions (mentioned in $\mathcal{F}_d^*$) on f_0 for KL support. Although the first prior is flexible in the weights and the second one in the atoms through their corresponding GP terms, S1, S2, T1 and T3 show that verification of KL property only requires that both the GP terms have continuous path realizations and desired approximation property. Moreover, for the second prior, any set of weights summing to one a.s. T2 suffices for showing KL property. Careful investigations of the prior for the GP kernel for the first model and the probability weights for the second one are required for strong consistency. For the first one we need the covariate dependence of the higher indexed GP terms in the weights to fade off. On the other hand, for the second model, the atoms can be i.i.d. realizations of a GP with Gaussian covariance kernel with inverse-Gamma bandwidth while limiting the model complexity through a sequence of probability weights which are allowed to decay rapidly. This suggests that full flexibility in the weights should be down-weighted by an appropriately chosen prior while full flexibility in the atoms should be accompanied by a restriction imposing fewer number of components. It would be interesting to see how the conditions on the bandwidth can be modified when we actually use a sieve Bayes prior, i.e. a prior with number of components k_n diverging to ∞.

Another interesting direction is to consider rates of convergence of the posterior and Bernstein von-Mises (BvM) type results. For infinite dimensional parameters [14], there has been quite a few positive BvM results very recently for linear functionals of a probability density function [36] and for general classes of linear and non-linear functionals in a Gaussian white noise model [5]. We conjecture that such BvM-type results hold for linear functionals of conditional density (e.g. conditional mean, conditional cdf) too under appropriate conditions on the prior and the true data generating conditional density.

Acknowledgments

This work was supported by Award Number R01ES017240 from the National Institute of Environmental Health Sciences. We also thank the Associate Editor and the referees for the comments which significantly improved the exposition of the paper.

Appendix A. A useful lemma

To prove Theorem 5.6, we need an auxiliary lemma which we state below.

Lemma A.1. *If* $\{\pi_h(\mathbf{x}), h = 1, \ldots, \infty\}$ *constructed as in* (5.1) *satisfies* S1 *and* S2 *then*

$$\mathcal{P}_{\mathcal{X}}\left\{\sup_{\mathbf{x}\in\mathcal{X}}|\pi_1(\mathbf{x}) - F_{\mathbf{x}}(A_1)| < \epsilon_1, \ldots, \sup_{\mathbf{x}\in\mathcal{X}}|\pi_k(\mathbf{x}) - F_{\mathbf{x}}(A_k)| < \epsilon_k\right\} > 0 \tag{A.1}$$

for a measurable partition $\{A_i, i = 1, \ldots, k\}$ *of* $\Re^p \times \Re^+$, $\epsilon_i > 0$ *and a conditional cdf* $\{F_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$.

Proof. Without loss of generality, let $0 < F_{\mathbf{x}}(A_i) < 1, i = 1, \ldots, k\ \forall\, \mathbf{x} \in \mathcal{X}$. We want to show that for any $\epsilon_i > 0, i = 1, \ldots, k$, (A.1) holds. Construct continuous functions $g_i : \mathcal{X} \mapsto \Re$, $0 < g_i(\mathbf{x}) < 1\ \forall \mathbf{x} \in \mathcal{X}, i = 1, \ldots, k-1$ such that

$$g_1(\mathbf{x}) = F_{\mathbf{x}}(A_1), \qquad g_i(\mathbf{x})\prod_{l<i}\{1 - g_l(\mathbf{x})\} = F_{\mathbf{x}}(A_i), \quad 2 \le i \le k-1, \qquad g_k(\mathbf{x}) = 1 \quad \forall \mathbf{x}. \tag{A.2}$$

As $0 < F_{\mathbf{x}}(A_i) < 1, i = 1, \ldots, k\ \forall\, \mathbf{x} \in \mathcal{X}$, it is trivial to find $g_i, i = 1, \ldots, k$ satisfying (A.2) since one can solve back for the g_i's from (A.2). $\sum_{i=1}^k F_{\mathbf{x}}(A_i) = 1$ enforces $g_k \equiv 1$. Since Φ is a continuous function, for any $\epsilon_i > 0, i = 1, \ldots, k-1$,

$$\mathcal{P}_{\mathcal{X}}\left\{\sup_{\mathbf{x}\in\mathcal{X}}|\Phi\{\alpha_i(\mathbf{x})\} - g_i(\mathbf{x})| < \epsilon_i\right\} > 0 \tag{A.3}$$

and for $i = k$,

$$\mathcal{P}_{\mathcal{X}}\left\{\sup_{\mathbf{x}\in\mathcal{X}}|\Phi\{\alpha_k(\mathbf{x})\} - 1| < \epsilon_k\right\} = \mathcal{P}_{\mathcal{X}}\left\{\inf_{\mathbf{x}\in\mathcal{X}}\alpha_k(\mathbf{x}) > \Phi^{-1}(1-\epsilon_k)\right\}. \tag{A.4}$$

Choose $M > \Phi^{-1}(1-\epsilon_k) + \epsilon_k$. We have $0 < M < 1$ and

$$\left\{\sup_{\mathbf{x}\in\mathcal{X}}|\alpha_k(\mathbf{x}) - M| < \epsilon_k\right\} \subset \left\{\inf_{\mathbf{x}\in\mathcal{X}}\alpha_k(\mathbf{x}) > \Phi^{-1}(1-\epsilon_k)\right\}.$$

Hence by assumption, $\mathcal{P}_{\mathcal{X}}\left\{\inf_{\mathbf{x}\in\mathcal{X}} \alpha_k(\mathbf{x}) > \Phi^{-1}(1-\epsilon_k)\right\} > 0$. Let S_{k-1} denote the k-dimensional simplex. For notational simplicity let $p_i(\mathbf{x}) = \Phi\{\alpha_i(\mathbf{x})\}$, $g_i(\mathbf{x}) = F_{\mathbf{x}}(A_i)$, $i = 1, \ldots, k-1$ and $g_k(\mathbf{x}) = 1$. Let $\boldsymbol{z} = (z_1, \ldots, z_p)'$, $f_i : S_{k-1} \to \Re$, $\boldsymbol{z} \mapsto z_i \prod_{l<i}(1-z_l)$, $i = 2, \ldots, k$ and $f_1(\boldsymbol{z}) = z_1$. Let $\boldsymbol{p}(\mathbf{x}) = (p_1(\mathbf{x}), \ldots, p_k(\mathbf{x}))$ and $\boldsymbol{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_k(\mathbf{x}))$. Then we need to show that

$$\mathcal{P}_{\mathcal{X}}\left\{\|f_1(\boldsymbol{p}) - f_1(\boldsymbol{g})\|_\infty < \epsilon_1, \ldots, \|f_{k-1}(\boldsymbol{p}) - f_{k-1}(\boldsymbol{g})\|_\infty < \epsilon_{k-1}, \|f_k(\boldsymbol{p}) - 1\|_\infty < \epsilon_k\right\} > 0.$$

Note that for $2 \le i \le k$,

$$\|f_i(\boldsymbol{p}) - f_i(\boldsymbol{g})\|_\infty \le (i-1)\,\|p_i - g_i\|_\infty + \sum_{l<i} \|f_l(\boldsymbol{p}) - f_l(\boldsymbol{g})\|_\infty .$$

Thus one can get $\epsilon_i^* > 0$, $i = 1, \ldots, k$, such that

$$\left\{\|p_i - g_i\|_\infty < \epsilon_i^*, i = 1, \ldots, k\right\} \subset \left\{\|f_1(\boldsymbol{p}) - f_1(\boldsymbol{g})\|_\infty < \epsilon_1, \ldots, \|f_{k-1}(\boldsymbol{p}) - f_{k-1}(\boldsymbol{g})\|_\infty < \epsilon_{k-1}, \|f_k(\boldsymbol{p}) - 1\|_\infty < \epsilon_k\right\}.$$

But since $\mathcal{P}_{\mathcal{X}}\{\|p_i - g_i\|_\infty < \epsilon_i^*, i = 1, \ldots, k\} = \prod_{i=1}^k \mathcal{P}_{\mathcal{X}}\{\|p_i - g_i\|_\infty < \epsilon_i^*\}$, the result follows immediately. □

Appendix B. Proof of Theorem 5.6

Fix $\{F_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} \in \mathcal{G}_{\mathcal{X}}^*$. Without loss of generality it is enough to show that for a uniformly continuous function $g : \Re^p \times \Re^+ \times \mathcal{X} \to [0,1]$ and $\epsilon > 0$,

$$\mathcal{P}_{\mathcal{X}}\left\{\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} : \left|\int_{\Re^p\times\Re^+\times\mathcal{X}} \left\{g(\boldsymbol{\beta}, \sigma, \mathbf{x})dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)q(\mathbf{x})d\mathbf{x} - g(\boldsymbol{\beta}, \sigma, \mathbf{x})dF_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)q(\mathbf{x})d\mathbf{x}\right\}\right| < \epsilon\right\} > 0.$$

Furthermore, it suffices to assume $g(\boldsymbol{\beta}, \sigma, \mathbf{x}) \to 0$ uniformly in $\mathbf{x} \in \mathcal{X}$ as $\|\boldsymbol{\beta}\| \to \infty, \sigma \to \infty$.

Fix $\epsilon > 0$, there exist $a, \underline{\sigma}, \overline{\sigma} > 0$ not depending on $\mathbf{x}$ such that $F_{\mathbf{x}}([-a,a]^p \times [\underline{\sigma}, \overline{\sigma}]) > 1 - \epsilon$ for all $\mathbf{x} \in \mathcal{X}$. Let $C = [-a,a]^p \times [\underline{\sigma}, \overline{\sigma}]$.

$$\begin{aligned}
&\int_{\Re^p\times\Re^+\mathcal{X}} \left\{g(\boldsymbol{\beta}, \sigma, \mathbf{x})dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma) - g(\boldsymbol{\beta}, \sigma, \mathbf{x})dF_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)\right\}q(\mathbf{x})d\mathbf{x} \\
&\le \int_{\mathcal{X}} \left\{\sum_{h=1}^\infty \pi_h(\mathbf{x})g(\boldsymbol{\beta}_h, \sigma_h, \mathbf{x}) - \int_C g(\boldsymbol{\beta}, \sigma, \mathbf{x})dF_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)\right\} q(\mathbf{x})d\mathbf{x} + \epsilon
\end{aligned}$$

where π_h's are specified by (5.1) with c_h satisfying S1 and S2 and $(\boldsymbol{\beta}_h, \sigma_h) \sim G_0$. Now for each $\mathbf{x} \in \mathcal{X}$, construct a Riemann sum approximation of $\int_C g(\boldsymbol{\beta}, \sigma, \mathbf{x})dF_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)$.

Let $\{A_{k,n}, k = 1, \ldots, n\}$ be a sequence of partitions of C with increasing refinement as n increases. Assume $\max_{1\le k\le n} \operatorname{diam}(A_{k,n}) \to 0$ as $n \uparrow \infty$. Fix $(\tilde{\boldsymbol{\beta}}_{k,n}, \tilde{\sigma}_{k,n}) \in A_{k,n}$, $k = 1, \ldots, n$. Then by DCT as $n \to \infty$,

$$\int_{\mathcal{X}} \left\{\sum_{k=1}^n g(\tilde{\boldsymbol{\beta}}_{k,n}, \tilde{\sigma}_{k,n}, \mathbf{x})F_{\mathbf{x}}(A_{k,n})\right\} q(\mathbf{x})d\mathbf{x} \to \int_{\mathcal{X}}\int_C g(\boldsymbol{\beta}, \sigma, \mathbf{x})dF_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)q(\mathbf{x})d\mathbf{x}. \tag{B.1}$$

Hence there exists n_1 such that for $n \ge n_1$

$$\begin{aligned}
&\left|\int_{\Re^p\times\Re^+\mathcal{X}} \left\{g(\boldsymbol{\beta}, \sigma, \mathbf{x})dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma) - g(\boldsymbol{\beta}, \sigma, \mathbf{x})dF_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)\right\}q(\mathbf{x})d\mathbf{x}\right| \\
&\le \left|\int_{\mathcal{X}} \left\{\sum_{h=1}^\infty \pi_h(\mathbf{x})g(\boldsymbol{\beta}_h, \sigma_h, \mathbf{x}) - \sum_{k=1}^n g(\tilde{\boldsymbol{\beta}}_{k,n}, \tilde{\sigma}_{k,n}, \mathbf{x})F_{\mathbf{x}}(A_{k,n})\right\} q(\mathbf{x})d\mathbf{x}\right| + 2\epsilon.
\end{aligned}$$

Consider the set

$$\Omega_1 = \left\{(\pi_h, h = 1, \ldots, \infty) : \sup_{\mathbf{x}\in\mathcal{X}} \left|\pi_1(\mathbf{x}) - F_{\mathbf{x}}(A_{1,n_1})\right| < \frac{\epsilon}{n_1}, \ldots, \sup_{\mathbf{x}\in\mathcal{X}} \left|\pi_{n_1}(\mathbf{x}) - F_{\mathbf{x}}(A_{n_1,n_1})\right| < \frac{\epsilon}{n_1}\right\}.$$

By Lemma A.1 which is proved in Appendix A, $\mathcal{P}_{\mathcal{X}}(\Omega_1) > 0$. Since $\sum_{h=1}^\infty \pi_h(\mathbf{x}) = 1$ a.s. there $\exists\ \Omega$ with $\mathcal{P}_{\mathcal{X}}(\Omega) = 1$, such that for each $\omega = \{\pi_h, h = 1, \ldots, \infty\} \in \Omega$, $g_n(\mathbf{x}) = \sum_{h=1}^n \pi_h(\mathbf{x}) \to 1$ as $n \to \infty$ for each $\mathbf{x}$ in $\mathcal{X}$. Note that this convergence is uniform since, $g_n(\cdot)$, $n \ge 1$ are continuous functions defined on a compact set monotonically increasing to a continuous function identically equal to 1. Hence for each $\omega = \{\pi_h, h = 1, \ldots, \infty\} \in \Omega$, $g_n(\mathbf{x}) \to 1$ uniformly in $\mathbf{x}$. By Egoroff's theorem, there exists a measurable subset Ω_2 of Ω_1 with $\mathcal{P}_{\mathcal{X}}(\Omega_2) > 0$ such that within this subset $g_n(\mathbf{x}) \to 1$ uniformly in $\mathbf{x}$ and uniformly in ω in Ω_2. Thus there exists a positive integer $n_\epsilon \ge n_1$ not depending on $\mathbf{x}$ and ω, such that

$\sum_{h=n_\epsilon+1}^{\infty}\pi_h(\mathbf{x}) < \epsilon$ on Ω_2. Moreover, one can find a $K > 0$ independent of $\mathbf{x}$ such that $g(\beta,\sigma,\mathbf{x}) < \epsilon$ if $\|\boldsymbol{\beta}\| > K$ and $\sigma > K$. Let $A_1 = \{(\boldsymbol{\beta},\sigma) : \|\boldsymbol{\beta}\| > K, \sigma > K\}$. Let $\Omega_3 = \Omega_2 \cap \{(\boldsymbol{\beta}_{n_1+1},\sigma_{n_1+1}) \in A_1,\ldots,(\boldsymbol{\beta}_{n_\epsilon-1},\sigma_{n_\epsilon-1}) \in A_1\}$. For $\omega \in \Omega_3$,

$$\left|\int_{\Re^p\times\Re^+\times\mathcal{X}} \{g(\boldsymbol{\beta},\sigma,\mathbf{x})dG_{\mathbf{x}}(\boldsymbol{\beta},\sigma) - g(\boldsymbol{\beta},\sigma,\mathbf{x})dF_{\mathbf{x}}(\boldsymbol{\beta},\sigma)\}q(\mathbf{x})d\mathbf{x}\right|$$
$$\le \int_{\mathcal{X}}\left\{\sum_{k=1}^{n_1}\left|\pi_k(\mathbf{x})g(\boldsymbol{\beta}_k,\sigma_k,\mathbf{x}) - g(\tilde{\boldsymbol{\beta}}_{k,n},\tilde{\sigma}_{k,n},\mathbf{x})F_{\mathbf{x}}(A_{k,n_1})\right|\right\} q(\mathbf{x})d\mathbf{x} + 4\epsilon$$

and

$$\int_{\mathcal{X}}\left\{\sum_{k=1}^{n_1}\left|\pi_k(\mathbf{x})g(\boldsymbol{\beta}_k,\sigma_k,\mathbf{x}) - g(\tilde{\boldsymbol{\beta}}_{k,n},\tilde{\sigma}_{k,n},\mathbf{x})F_{\mathbf{x}}(A_{k,n_1})\right|\right\} q(\mathbf{x})d\mathbf{x}$$
$$\le \sum_{k=1}^{n_1}\int_{\mathcal{X}}\pi_k(\mathbf{x})\left|g(\boldsymbol{\beta}_k,\sigma_k,\mathbf{x}) - g(\tilde{\boldsymbol{\beta}}_{k,n},\tilde{\sigma}_{k,n},\mathbf{x})\right| q(\mathbf{x})d\mathbf{x} + \epsilon.$$

There exists sets B_k, $k = 1,\ldots,n_1$ depending on n_1 but independent of $\mathbf{x}$ such that if $(\boldsymbol{\beta}_k,\sigma_k) \in B_k$, $\left|g(\boldsymbol{\beta}_k,\sigma_k,\mathbf{x}) - g(\tilde{\boldsymbol{\beta}}_{k,n_1},\tilde{\sigma}_{k,n_1},\mathbf{x})\right| < \epsilon$. So for $\omega \in \Omega_4 = \Omega_3 \cap \{(\boldsymbol{\beta}_1,\sigma_1) \in B_1,\ldots,(\boldsymbol{\beta}_{n_1},\sigma_{n_1}) \in B_{n_1}\}$,

$$\left|\int_{\Re^p\times\Re^+\times\mathcal{X}} \{g(\boldsymbol{\beta},\sigma,\mathbf{x})dG_{\mathbf{x}}(\boldsymbol{\beta},\sigma) - g(\boldsymbol{\beta},\sigma,\mathbf{x})dF_{\mathbf{x}}(\boldsymbol{\beta},\sigma)\}q(\mathbf{x})d\mathbf{x}\right| < 5\epsilon.$$

Now since $\mathcal{P}_{\mathcal{X}}(\Omega_2) > 0$ and the sets $\{(\boldsymbol{\beta}_{n_1+1},\sigma_{n_1+1}) \in A_1,\ldots,(\boldsymbol{\beta}_{n_\epsilon-1},\sigma_{n_\epsilon-1}) \in A_1\}$ and $\{(\boldsymbol{\beta}_1,\sigma_1) \in B_1,\ldots,(\boldsymbol{\beta}_{n_1},\sigma_{n_1}) \in B_{n_1}\}$ are independent from Ω_2 and have positive probability, it follows that $\mathcal{P}_{\mathcal{X}}(\Omega_4) > 0$. □

Appendix C. Proof of Theorem 5.3

Without loss of generality, assume that the covariate space $\mathcal{X}$ is $[\zeta,1]^p$ for some $0 < \zeta < 1$. The proof is essentially along the lines of Theorem 3.2 of [43]. The $\tilde{f}$ in (5.2) will be constructed so as to satisfy the assumptions of Lemma 5.5 and such that $\int_{\mathcal{X}}\int_{\mathcal{Y}} f_0(y \mid \mathbf{x})\log\frac{f_0(y|\mathbf{x})}{\tilde{f}(y|\mathbf{x})}dyq(\mathbf{x})d\mathbf{x} < \frac{\epsilon}{2}$ for any $\epsilon > 0$. Define a sequence of conditional densities $f_n(y \mid \mathbf{x}) = \int \frac{1}{\sigma}\phi(\frac{y-\mathbf{x}'\boldsymbol{\beta}}{\sigma})d\tilde{G}_{n,\mathbf{x}}(\boldsymbol{\beta},\sigma)$, $n \ge 1$ where for $\sigma_n = n^{-\eta}$,

$$dG_{n,\mathbf{x}}(\boldsymbol{\beta},\sigma) = \frac{I_{\beta_1\in[-n,n]}f_0(\mathbf{x}'\beta \mid \mathbf{x})\prod_{j=2}^{p}\delta_0(\beta_j)\delta_{\sigma_n}(\sigma)}{\int_{-n}^{n}f_0(x_1\beta_1 \mid \mathbf{x})d\beta_1}. \tag{C.1}$$

Define

$$f_n(y \mid \mathbf{x}) = \frac{\int_{-nx_1}^{nx_1}\frac{1}{\sigma_n}\phi\left(\frac{y-t}{\sigma_n}\right)f_0(t \mid \mathbf{x})dt}{\int_{-nx_1}^{nx_1}f_0(t \mid \mathbf{x})dt}. \tag{C.2}$$

Proceeding as in Theorem 3.2 of [43], an application of DCT using the conditions A1–A5 yields

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} f_0(y \mid \mathbf{x})\log\frac{f_0(y \mid \mathbf{x})}{f_n(y \mid \mathbf{x})}dyq(\mathbf{x})d\mathbf{x} \to 0 \quad \text{as } n \to \infty.$$

Therefore one can simply choose $\tilde{f} = f_{n_0}$ for sufficiently large n_0. f_{n_0} satisfies the assumptions of Lemma 5.5 since $\{G_{n_0,\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ is compactly supported. Also $\{G_{n_0,\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} \in \mathcal{G}^*_{\mathcal{X}}$ as $\mathbf{x} \to G_{n_0,\mathbf{x}}(A)$ is continuous. Hence there exists a finite intersection W of neighborhoods of $\{G_{n_0,\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ the type (5.5) such that for any $\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} \in W$, the second term of (5.2) is arbitrarily small. The conclusion of the theorem follows immediately from Corollary 5.8. □

Appendix D. Proof of Theorem 5.10

Consider the sequence of sieves defined by (5.9) for given $\epsilon > 0$ and for sequences $a_n, h_n, l_n, M_n, m_n, r_n$ to be chosen later with $\delta_n = K_1\epsilon/(M_n m_n^2)$ for some constant K_1. We will first show that given $\xi > 0$, there exist $c_1, c_2 > 0$ and sequences m_n and M_n, such that $\Pi_{\mathcal{X}}(\mathcal{F}_n^c) \le c_1 e^{-nc_2}$ and $\log N(\delta,\mathcal{F}_n,\|\cdot\|) < n\xi$.

For $f_1, f_2 \in \mathcal{F}_n$, we have for each $\mathbf{x} \in \mathcal{X}$,

$$\|f_1(\cdot \mid \mathbf{x}) - f_2(\cdot \mid \mathbf{x})\|_1 \le \sum_{h=1}^{m_n}\left\|\pi_h^{(1)} - \pi_h^{(2)}\right\|_\infty + 2\epsilon.$$

Let $\Theta_{\pi,n} = \{\pi^{m_n} = (\pi_1, \pi_2, \ldots, \pi_{m_n}) : \alpha_h \in B_{h,n}, h = 1, \ldots, m_n\}$. Fix $\pi_1^{m_n}, \pi_2^{m_n} \in \Theta_{\pi,n}$. Note that since $|\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2)| < K_2 |\mathbf{x}_1 - \mathbf{x}_2|$ for a global constant $K_2 > 0$, we have

$$\left\|\Phi(\alpha_{h,1}) - \Phi(\alpha_{h,2})\right\|_\infty \leq K_2 \left\|\alpha_{h,1} - \alpha_{h,2}\right\|_\infty.$$

The above fact together with the proof of Lemma A.1 show that if we can make $\left\|\alpha_{h,1} - \alpha_{h,2}\right\|_\infty < \frac{\epsilon}{m_n^2}$, $h = 1, \ldots, m_n$, we would have $\sum_{h=1}^{m_n} \left\|\pi_h^{(1)} - \pi_h^{(2)}\right\|_\infty < \epsilon$. From the proof of Theorem 3.1 in [48] it follows that for $h = 1, \ldots, m_n^\eta$ and for sufficiently large M_n, r_n,

$$\log N(2\epsilon/m_n^2, B_{h,n}, \|\cdot\|_\infty) \leq K_3 r_n^p \log\left(\frac{M_n m_n^2 \sqrt{r_n/\delta_n}}{\epsilon}\right)^{p+1} + 2\log\frac{K_4 M_n m_n^2}{\epsilon} \tag{D.1}$$

for global constants $K_3, K_4 > 0$. For $M_n^2 > 16K_5 r_n^p (\log(r_n/\epsilon))^{1+p}$, $r_n > 1$ we have for $h = 1, \ldots, m_n^\eta$,

$$P(\alpha_h \notin B_{h,n}) \leq P(A_h > r_n) + e^{-M_n^2/2}. \tag{D.2}$$

Hence for sufficiently large M_n, we have for $h = m_n^\eta + 1, \ldots, m_n$,

$$\log N(3\epsilon/m_n^2, B_{h,n}, \|\cdot\|_\infty) \leq 2\log\frac{K_4 M_n m_n^2}{\epsilon}. \tag{D.3}$$

For $h = m_n^\eta + 1, \ldots, m_n$,

$$\begin{aligned} P(\alpha_h \notin B_{h,n}) &\leq P(A_h > \delta_n) + \int_{a=0}^{\delta_n} P(\alpha_h \notin B_{h,n} \mid A_h = a) g_{A_h}(a) da \\ &\leq P(A_h > \delta_n) + (1 - \Phi(\Phi^{-1}(e^{-\phi_0^{\delta_n}(\epsilon/m_n^2)}) + M_n)) \end{aligned}$$

where $\phi_0^\kappa(\epsilon)$ denotes the concentration function of the Gaussian process with covariance kernel $c(\mathbf{x}, \mathbf{x}') = \tau^2 e^{-\kappa\|\mathbf{x}-\mathbf{x}'\|^2}$. Now

$$\phi_0^{\delta_n}(\epsilon/m_n^2) \leq -\log P(|W_0| \leq \epsilon/m_n^2) = K_6 \left|\log(\epsilon/m_n^2)\right|$$

for some constant $K_6 > 0$. Hence if $M_n \geq K_7 \left|\log(\epsilon/m_n^2)\right|$ for some $K_7 > 0$, then it follows from the proof of Theorem 3.1 in [48] that

$$P(\alpha_h \notin B_{h,n}) \leq P(A_h > \delta_n) + e^{-M_n^2/2}. \tag{D.4}$$

From (D.1) and (D.3),

$$\log(N(\epsilon, B_{1,n} \times \cdots \times B_{m_n,n}, \|\cdot\|_\infty)) \leq 2m_n \log\frac{K_4 M_n m_n^2}{\epsilon} + m_n^\eta r_n^p \log\left(\frac{M_n m_n^2 \sqrt{r_n/\delta_n}}{\epsilon}\right)^{p+1}. \tag{D.5}$$

Also from (D.2) and (D.4),

$$\sum_{h=1}^{m_n} P(\alpha_h \notin B_{h,n}) \leq m_n e^{-M_n^2/2} + \sum_{h=1}^{m_n^\eta} P(A_h > r_n) + \sum_{h=m_n^\eta+1}^{m_n} P(A_h > \delta_n).$$

We will show that with $m_n = O(\frac{n}{\log n})$, $\Pi_{\mathcal{X}}(\mathcal{F}_n^c) < e^{-n\xi_0}$ for some ξ_0. By assumption C1, we have

$$\Pi_{\mathcal{X}}(\Theta_{a_n,h_n,l_n}^c) \precsim m_n O(e^{-n}) \precsim O(e^{-n}). \tag{D.6}$$

With $m_n = O(n/\log n)$, $\sum_{h=1}^{m_n^\eta} P(A_h > r_n) \leq m_n^\eta e^{-n} \precsim e^{-n}$, $\sum_{h=m_n^\eta+1}^{m_n} P(A_h > \delta_n) \leq (m_n - m_n^\eta) e^{-n^{-\eta_0} m_n^{\eta_0+2} \log m_n} \precsim e^{-m_n \log m_n}$.

With $m_n = \frac{n}{\log n}$, $m_n \log m_n > \frac{n}{2}$ for large enough n and it follows from Lemma 5.15 that

$$\Pi_{\mathcal{X}}\left(\sup_{\mathbf{x}\in\mathcal{X}} \sum_{h=m_n+1}^{\infty} \pi_h(\mathbf{x}) > \epsilon\right) \precsim O(e^{-n/2}). \tag{D.7}$$

Thus with $M_n = O(n^{1/2})$,

$$\sum_{h=1}^{m_n} P(\alpha_h \notin B_{h,n}) \precsim e^{-n}. \tag{D.8}$$

Eqs. (D.6)–(D.8) together imply that $\Pi_{\mathcal{X}}(\mathcal{F}_n^c) \precsim O(e^{-n})$.

Also $m_n^\eta r_n^p \log\left(\frac{M_n\sqrt{r_n/\delta_n}}{\epsilon}\right)^{p+1} = o(n)$ for the choice of the sequence r_n. With $m_n = n/(C\log n)$ for some large $C > 0$, one can make

$$\log(N(\epsilon, B_{1,n} \times \cdots \times B_{m_n,n}, \|\cdot\|_\infty)) < n\xi \tag{D.9}$$

for any $\xi > 0$. Also from Lemma 5.14,

$$m_n \log N(\mathcal{O}_{a_n,h_n,l_n}, \epsilon, \|\cdot\|_\infty) \le m_n \log\left\{d_1\left(\frac{a_n}{l_n}\right)^p + d_2 \log\frac{h_n}{l_n} + 1\right\} < n\xi \tag{D.10}$$

for any $\xi > 0$. Combining (D.9) and (D.10), $\log N(\mathcal{F}_n, 4\epsilon, \|\cdot\|_1) < n\xi$ for any $\xi > 0$. □

Appendix E. Another useful lemma

We state without proof the following lemma needed to prove Theorem 6.1.

Lemma E.1. *For non-negative r.v.s. A_i, B_i, if $P(A_i \le u) \le C_i P(B_i \le u)$ for $u \in (0, t_0)$, $t_0 > 0$, $i = 1, 2$, $P(A_1 + A_2 \le t_0) \le C_1 C_2 P(B_1 + B_2 \le t_0)$.*

Appendix F. Proof of Theorem 6.1

Proof. Once again we approximate $f_0(y \mid \mathbf{x})$ by $\tilde{f}(y \mid \mathbf{x}) = \int \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) d\tilde{G}_{\mathbf{x}}(\mu, \sigma)$, so that the first term of (5.2) is arbitrarily small. We construct such an $\tilde{f}$ analogous to that in Theorem 5.3. Lemma F.1 is a variant of Lemma 5.5 which ensures that the second term in (5.2) is also sufficiently small. Before that we need a different notion of neighborhood of $\{F_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ which we formulate below.

$$\left\{\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} : \sup_{\mathbf{x}\in\mathcal{X}} \left| \int_{\Re\times\Re^+} \left\{g(\mu, \sigma) dG_{\mathbf{x}}(\mu, \sigma) - g(\mu, \sigma) dF_{\mathbf{x}}(\mu, \sigma)\right\} \right| < \epsilon \right\}. \tag{F.1}$$

Lemma F.1. *Assume that $f_0 \in \mathcal{F}_d$ satisfies $\int_{\mathcal{X}} \int_{\mathcal{Y}} y^2 f_0(y \mid \mathbf{x}) dy q(\mathbf{x}) d\mathbf{x} < \infty$. Suppose $\tilde{f}(y \mid \mathbf{x}) = \int \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) d\tilde{G}_{\mathbf{x}}(\mu, \sigma)$, where $\exists\, a > 0$ and $0 < \underline{\sigma} < \overline{\sigma}$ such that*

$$\tilde{G}_{\mathbf{x}}\big([-a, a] \times (\underline{\sigma}, \overline{\sigma})\big) = 1 \quad \forall\, \mathbf{x} \in \mathcal{X}, \tag{F.2}$$

so that $\tilde{G}_{\mathbf{x}}$ has compact support for each $\mathbf{x} \in \mathcal{X}$. Then given any $\epsilon > 0$, $\exists$ a neighborhood W of $\{\tilde{G}_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ which is a finite intersection of neighborhoods of the type (F.1) such that for any conditional density $f(y \mid \mathbf{x}) = \int \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) dG_{\mathbf{x}}(\mu, \sigma)$, $\mathbf{x} \in \mathcal{X}$, with $\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} \in W$,

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} f_0(y \mid \mathbf{x}) \log\frac{\tilde{f}(y \mid \mathbf{x})}{f(y \mid \mathbf{x})} dy q(\mathbf{x}) d\mathbf{x} < \epsilon. \tag{F.3}$$

The proof of Lemma F.1 is similar to that of Lemma 5.5 and is omitted here. To characterize the support of $\mathcal{P}_{\mathcal{X}}$, we define a collection of fixed conditional probability measures $\{F_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ on $(\Re \times \Re^+, \mathcal{B}(\Re \times \Re^+))$ denoted by $\mathcal{G}^{**}_{\mathcal{X}}$ satisfying $\mathbf{x} \mapsto \int_{\Re\times\Re^+} g(\mu, \sigma) dF_{\mathbf{x}}(\mu)$ is a continuous function of $\mathbf{x}$ for all bounded uniformly continuous functions $g : \Re \times \Re^+ \to [0, 1]$.

Theorem F.2. *Assume the following holds.*

T1. *G_0 is specified by $\mu_h \sim \mathrm{GP}(\mu, c)$, $\sigma_h \sim G_{0,\sigma}$ where c is chosen so that $\mathrm{GP}(0, c)$ has continuous path realizations and Π_σ is absolutely continuous w.r.t. Lebesgue measure on $\Re^+$.*

T2. *For every $k \ge 2$, $(\pi_1, \ldots, \pi_k)$ is absolutely continuous w.r.t. to the Lebesgue measure on S_{k-1}.*

T3. *For any continuous function $g : \mathcal{X} \mapsto \Re$,*

$$\mathcal{P}_{\mathcal{X}}\left\{\sup_{\mathbf{x}\in\mathcal{X}} |\mu_h(\mathbf{x}) - g(\mathbf{x})| < \epsilon\right\} > 0$$

$h = 1, \ldots, \infty$ and for any $\epsilon > 0$.

Then for a bounded uniformly continuous function $g : \Re \times \Re^+ : [0, 1]$ satisfying $g(\mu, \sigma) \to 0$ as $|\mu| \to \infty, \sigma \to \infty$,

$$\mathcal{P}_{\mathcal{X}}\left\{\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} : \sup_{\mathbf{x}\in\mathcal{X}} \left| \int_{\Re\times\Re^+} \left\{g(\mu, \sigma) dG_{\mathbf{x}}(\mu, \sigma) - g(\mu, \sigma) dF_{\mathbf{x}}(\mu, \sigma)\right\} \right| < \epsilon \right\} > 0. \tag{F.4}$$

Proof. It suffices to assume that g is coordinatewise monotonically increasing on $\Re \times \Re^+$. Let $\epsilon > 0$ be given and $\psi(\mathbf{x}) = \int_{\Re\times\Re^+} g(\mu, \sigma) dF_{\mathbf{x}}(\mu, \sigma)$. Let n_ϵ be such that $\mathcal{P}_{\mathcal{X}}(\Omega_1) > 0$ where $\Omega_1 = \{\sum_{h=n_\epsilon+1}^{\infty} \pi_h < \epsilon\}$. Then in Ω_1,

$$\left| \int_{\Re\times\Re^+} \left\{g(\mu, \sigma) dG_{\mathbf{x}}(\mu, \sigma) - \psi(\mathbf{x})\right\} \right| \le \sum_{k=1}^{n_\epsilon} \pi_k |g(\mu_k(\mathbf{x}), \sigma_k) - \psi(\mathbf{x})| + \epsilon.$$

Define $\Omega_2 = \{\sup_{\mathbf{x}\in\mathcal{X}} |g(\mu_k(\mathbf{x}), \sigma_k) - \psi(\mathbf{x})| < \epsilon, k = 1, \ldots, n_\epsilon\}$. For a fixed σ_k, there exists a δ such that $\sup_{\mathbf{x}\in\mathcal{X}} |g(\mu_k(\mathbf{x}), \sigma_k) - \psi(\mathbf{x})| < \epsilon/2$ if $\sup_{\mathbf{x}\in\mathcal{X}} \left|\mu_k(\mathbf{x}) - g^{-1}_{\sigma_k}\psi(\mathbf{x})\right| < \delta$ where $g^{-1}_{\sigma_k}$ denotes the inverse of $g(\cdot, \sigma_k)$ for

fixed σ_k. Hence there exists a neighborhood B_k of σ_k such that for $\sigma_k \in B_k$ and $\sup_{\mathbf{x}\in\mathcal{X}} \left|\mu_k(\mathbf{x}) - g_{\sigma_k}^{-1}\psi(\mathbf{x})\right| < \delta$, we have $\sup_{\mathbf{x}\in\mathcal{X}} |g(\mu_k(\mathbf{x}), \sigma_k) - \psi(\mathbf{x})| < \epsilon$. Since for each $k = 1, \ldots, n_\epsilon$, $\mathcal{P}_{\mathcal{X}}\left\{\sigma_k \in B_k, \sup_{\mathbf{x}\in\mathcal{X}} \left|\mu_k(\mathbf{x}) - g_{\sigma_k}^{-1}\psi(\mathbf{x})\right| < \delta\right\} =$

$$\int_{\sigma_k\in B_k} \mathcal{P}_{\mathcal{X}}\left\{\sup_{\mathbf{x}\in\mathcal{X}} \left|\mu_k(\mathbf{x}) - g_{\sigma_k}^{-1}\psi(\mathbf{x})\right| < \delta\right\} dG_{0,\sigma}(\sigma_k) > 0,$$

$\mathcal{P}_{\mathcal{X}}(\Omega_2) > 0$. The conclusion of the theorem follows from the independence of Ω_1 and Ω_2. □

$\tilde{f}$ in (5.2) will be constructed so as to satisfy the assumptions of Lemma F.1 and such that $\int_{\mathcal{X}} \int_{\mathcal{Y}} f_0(y \mid \mathbf{x}) \log \frac{f_0(y|\mathbf{x})}{\tilde{f}(y|\mathbf{x})} dy\, q(\mathbf{x})d\mathbf{x} < \frac{\epsilon}{2}$ for any $\epsilon > 0$. Define a sequence of conditional densities $f_n(y \mid \mathbf{x}) = \int \frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma}) d\tilde{G}_{n,\mathbf{x}}(\mu, \sigma)$, $n \geq 1$ where for $\sigma_n = n^{-\eta}$,

$$dG_{n,\mathbf{x}}(\mu, \sigma) = \frac{I_{\mu\in[-n,n]} f_0(\mu \mid \mathbf{x})\delta_{\sigma_n}(\sigma)}{\int_{-n}^{n} f_0(\mu \mid \mathbf{x})}. \tag{F.5}$$

As before define the approximator

$$f_n(y \mid \mathbf{x}) = \frac{\int_{-n}^{n} \frac{1}{\sigma_n}\phi\left(\frac{y-t}{\sigma_n}\right) f_0(t \mid \mathbf{x})dt}{\int_{-n}^{n} f_0(t \mid \mathbf{x})dt}. \tag{F.6}$$

$\tilde{f}$ will be chosen to be f_{n_0} for some large n_0. f_{n_0} satisfies the assumptions of Lemma F.1 since $\{G_{n_0,\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ is compactly supported. Moreover $\{G_{n_0,\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} \in \mathcal{G}_{\mathcal{X}}^{**}$ as $\mathbf{x} \to \int_{\Re\times\Re^+} g(\mu, \sigma) dG_{n_0,\mathbf{x}}(\mu, \sigma)$ is continuous function of $\mathbf{x}$ for all bounded uniformly continuous function g. Hence there exists a finite intersection W of neighborhoods of $\{G_{n_0,\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ the type (F.1) such that for any $\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} \in W$, the second term of (5.2) is arbitrarily small. The conclusion of the theorem follows immediately from a variant of Corollary 5.8 applied to neighborhoods of the type (F.1). □

Appendix G. Proof of Theorem 6.2

Proof. As before we establish q-integrated L_1 consistency of Gaussian mixtures of fixed-π dependent processes by verifying the conditions of Theorem 5.9. Let $\phi_{\mu,\sigma}(\mathbf{x}, y) := \frac{1}{\sigma}\phi\left(\frac{y-\mu(\mathbf{x})}{\sigma}\right)$ for $y \in \mathcal{Y}$ and $\mathbf{x} \in \mathcal{X}$. Construct B_n as

$$B_n = \left(M_n\sqrt{\frac{r_n}{\delta_n}}\mathbb{H}_1^{r_n} + \frac{\epsilon l_n\sqrt{\pi}}{4\sqrt{2}}\mathbb{B}_1\right) \cup \left(\cup_{a<\delta_n} M_n\mathbb{H}_1^{a} + \frac{\epsilon l_n\sqrt{\pi}}{4\sqrt{2}}\mathbb{B}_1\right)$$

with $\delta_n = \frac{K_1\epsilon l_n}{M_n}$ for some constant $K_1 > 0$. Let

$$\Theta_n = \left\{\phi_{\mu,\sigma} : \|\boldsymbol{\beta}\| \leq a_n, \eta \in B_n, l_n \leq \sigma \leq h_n\right\}. \tag{G.1}$$

It is easy to see that

$$\log N(\mathcal{F}_n, 4\epsilon, \|\cdot\|) \leq K_2 m_n r_n^p \left\{\log\left(\frac{8\sqrt{2}M_n\sqrt{r_n/\delta_n}}{\epsilon\sqrt{\pi}l_n}\right)\right\}^{p+1} + m_n \log\frac{K_4 m_n^2}{\epsilon} m_n \log\frac{K_3 M_n}{\epsilon l_n} + m_n \log\left\{d_1\left(\frac{a_n}{l_n}\right)^p + d_2 \log\frac{h_n}{l_n} + 1\right\}. \tag{G.2}$$

Note that $\Pi_{\mathcal{X}}(\mathcal{F}_n^c) \leq m_n P(\Theta_n^c) + P(\sum_{h=m_n}^{\infty} \pi_h > \epsilon)$ and $P(\Theta_n^c) \leq \left\{P(\|\boldsymbol{\beta}\| > a_n) + P(\sigma \in [l_n, h_n]^c) + P(\eta \in B_n^c)\right\}$. It follows from the proof of Theorem 3.1 of [48] that

$$P(\eta \in B_n^c) \leq P(A > r_n) + e^{-M_n^2/2}$$

if $M_n^2 > r_n^p\left\{\log\left(\frac{8\sqrt{2}M_n\sqrt{r_n/\delta_n}}{\epsilon\sqrt{\pi}l_n}\right)\right\}$. Since $A^{p(1+\eta_2)/\eta_2} \sim \text{Ga}(a, b)$, Lemma 4.9 of [48] indicates that $P(A > r_n) \precsim \exp\{-r_n^{p(1+\eta_2)/\eta_2}\}$. Hence with $M_n = O(n^{1/2})$, $m_n = O\{n/(\log n)^{p+1}\}^{1/(1+\eta_2)}$ and $r_n^p = O\{n^{\eta_2/(1+\eta_2)}\}$, $P(\Theta_n^c) \precsim e^{-n}$ and

$$P\left(\sum_{h=m_n}^{\infty} \pi_h > \epsilon\right) \precsim \exp\{-m_n^{1+\eta_2}(\log m_n)^{(p+1)}\} \precsim e^{-n}. \tag{G.3}$$

Also, the first term in the right hand side of (G.2) can be made smaller than $n\xi$ since $m_n r_n^p = O(n/(\log n)^{p+1})$. Also by F1, the last two terms of the right hand side of (G.2) can be made to grow at $o(n)$. □

References

[1] R. Adler, An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes, Vol. 12, Institute of Mathematical Statistics, 1990.
[2] F. Barrientos, A. Jara, F. Quintana, On the support of MacEachern's dependent Dirichlet processes, Bayesian Analysis 7 (2012) 1–34.
[3] A. Barron, M. Schervish, L. Wasserman, The consistency of posterior distributions in nonparametric problems, The Annals of Statistics 27 (1999) 536–561.
[4] A. Bhattacharya, D. Dunson, Strong consistency of nonparametric Bayes density estimation on compact metric spaces with applications to specific manifolds, Annals of the Institute of Statistical Mathematics (2011) 1–28.
[5] I. Castillo, R. Nickl, Nonparametric bernstein-von mises theorems, 2012. Preprint arXiv:1208.3862.
[6] Y. Chung, D. Dunson, Nonparametric Bayes conditional distribution modeling with variable selection, Journal of the American Statistical Association 104 (2009) 1646–1660.
[7] M. De Iorio, W. Johnson, P. Müller, G. Rosner, Bayesian nonparametric nonproportional hazards survival modeling, Biometrics 65 (2009) 762–771.
[8] M. De Iorio, P. Mueller, G. Rosner, S. MacEachern, An ANOVA model for dependent random measures, Journal of the American Statistical Association 99 (2004) 205–215.
[9] D. Dunson, J. Park, Kernel stick-breaking processes, Biometrika 95 (2008) 307–323.
[10] D. Dunson, N. Pillai, J. Park, Bayesian density regression, Journal of the Royal Statistical Society: Series B 69 (2007) 163–183.
[11] J. Fan, Q. Yao, H. Tong, Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems, Biometrika 83 (1996) 189–206.
[12] T. Ferguson, A Bayesian analysis of some nonparametric problems, The Annals of Statistics 1 (1973) 209–230.
[13] T. Ferguson, Prior distributions on spaces of probability measures, The Annals of Statistics 2 (1974) 615–629.
[14] D. Freedman, Wald lecture: on the Bernstein–von Mises theorem with infinite-dimensional parameters, The Annals of Statistics 27 (1999) 1119–1141.
[15] A. Gelfand, A. Kottas, S. MacEachern, Bayesian nonparametric spatial modeling with Dirichlet process mixing, Journal of the American Statistical Association 100 (2005) 1021–1035.
[16] S. Ghosal, J. Ghosh, R. Ramamoorthi, Posterior consistency of Dirichlet mixtures in density estimation, The Annals of Statistics 27 (1999) 143–158.
[17] S. Ghosal, A. van der Vaart, Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities, The Annals of Statistics 29 (2001) 1233–1263.
[18] S. Ghosal, A. van der Vaart, Posterior convergence rates of Dirichlet mixtures at smooth densities, The Annals of Statistics 35 (2007) 697–723.
[19] J. Griffin, M. Steel, Order-based dependent Dirichlet processes, Journal of the American Statistical Association 101 (2006) 179–194.
[20] J. Griffin, M. Steel, Bayesian nonparametric modelling with the dirichlet process regression smoother, Statistica Sinica 20 (2010) 1507–1527.
[21] H. Ishwaran, L. James, Gibbs sampling methods for stick-breaking priors, Journal of the American Statistical Association 96 (2001) 161–173.
[22] S. Jain, R. Neal, A split-merge Markov Chain Monte Carlo procedure for the dirichlet process mixture model, Journal of Computational and Graphical Statistics 13 (2004) 158–182.
[23] A. Jara, E. Lesaffre, M. De Iorio, F. Quintana, Bayesian semiparametric inference for multivariate doubly-interval-censored data, The Annals of Applied Statistics 4 (2010) 2126–2149.
[24] M. Kalli, J. Griffin, S. Walker, Slice sampling mixture models, Statistics and computing (2010) 1–13.
[25] W. Kruijer, J. Rousseau, A. Van Der Vaart, Adaptive Bayesian density estimation with location-scale mixtures, Electronic Journal of Statistics 4 (2010) 1225–1257.
[26] A. Lo, On a class of Bayesian nonparametric estimates: I. Density estimates, The Annals of Statistics 12 (1984) 351–357.
[27] S.N. MacEachern, Dependent nonparametric processes, in: ASA Proceedings of the Section on Bayesian Statistical Science, American Statistical Association, Alexandria, VA, 1999, pp. 50–55.
[28] T.P. Minka, Expectation propagation for approximate Bayesian inference, in: Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.
[29] P. Müller, A. Erkanli, M. West, Bayesian curve fitting using multivariate normal mixtures, Biometrika 83 (1996) 67–79.
[30] A. Norets, Approximation of conditional densities by smooth mixtures of regressions, The Annals of Statistics 38 (2010) 1733–1766.
[31] A. Norets, J. Pelenis, Posterior consistency in conditional density estimation by covariate dependent mixtures, Unpublished Manuscript, Princeton Univ, 2010.
[32] O. Papaspiliopoulos, A note on posterior sampling from Dirichlet mixture models, Technical Report, 2008.
[33] O. Papaspiliopoulos, G. Roberts, Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models, Biometrika 95 (2008) 169.
[34] O. Papaspiliopoulos, G. Roberts, Retrospective Markov Chain Monte Carlo methods for Dirichlet process hierarchical models, Biometrika 95 (2008) 169–183.
[35] D. Pati, D. Dunson, Bayesian nonparametric regression with varying residual density, Unpublished Paper, 2009.
[36] V. Rivoirard, J. Rousseau, Bernstein–von Mises theorem for linear functionals of the density, The Annals of Statistics 40 (2012) 1489–1523.
[37] A. Rodriguez, D. Dunson, Nonparametric Bayesian models through probit stick-breaking processes, Bayesian Analysis 6 (2011) 145–178.
[38] A. Rojas, C. Genovese, C. Miller, R. Nichol, L. Wasserman, Conditional density estimation using finite mixture models with an application to astrophysics, 2005.
[39] C. Scricciolo, Posterior rates of convergence for dirichlet mixtures of exponential power densities, Electronic Journal of Statistics 5 (2011) 270–308.
[40] W. Shen, S. Tokdar, S. Ghosal, Adaptive Bayesian multivariate density estimation with Dirichlet mixtures, 2011. Preprint arXiv:1109.6406.
[41] Y. Tang, S. Ghosal, A consistent nonparametric Bayesian procedure for estimating autoregressive conditional densities, Computational Statistics & Data Analysis 51 (2007) 4424–4437.
[42] Y. Tang, S. Ghosal, Posterior consistency of Dirichlet mixtures for estimating a transition density, Journal of Statistical Planning and Inference 137 (2007) 1711–1726.
[43] S. Tokdar, Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression, Sankhyā: The Indian Journal of Statistics 67 (2006) 90–110.
[44] S. Tokdar, Adaptive convergence rates in Dirichlet process mixtures of multivariate normals, 2011. Preprint arXiv:1111.4148.
[45] S. Tokdar, J. Ghosh, Posterior consistency of logistic Gaussian process priors in density estimation, Journal of Statistical Planning and Inference 137 (2007) 34–42.
[46] S. Tokdar, Y. Zhu, J. Ghosh, Bayesian density regression with logistic Gaussian process and subspace projection, Bayesian Analysis 5 (2010) 1–26.
[47] A. van der Vaart, J. van Zanten, Reproducing kernel Hilbert spaces of Gaussian priors, IMS Collections 3 (2008) 200–222.
[48] A. van der Vaart, J. van Zanten, Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth, The Annals of Statistics 37 (2009) 2655–2675.
[49] S. Walker, Sampling the Dirichlet mixture model with slices, Communications in Statistics—Simulation and Computation 36 (2007) 45–54.
[50] Y. Wu, S. Ghosal, Kullback Leibler property of kernel mixture priors in Bayesian density estimation, Electronic Journal of Statistics 2 (2008) 298–331.
[51] Y. Wu, S. Ghosal, The $L1$-consistency of Dirichlet mixtures in multivariate Bayesian density estimation, Journal of Multivariate Analysis (2010) 2411–2419.
[52] J. Yoon, Bayesian analysis of conditional density functions: a limited information approach, Unpublished Manuscript, Claremont Mckenna College, 2009.