

# Posterior consistency in conditional distribution estimation

Debdeep Pati<sup>\*</sup>, David Dunson<sup>†</sup> and Surya Tokdar<sup>†</sup>

*BOX 90251, Old Chemistry Building  
Durham, NC 27708  
e-mail: [dp55@stat.duke.edu](mailto:dp55@stat.duke.edu)*

[dunson@stat.duke.edu](mailto:dunson@stat.duke.edu); [tokdar@stat.duke.edu](mailto:tokdar@stat.duke.edu)

**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. In particular, we have shown posterior consistency using the supremum of the  $L_1$  neighborhoods of the conditional densities across the covariate space. 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 constructed through mapping continuous stochastic processes to the unit interval using a monotone differentiable link function. Probit stick-breaking processes provide a computationally convenient special case. We also provide a set of sufficient conditions to ensure posterior consistency using Gaussian mixtures of fixed- $\pi$  dependent processes.

**AMS 2000 subject classifications:** Primary 62G07, 62G20; secondary 60K35.

**Keywords and phrases:** Asymptotics, Bayesian nonparametrics, Density regression, Large support, Probit stick-breaking process, Dependent Dirichlet process.

## 1. Introduction

There is a rich literature on Bayesian methods for density estimation using mixture models of the form

$$y_i \sim f(\theta_i), \quad \theta_i \sim P, \quad P \sim \Pi, \quad (1.1)$$

where  $f(\cdot)$  is a parametric density and  $P$  is an unknown mixing distribution assigned a prior  $\Pi$ . The most common choice of  $\Pi$  is Dirichlet process prior, first introduced by [Ferguson \(1973, 1974\)](#). [Barron, Schervish and Wasserman](#)

---

<sup>\*</sup>PhD. Student, Department of Statistical Science

<sup>†</sup>Professor, Department of Statistical Science

(1999); Ghosal, Ghosh and Ramamoorthi (1999) used upper bracketing and  $L_1$ -metric entropy bounds respectively to derive sufficient conditions on the prior on  $f$  and the true data generating  $f$  for obtaining strong posterior consistency in Bayesian density estimation. Ghosal, Ghosh and Ramamoorthi (1999) also provided sufficient conditions for posterior consistency in univariate density estimation using Dirichlet process location mixtures of normals. Tokdar (2006) significantly relaxed their conditions in a Dirichlet process location-scale mixture of normals setting, requiring existence of only weak moments of the true  $f$ . Ghosal and van der Vaart (2001, 2007) provided rates of convergence for Bayesian univariate density estimation using a Dirichlet process mixture of normals. Bhattacharya and Dunson (2010) provided conditions for strong consistency of kernel mixture priors for densities on compact metric spaces and manifolds.

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  $f(y | \mathbf{x})$  as a function valued parameter and allows its center, spread, skewness, modality and other such features to vary with  $\mathbf{x}$ . For data  $\{(y_i, \mathbf{x}_i), i = 1, \dots, n\}$  let

$$y_i | \mathbf{x}_i \sim f(y | \mathbf{x}_i), \quad \{f(\cdot | \mathbf{x}), \mathbf{x} \in X\} \sim \Pi_{\mathcal{X}}, \quad (1.2)$$

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 Müller, Erkanli and West (1996); Griffin and Steel (2006, 2008); Dunson, Pillai and Park (2007); Dunson and Park (2008); Chung and Dunson (2009) and Tokdar, Zhu and Ghosh (2010) among others.

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

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

where  $\phi$  is the standard normal density,  $\{\pi_h(\mathbf{x}), h = 1, 2, \dots\}$  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  $\mathcal{F}_{\mathcal{X}} \times \mathbb{R}^+$ ,  $\mathcal{F}_{\mathcal{X}} \subset \mathcal{X}^{\mathbb{R}}$ . However, there is a dearth of results on support properties of prior distributions for conditional distributions and on general theorems providing conditions for weak and strong posterior consistency. To our knowledge, only Barrientos, Jara and Quintana (2011) have considered formalizing the notions of weak and KL-support for dependent stick-breaking processes. However, our current approach is completely independent of theirs and focuses on providing sufficient conditions for supports based on stronger topologies. We focus on a broad class of generalized stick-breaking 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 (MacEachern, 1999) as a special case.

To our knowledge, only a few papers have considered posterior consistency in conditional density estimation. Tokdar, Zhu and Ghosh (2010) considers posterior consistency in estimating conditional distributions focusing exclusively on logistic Gaussian process priors (Tokdar and Ghosh, 2007). Such priors have beautiful theoretical properties but lack the computational simplicity of the countable mixture priors in (1.3). In addition, (1.3) has the appealing side effect of inducing predictor-dependent clustering, which is often of interest in itself and is an aid to interpretation and inferences. Yoon (2009) considers posterior consistency in conditional distribution estimation through a limited information approach by approximating the likelihood by the quantiles of the true distribution. Tang and Ghosal (2007a,b) 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. While Tokdar, Zhu and Ghosh (2010) focussed on  $L_1$ -neighborhoods of the induced joint densities, Tang and Ghosal (2007a,b) defined several topologies suitable for transition densities which are also relevant for conditional densities.

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,  $L_1$ -integrated and sup- $L_1$  neighborhoods that are appropriate for conditional distribution modeling. We then show that the sufficient conditions are satisfied for a novel class of generalized stick-breaking priors that construct the stick-breaking lengths through mapping continuous stochastic processes to the unit interval using a monotone differentiable link function. 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 Chung and Dunson (2009) and Rodriguez and Dunson (2011). We also considered Gaussian mixtures of fixed- $\pi$  dependent processes (MacEachern, 1999; De Iorio *et al.*, 2004).

The fundamental contributions of this article are 1) showing consistency using a general topology for conditional densities and 2) the development of a novel method of constructing a sieve for the proposed class of priors. The joint  $L_1$  topology concerns average accuracy for prediction of future  $y$  values when the future  $\mathbf{x}$  values are drawn from the same covariate distribution  $Q$  that generate the data  $\mathbf{x}$ 's. A better measure of learning the conditional density obtains if similar average accuracies can be guaranteed when the future  $\mathbf{x}$ 's are generated from any arbitrary measure  $\nu$  whose support is a subset of the support of  $Q$ . In this article, we have focused on a topology using supremum of the  $L_1$  neighborhoods (Tang and Ghosal, 2007a,b) of the true conditional density for posterior consistency. The sup- $L_1$  topology is even a stronger form of assuring these appealing ways of learning a conditional density.

Our next contribution is the construction of a sieve suited to predictor dependent mixture priors. It has been noted by Wu and Ghosal (2010) 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<sup>1</sup> avoids the pitfall mentioned by [Wu and Ghosal \(2010\)](#) in showing consistency using multivariate mixtures.

## 2. Notations

Throughout the paper, Lebesgue measure on  $\mathfrak{R}$  or  $\mathfrak{R}^p$  is denoted by  $\lambda$  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  $\nu$ -integrable  $p$ th 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  $\epsilon$ -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  $\epsilon$  needed to cover  $M$ . The logarithm of the covering number is referred to as the entropy. “ $\gtrsim$ ” stands for inequality up to a constant multiple or if the constant multiple is irrelevant to the given situation.  $\langle 0 \rangle$  stands for a distribution degenerate at 0 and  $\text{supp}(\nu)$  for the support of a measure  $\nu$ .

## 3. Gaussian process priors

In this section, we first recall the definition of the RKHS of a Gaussian process prior. [van der Vaart and van Zanten \(2008\)](#) reviews facts that are relevant to the present applications.

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 completion of the linear space of functions  $t \mapsto EW(t)H$  relative to the inner product

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

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  $W$ . 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  $\|\cdot\|$ .

<sup>1</sup>A similar sieve appears in [Norets and Pelenis \(2010\)](#) with a citation to an earlier draft of our paper.

#### 4. 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  $\mathcal{Y} = \mathbb{R}$  be the response space and  $\mathcal{X}$  be the covariate space which is a compact subset of  $\mathbb{R}^p$ . 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 the space of all conditional densities,

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

Suppose  $y_i$  is observed independently given the covariates  $\mathbf{x}_i$ ,  $i = 1, 2, \dots$  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  $h(\mathbf{x}, y) = q(\mathbf{x})f(y | \mathbf{x})$  and  $h_0(\mathbf{x}, y) = q(\mathbf{x})f_0(y | \mathbf{x})$  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  $\{f_0(\cdot | \mathbf{x}), \mathbf{x} \in \mathcal{X}\}$  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

$$f(y | \mathbf{x}) = \int \frac{1}{\sigma} \phi\left(\frac{y - \mu}{\sigma}\right) dG_{\mathbf{x}}(\psi), \quad (4.1)$$

where  $\psi = (\mu, \sigma)$ ,  $\phi$  is the standard normal pdf and

$$G_{\mathbf{x}} = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \delta_{\{\mu_h(\mathbf{x}), \sigma_h\}}, \quad (\mu_h, \sigma_h) \sim G_0, \quad (4.2)$$

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  $\mathbf{x} \in \mathcal{X}$ .  $\{\mu_h(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}_{h=1}^{\infty}$  are i.i.d. realizations of a real valued stochastic process, i.e.,  $G_0$  is a probability distribution over  $\mathcal{F}_{\mathcal{X}} \times \mathbb{R}^+$ , where  $\mathcal{F}_{\mathcal{X}}$  is a function space. Hence for each  $\mathbf{x} \in \mathcal{X}$ ,  $G_{\mathbf{x}}$  is a random probability measure over the measurable Polish space  $(\mathbb{R} \times \mathbb{R}^+, \mathcal{B}(\mathbb{R} \times \mathbb{R}^+))$ . We are interested in Bayesian posterior consistency for a broad class of predictor-dependent stick-breaking mixtures including the following two important special cases.

#### 4.1. Predictor dependent countable mixtures of Gaussian linear regressions

We define the predictor dependent countable mixtures of Gaussian linear regressions (MGLR<sub>**x**</sub>) as

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

where  $\phi$  is the standard normal pdf and

$$G_{\mathbf{x}} = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \delta_{(\boldsymbol{\beta}_h, \sigma_h)}, \quad (\boldsymbol{\beta}_h, \sigma_h) \sim G_0 \quad (4.3)$$

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  $\mathbb{R}^p \times \mathbb{R}^+$  where  $G_{0,\boldsymbol{\beta}}$  and  $G_{0,\sigma}$  are probability distributions on  $\mathbb{R}^p$  and  $\mathbb{R}^+$  respectively.

#### 4.2. Gaussian mixtures of fixed- $\pi$ dependent processes

In (4.1), set  $G_{\mathbf{x}}$  as in (4.2) 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  $\{\mu_h(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}_{h=1}^{\infty}$  are as in (4.2). Examples are fixed- $\pi$  dependent Dirichlet process mixtures of Gaussians (MacEachern, 1999).

Probit stick-breaking mixtures of Gaussians has been previously applied to real data (Chung and Dunson, 2009; Rodriguez and Dunson, 2011; Pati and Dunson, 2009). The latter two articles considered probit transformations of Gaussian processes in constructing the stick-breaking weights. Such latent Gaussian processes can be updated using data augmentation Gibbs sampling as in continuation-ratio probit models for survival analysis (Albert and Chib, 2001). The infinite mixture of normals can be handled via a novel combination of the slice sampler as discussed in Walker (2007) and the retrospective sampler in Papaspiliopoulos and Roberts (2008). On the other hand, versions of the fixed  $\pi$ -DDP have been applied to ANOVA (De Iorio *et al.*, 2004), survival analysis (De Iorio *et al.*, 2009; Jara *et al.*, 2010), spatial modeling (Gelfand, Kottas and MacEachern, 2005), and many more.

### 5. Notions of neighborhoods in conditional density estimation

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

$$W_{\epsilon,g}(f_0) = \left\{ f : f \in \mathcal{F}_d, \left| \int_{\mathcal{X} \times \mathcal{Y}} gh - \int_{\mathcal{X} \times \mathcal{Y}} gh_0 \right| < \epsilon \right\}, \quad (5.1)$$

for a bounded continuous function  $g : \mathcal{Y} \times \mathcal{X} \rightarrow \mathfrak{R}$ . A weak neighborhood base is formed by finite intersections of neighborhoods of the type (5.1). Define a  $\nu$ -integrated  $L_1$  neighborhood

$$S_\epsilon(f_0; \nu) = \{f : f \in \mathcal{F}_d, \int \|f(\cdot | \mathbf{x}) - f_0(\cdot | \mathbf{x})\|_1 \nu(\mathbf{x}) d\mathbf{x} < \epsilon\} \quad (5.2)$$

for any measure  $\nu$  with  $\text{supp}(\nu) \subset \mathcal{X}$ . Observe that under the topology in (5.2),  $\mathcal{F}_d$  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. For  $f_1, f_2 \in \mathcal{F}_d$ , let  $d_{SS}(f_1, f_2) = \sup_{\mathbf{x} \in \mathcal{X}} \|f_1(\cdot | \mathbf{x}) - f_2(\cdot | \mathbf{x})\|_1$  and define the sup- $L_1$  neighborhood

$$SS_\epsilon(f_0) = \{f : f \in \mathcal{F}_d, d_{SS}(f, f_0) < \epsilon\}. \quad (5.3)$$

Under the sup- $L_1$  topology,  $\mathcal{F}_d$  can be viewed as a closed subset of the separable Banach space of continuous functions from  $\mathcal{X} \rightarrow L_1(\lambda, \mathcal{Y})$  which are norm bounded and hence a complete separable metric space. Thus measurability issues won't 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 5.1.** *For any  $f_0 \in \mathcal{F}_d$ , such that  $h_0(\mathbf{x}, y) = q(\mathbf{x})f_0(y | \mathbf{x})$  is the true joint data-generating density, we define an  $\epsilon$ -sized KL neighborhood around  $f_0$  as*

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

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

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

**Definition 5.2.** *The posterior  $\Pi_{\mathcal{X}}(\cdot | \mathbf{y}^n, \mathbf{x}^n)$  is consistent weakly, strongly in the  $\nu$ -integrated  $L_1$  topology or strongly in the sup- $L_1$  topology at  $\{f_0(\cdot | \mathbf{x}), \mathbf{x} \in \mathcal{X}\}$  if  $\Pi_{\mathcal{X}}(U^c | \mathbf{y}^n, \mathbf{x}^n) \rightarrow 0$  a.s. for any  $\epsilon > 0$  with  $U = W_\epsilon(f_0), S_\epsilon(f_0; \nu)$  and  $SS_\epsilon(f_0)$  respectively.*

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

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  $S \times \mathfrak{R}^+$  for each  $\mathbf{x} \in \mathcal{X}$ . Here  $S = \mathfrak{R}^p$  or  $\mathfrak{R}$  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 5.3.** For a bounded continuous function  $g : S \times \mathbb{R}^+ \times \mathcal{X} \rightarrow \mathbb{R}$  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

$$\left\{ \{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} : \left| \int_{S \times \mathbb{R}^+ \times \mathcal{X}} g(s, \sigma, \mathbf{x}) dG_{\mathbf{x}}(s, \sigma) q(\mathbf{x}) d\mathbf{x} - \int_{S \times \mathbb{R}^+ \times \mathcal{X}} g(s, \sigma, \mathbf{x}) dF_{\mathbf{x}}(s, \sigma) q(\mathbf{x}) d\mathbf{x} \right| < \epsilon \right\} \quad (5.4)$$

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 (5.4). In the sequel, we will also consider a neighborhood of the form

$$\left\{ \{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} : \sup_{\mathbf{x} \in \mathcal{X}} \left| \int_{S \times \mathbb{R}^+} \{g(s, \sigma) dG_{\mathbf{x}}(s, \sigma) - g(s, \sigma) dF_{\mathbf{x}}(s, \sigma)\} \right| < \epsilon \right\}. \quad (5.5)$$

for a bounded continuous function  $g : S \times \mathbb{R}^+ \rightarrow \mathbb{R}$ .

## 6. Posterior consistency in MGLR<sub>x</sub> mixture of Gaussians

### 6.1. Kullback-Leibler property

We will work with a specific choice of  $\mathcal{P}_{\mathcal{X}}$  motivated by the probit stick breaking process construction in [Chung and Dunson \(2009\)](#) but using Gaussian process transforms instead of Gaussian transforms. Let

$$\pi_h(\mathbf{x}) = \Phi\{\alpha_h(\mathbf{x})\} \prod_{l < h} [1 - \Phi\{\alpha_l(\mathbf{x})\}], \quad (6.1)$$

where  $\alpha_h \sim \text{GP}(0, c_h)$ , for  $h = 1, 2, \dots, \infty$ .

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  $f_0(y | \mathbf{x})$  and approximate it by  $\tilde{f}(y | \mathbf{x}) = \int \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{x}'\beta}{\sigma}\right) d\tilde{G}_{\mathbf{x}}(\beta, \sigma)$ , where  $\{\tilde{G}_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$  is compactly supported. Observe that,

$$KL(h_0, h) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f_0(y | \mathbf{x}) \log \frac{f_0(y | \mathbf{x})}{\tilde{f}(y | \mathbf{x})} dy q(\mathbf{x}) d\mathbf{x} + \int_{\mathcal{X}} \int_{\mathcal{Y}} f_0(y | \mathbf{x}) \log \frac{\tilde{f}(y | \mathbf{x})}{f(y | \mathbf{x})} dy q(\mathbf{x}) d\mathbf{x}. \quad (6.2)$$

We construct such an  $\tilde{f}$  in [Theorem 6.6](#) which makes the first term in the right hand side of (6.2) sufficiently small. The following lemma (which is similar to [Lemma 3.1](#) in [Tokdar \(2006\)](#) and [Theorem 3](#) in

Ghosal, Ghosh and Ramamoorthi (1999) guarantees that the second term in the right hand side of (6.2) 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 (5.5).

**Lemma 6.1.** *Assume that  $f_0 \in \mathcal{F}_d$  satisfies  $\int_{\mathcal{X}} \int_{\mathcal{Y}} y^2 f_0(y | \mathbf{x}) dy q(\mathbf{x}) d\mathbf{x} < \infty$ . Suppose  $\tilde{f}(y | \mathbf{x}) = \int \frac{1}{\sigma} \phi(\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}) d\tilde{G}_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)$ , where  $\exists a > 0$  and  $0 < \underline{\sigma} < \bar{\sigma}$  such that*

$$\tilde{G}_{\mathbf{x}}([-a, a]^p \times (\underline{\sigma}, \bar{\sigma})) = 1 \quad \forall \mathbf{x} \in \mathcal{X}, \quad (6.3)$$

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 (5.5) such that for any conditional density  $f(y | \mathbf{x}) = \int \frac{1}{\sigma} \phi(\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}) dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)$ ,  $\mathbf{x} \in \mathcal{X}$ , with  $\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} \in W$ ,

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} f_0(y | \mathbf{x}) \log \frac{\tilde{f}(y | \mathbf{x})}{f(y | \mathbf{x})} dy q(\mathbf{x}) d\mathbf{x} < \epsilon. \quad (6.4)$$

The proof of Lemma 6.1 is provided in Appendix A. 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 6.1, we define a collection of fixed conditional probability measures on  $(\mathbb{R}^p \times \mathbb{R}^+, \mathcal{B}(\mathbb{R}^p \times \mathbb{R}^+))$  denoted by  $\mathcal{G}_{\mathcal{X}}^*$  satisfying

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

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

**Theorem 6.2.** *Assume the following holds.*

- S1.  $c_h$  is chosen so that  $\alpha_h \sim GP(0, c_h)$  has continuous path realizations and
- S2. for any continuous function  $g : \mathcal{X} \mapsto \mathbb{R}$ ,

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

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

- S3.  $G_0$  is absolutely continuous with respect to  $\lambda(\mathbb{R}^p \times \mathbb{R}^+)$ .

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

$$\mathcal{P}_{\mathcal{X}} \left\{ \{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} : \sup_{\mathbf{x} \in \mathcal{X}} \left| \int_{\mathbb{R}^p \times \mathbb{R}^+} g(\boldsymbol{\beta}, \sigma) dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma) - g(\boldsymbol{\beta}, \sigma) dF_{\mathbf{x}}(\boldsymbol{\beta}, \sigma) \right| < \epsilon \right\} > 0. \quad (6.5)$$

*Proof.* The proof is similar to Theorem 6.2 with the  $L_1$  convergence in (C.1) replaced by convergence uniformly in  $\mathbf{x}$ . This is because under the assumptions of Corollary 6.3, the uniformly continuous sequence of functions  $\sum_{k=1}^n g(\tilde{\beta}_{k,n}, \tilde{\sigma}_{k,n}) F_{\mathbf{x}}(A_{k,n})$  on  $\mathcal{X}$  monotonically decreases to  $\int_C g(\beta, \sigma) dF_{\mathbf{x}}(\beta, \sigma)$  as  $n \rightarrow \infty$  where  $C$  is given by  $[-a, a]^p \times [\underline{\sigma}, \underline{\sigma}]$ .  $\square$

The proof of the following corollary is along the lines of the proof of Theorem 6.2 and is omitted here.

**Corollary 6.4.** *Under the assumptions of Corollary 6.3 for any  $k_0 \geq 1$ ,*

$$\mathcal{P}_{\mathcal{X}} \left\{ \bigcap_{j=1}^{k_0} U_j \right\} > 0, \quad (6.6)$$

where  $U_j$ 's are neighborhoods of the type (6.5).

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

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

**Remark 6.5.** *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 residual density as heavy-tailed as a  $t$ -density with 3 degrees of freedom ( $T_3$ ) satisfies A4. However conditions A2 and A3 require case by case verification. Here we focus on a general class of models which satisfies A2 and A3 together with the others. Let  $y_i | \mathbf{x}_i \sim \mu(x_i) + \epsilon_i$ ,  $\mu(x)$  is a continuous function from  $\mathcal{X} \rightarrow \mathfrak{R}$ ,  $\epsilon_i \sim f_{\mathbf{x}_i}$ , where  $f_{\mathbf{x}}(y) = \sum_{h=1}^H \pi_h(\mathbf{x}) \psi(y; \mu_h, \sigma_h^2)$  for some  $H \geq 1$  and  $\sum_{h=1}^H \pi_h(x) = 1$ ,  $\pi_h : \mathcal{X} \rightarrow [0, 1]$  being continuous. Then A1-A5 are satisfied with  $\psi$  chosen to be Gaussian or  $T_\nu$  with  $\nu \geq (2 + \delta)$  for some  $\delta > 0$ .*

The following theorem characterizes the subset of  $\mathcal{F}_d$  for which  $\Pi_{\mathcal{X}}$  has the KL property. The proof of Theorem 6.6 is provided in Appendix D.

**Theorem 6.6.**  $f_0 \in KL(\Pi_{\mathcal{X}})$  for each  $f_0$  in  $\mathcal{F}_d^*$  if  $\mathcal{P}_{\mathcal{X}}$  satisfies S1-S3.

**Remark 6.7.** *Although we have demonstrated the theory using a probit transformation of Gaussian processes, the conditions are satisfied for a class of generalized stick-breaking 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.*

### 6.2. Strong Consistency with the sup- $L_1$ neighborhood

To obtain strong consistency in the sup- $L_1$  topology, we need an extension of [Schwartz \(1965\)](#) theorem as below. The proof of the theorem is similar to a similar one in [Schwartz \(1965\)](#); [Tang and Ghosal \(2007b\)](#) and we omit the details.

**Theorem 6.8.** *Suppose there exists a sequence of test functions  $\Phi_n = \Phi\{(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\}$  for testing  $H_0 : f = f_0$  against  $H_1 : f \in SS_\epsilon(f_0)^c$  and subsets  $\mathcal{F}_n$  such that*

1.  $f_0 \in KL(\Pi_{\mathcal{X}})$ ,
2.  $\Phi_n \rightarrow 0$  a.s,
3.  $\sup_{f \in SS_\epsilon(f_0)^c \cap \mathcal{F}_n} \int_{\mathcal{X}^n} \int_{\mathcal{Y}^n} (1 - \Phi_n) \prod_{i=1}^n f(y_i | \mathbf{x}_i) q(\mathbf{x}_i) dy_i d\mathbf{x}_i < c_1 e^{-n\beta_1}$   
for some  $c_1, \beta_1 > 0$ ,
4.  $\Pi_{\mathcal{X}}(\mathcal{F}_n^c) \leq c_2 e^{-n\beta_2}$  for some  $c_2, \beta_2 > 0$ ,

then the posterior is strongly consistent with respect to the sup- $L_1$  neighborhood.

Let  $\phi_{\beta, \sigma}(\mathbf{x}, y) := \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{x}'\beta}{\sigma}\right)$  for  $y \in \mathcal{Y}$  and  $\mathbf{x} \in \mathcal{X}$ . From [Tokdar \(2006\)](#), we obtain for  $\sigma_2 > \sigma_1 > \frac{\sigma_2^2}{2}$  and for each  $\mathbf{x} \in \mathcal{X}$ ,

$$\int_{\mathcal{Y}} |\phi_{\beta_1, \sigma_1}(\mathbf{x}, y) - \phi_{\beta_2, \sigma_2}(\mathbf{x}, y)| dy \leq \left(\frac{2}{\pi}\right)^{1/2} \frac{\|\beta_2 - \beta_1\| \sqrt{p}}{\sigma_2} + \frac{3(\sigma_2 - \sigma_1)}{\sigma_1}$$

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

$$\Theta_{a, h, l} = \{\phi_{\beta, \sigma} : \|\beta\| \leq a, l \leq \sigma \leq h\}. \quad (6.7)$$

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 [Tokdar \(2006\)](#).

**Lemma 6.9.** *There exists constants  $d_1, d_2 > 0$  such that  $N(\Theta_{a, h, l}, \epsilon, d_{SS}) \leq d_1 \left(\frac{a}{l}\right)^p + d_2 \log \frac{h}{l} + 1$ .*

Before stating the main theorem on strong consistency, we consider a hierarchical extension of 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  $\alpha_h$ ,  $h \geq 1$  each having  $\mathfrak{R}^+$  as its support. Note that the proportion with which higher indexed atoms are selected can only explain a small fraction of the variability with respect to the covariate. As the index  $h$  increases, we define a new cut from the interval  $[\sum_{i=1}^{h-1} \pi_i(\mathbf{x}), 1]$  by the random variable  $\Phi\{\alpha_h(\mathbf{x})\}$  for each  $\mathbf{x} \in \mathcal{X}$ . Hence we would expect that the variability in the stochastic process  $\Phi\{\alpha_h\}$  due to the covariate decreases as  $h$  increases. This is maneuvered through the prior for the covariance kernel  $c_h$  of the Gaussian process  $\alpha_h$ .

Let  $\alpha_0$  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  $\alpha_h$  with respect to the covariate is shrunk or stretched to the rectangle  $[0, A_h^{1/2}]^p$  as  $A_h$  decreases or increases. We want  $A_h$ 's to be stochastically decreasing to  $\delta_0$  so that the covariate variability fades out for the higher indexed weights. A simple prior construction can be achieved by letting  $A_h^d \sim$

$\text{Ga}(a, b), h = 1, \dots, L$  for some integer  $d$  and  $A_h \equiv \delta_0$  for  $h > L$  where  $L$  is a random variable with support on  $\mathbb{N}$ . The appropriate tail condition on the random truncation  $L$  required to achieve strong posterior consistency with the sup- $L_1$  topology is discussed in Remark 6.15. However, there are computational challenges in updating  $L$ . Instead of reducing the domain of covariate dependence  $[0, A_h^{1/2}]^p$  abruptly to 0 after some random index, we shrunk the interval to 0 gradually in the following more general construction. We will focus on this general version of the prior in Lemma 6.10 and Theorem 6.11 and give a brief sketch of the assumptions required for the truncated version in Remark 6.15.

$A_h \rightarrow \langle 0 \rangle$  in distribution and there exist  $0 < \eta, \eta_0 > 0$  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 6.14 following Theorem 6.11.

The following lemma is crucial to the proof of Theorem 6.11 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 6.10.** *Let  $\pi_h$ 's satisfy (6.1) with  $\alpha_h \sim \text{GP}(0, c_h)$  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  $C_\tau > 0$ ,*

$$\Pi_{\mathcal{X}} \left( \left\| \sum_{h=m_n+1}^{\infty} \pi_h \right\|_{\infty} > \epsilon \right) \leq e^{-C_\tau m_n \log m_n} + \sum_{h=m_n+1}^{m_n} P(A_h > \delta_n). \quad (6.8)$$

*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  $\gamma_0$  in the sequel. Let  $t_0 = -\log \epsilon > 0$ . Observe that

$$\begin{aligned} \Pi_{\mathcal{X}} \left( \left\| \sum_{h=m_n+1}^{\infty} \pi_h \right\|_{\infty} > \epsilon \right) &= \Pi_{\mathcal{X}} \left( \sup_{\mathbf{x} \in \mathcal{X}} \prod_{h=1}^{m_n} [1 - \Phi\{\alpha_h(\mathbf{x})\}] > \epsilon \right) \\ &\leq \Pi_{\mathcal{X}} \left( \prod_{h=m_n+1}^{m_n} \{1 - \Phi(\alpha'_h)\} > \epsilon \right) = \Pi_{\mathcal{X}} \left( - \sum_{h=m_n+1}^{m_n} \log\{1 - \Phi(\alpha'_h)\} < t_0 \right). \end{aligned}$$

Note that if we had  $\alpha_h(\mathbf{x}) \equiv \alpha_h \sim \text{N}(0, 1)$ , then the right hand side above equals

$$\Pi_{\mathcal{X}} \left( - \sum_{h=1}^{m_n} \log\{1 - \Phi(\alpha_h)\} < t_0 \right) = \Pi_{\mathcal{X}}(\Lambda_h < t_0)$$

where  $\Lambda_h \sim \text{Ga}(m_n, 1)$ . Then its easy to show that  $\Pi_{\mathcal{X}}(\Lambda_h < t_0) \lesssim e^{-m_n \log m_n}$ . However, the calculation gets complicated when  $\alpha_h$ '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  $\gamma_0$ , depending on  $\epsilon$  and  $\tau$  but not depending

on  $n$ , such that

$$\begin{aligned} \Pi_{\mathcal{X}}\left(\sum_{h=m_n^\eta+1}^{m_n} W_h < t_0\right) &\leq \xi(\delta_n)^{m_n-m_n^\eta} \Pi_{\mathcal{X}}\left(\sum_{h=m_n^\eta+1}^{m_n} Z_h < t_0\right) \\ &+ \sum_{h=m_n^\eta+1}^{m_n} P(A_h > \delta_n). \end{aligned} \quad (6.9)$$

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

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  $\tau$  as  $\tau'$ . Hence without loss of generality we assume  $\tau < 1$ .

Define  $g : [0, t_0] \rightarrow \mathfrak{R}, t \mapsto -\Phi^{-1}(1 - e^{-t})$ . 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  $h = m_n^\eta + 1, \dots, m_n$ ,

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

Below we estimate  $P(\sup_{\mathbf{x} \in \sqrt{\delta_n} \mathcal{X}} \alpha_0(\mathbf{x}) \geq \lambda)$  for large enough  $\lambda$  following Theorem 5.2 of [Adler \(1990\)](#). However extra care is required to identify the role of  $\delta_n$ . 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

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

for constants  $C_3, C_4 > 0$ . The last inequality holds for all large  $\lambda$  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

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

for any  $\gamma_0 > 1$ . Further choose  $\gamma_0$  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 (6.9) 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.  $\square$

The following theorem provides sufficient conditions for strong posterior consistency in the sup- $L_1$  topology.

**Theorem 6.11.** *Let  $\pi_h$ 's satisfy (6.1) 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 exists 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 last paragraph before Lemma 6.10.*

then  $f_0 \in KL(\Pi_{\mathcal{X}})$  implies that  $\Pi_{\mathcal{X}}$  achieves strong posterior consistency in sup- $L_1$  topology at  $f_0$ .

*Proof.* We will verify the sufficient conditions 2, 3 and 4 of Theorem 6.8. First we describe the construction of a sequence of sieves  $\mathcal{F}_n$ . Assume  $\epsilon > 0$  be 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 sequences  $M_n$  and  $m_n$  to be chosen later, let  $\delta_n = K_1 \epsilon / (M_n m_n^2)$ . Construct a sequence of subsets  $\{B_{h,n}, h = 1, \dots, m_n\}$  of  $\mathbb{C}[0, 1]^p$  as follows.

$$B_{h,n} = \begin{cases} (M_n \sqrt{r_n / \delta_n} \mathbb{H}_1^{r_n} + \frac{\epsilon}{m_n^2} \mathbb{B}_1) \cup (\cup_{a < \delta_n} M_n \mathbb{H}_1^a + \frac{\epsilon}{m_n^2} \mathbb{B}_1), & \text{if } h = 1, \dots, m_n^\eta \\ \cup_{a < \delta_n} M_n \mathbb{H}_1^a + \frac{\epsilon}{m_n^2} \mathbb{B}_1, & \text{if } h = m_n^\eta + 1, \dots, m_n. \end{cases}$$

Consider the sequence of sieves defined by

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

We will first show that given  $\xi > 0$ , there exists  $c_1, c_2 > 0$  and sequences  $m_n$  and  $M_n$ , such that  $\Pi_{\mathcal{X}}(\mathcal{F}_n^c) \leq c_1 e^{-nc_2}$  and  $\log N(\delta, \mathcal{F}_n, d_{SS}) < n\xi$ .

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

$$\begin{aligned} \|f_1(\cdot | \mathbf{x}) - f_2(\cdot | \mathbf{x})\|_1 &\leq \int_{\mathcal{Y}} \sum_{h=1}^{m_n} \pi_h^{(1)}(\mathbf{x}) \left| \phi_{\beta_h^{(1)}, \sigma_h^{(1)}}(\mathbf{x}, y) - \phi_{\beta_h^{(2)}, \sigma_h^{(2)}}(\mathbf{x}, y) \right| dy \\ &\quad + \int_{\mathcal{Y}} \sum_{h=1}^{m_n} \left| \pi_h^{(1)}(\mathbf{x}) - \pi_h^{(2)}(\mathbf{x}) \right| \phi_{\beta_h^{(2)}, \sigma_h^{(2)}}(\mathbf{x}, y) dy \\ &\quad + \sum_{h=m_n+1}^{\infty} \{ \pi_h^{(1)}(\mathbf{x}) + \pi_h^{(2)}(\mathbf{x}) \} \\ &\leq \sum_{h=1}^{m_n} \pi_h(\mathbf{x}) \left\{ \left( \frac{2}{\pi} \right)^{1/2} \frac{\| \beta_h^{(2)} - \beta_h^{(1)} \| \sqrt{p}}{\sigma_h^{(2)}} + \frac{3(\sigma_h^{(2)} - \sigma_h^{(1)})}{\sigma_h^{(1)}} \right\} \\ &\quad + \sum_{h=1}^{m_n} \left\| \pi_h^{(1)} - \pi_h^{(2)} \right\|_{\infty} + 2\epsilon. \end{aligned}$$

Let  $\Theta_{\pi, n} = \{ \pi^{m_n} = (\pi_1, \pi_2, \dots, \pi_{m_n}) : \alpha_h \in B_{h, n}, h = 1, \dots, m_n \}$ . 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

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

The above fact together with the proof of Lemma B.1 show that if we can make  $\| \alpha_{h,1} - \alpha_{h,2} \|_{\infty} < \frac{\epsilon}{m_n^2}, h = 1, \dots, m_n$ , 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 van der Vaart and van Zanten (2009) it follows that for  $h = 1, \dots, m_n^{\eta}$  and for sufficiently large  $M_n, r_n$ ,

$$\begin{aligned} \log N(2\epsilon/m_n^2, B_{h,n}, \|\cdot\|_{\infty}) &\leq K_3 r_n^p \log \left( \frac{M_n m_n^2 \sqrt{r_n/\delta_n}}{\epsilon} \right)^{p+1} + \\ &\quad 2 \log \frac{K_4 M_n m_n^2}{\epsilon}. \end{aligned} \tag{6.10}$$

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  $h = 1, \dots, m_n^{\eta}$ ,

$$P(\alpha_h \notin B_{h,n}) \leq P(A_h > r_n) + e^{-M_n^2/2}. \tag{6.11}$$

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

$$\log N(3\epsilon/m_n^2, B_{h,n}, \|\cdot\|_{\infty}) \leq 2 \log \frac{K_4 M_n m_n^2}{\epsilon}. \tag{6.12}$$

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

$$\begin{aligned} P(\alpha_h \notin B_{h,n}) &\leq P(A_h > \delta_n) + \int_{a=0}^{\delta_n} P(\alpha_h \notin B_{h,n} \mid A_h = a) g_{A_h}(a) da \\ &\leq P(A_h > \delta_n) + \int_{a=0}^{\delta_n} P(\alpha_h \notin M_n \mathbb{H}_1^a + \epsilon \mathbb{B}_1 \mid A_h = a) g_{A_h}(a) da \\ &\leq P(A_h > \delta_n) + (1 - \Phi(\Phi^{-1}(e^{-\phi_0^{\delta_n}(\epsilon/m_n^2)} + M_n))). \end{aligned}$$

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  $\phi_0^{\delta_n}(\epsilon/m_n^2) \leq -\log P(|W_0| \leq \epsilon/m_n^2) = \frac{K_6}{m_n^2}$  for some constant  $K_6 > 0$ . Hence if  $M_n \geq K_7/m_n$  for some  $K_7 > 0$ , then it follows from the proof of Theorem 3.1 in van der Vaart and van Zanten (2009) that

$$P(\alpha_h \notin B_{h,n}) \leq P(A_h > \delta_n) + e^{-M_n^2/2}. \quad (6.13)$$

From (6.10) and (6.12),

$$\begin{aligned} \log(N(\epsilon, B_{1,n} \times \dots \times B_{m_n,n}, \|\cdot\|_\infty)) &\leq 2m_n \log \frac{K_4 M_n m_n^2}{\epsilon} + \\ &\quad m_n^\eta r_n^p \log \left( \frac{M_n m_n^2 \sqrt{r_n/\delta_n}}{\epsilon} \right)^{p+1}. \end{aligned} \quad (6.14)$$

Also from (6.11) and (6.13),

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

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  $\xi_0$ . By assumption C1, we have

$$\Pi_{\mathcal{X}}(\Theta_{a_n, h_n, l_n}^c) \lesssim m_n O(e^{-n}) \lesssim O(e^{-n}). \quad (6.15)$$

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

With  $m_n = \frac{n}{\log n}$ ,  $m_n \log m_n > \frac{n}{2}$  for large enough  $n$  and it follows from Lemma 6.10 that

$$\Pi_{\mathcal{X}} \left( \sup_{\mathbf{x} \in \mathcal{X}} \sum_{h=m_n+1}^{\infty} \pi_h(\mathbf{x}) > \epsilon \right) \lesssim O(e^{-n/2}). \quad (6.16)$$

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

$$\sum_{h=1}^{m_n} P(\alpha_h \notin B_{h,n}) \lesssim e^{-n}. \quad (6.17)$$

(6.15), (6.16) and (6.17) together imply that  $\Pi_{\mathcal{X}}(\mathcal{F}_n^c) \lesssim O(e^{-n})$ .

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

$$\log(N(\epsilon, B_{1,n} \times \cdots \times B_{m_n,n}, \|\cdot\|_\infty)) < n\xi \quad (6.18)$$

for any  $\xi > 0$ . Also from Lemma 6.9,

$$\begin{aligned} m_n \log N(\Theta_{a_n, h_n, l_n}, \epsilon, \|\cdot\|_1) &\leq m_n \log \left\{ d_1 \left( \frac{a_n}{l_n} \right)^p + d_2 \log \frac{h_n}{l_n} + 1 \right\} \\ &< n\xi \end{aligned} \quad (6.19)$$

for any  $\xi > 0$ . Combining (6.18) and (6.19),  $\log N(\mathcal{F}_n, 4\epsilon, d_{SS}) < n\xi$  for any  $\xi > 0$ .

Next we turn to proving Theorem 6.11. Let  $\delta > 0$  be given. Let  $SS(f_0)^c \cap \mathcal{F}_n \subset \cup_{k=1}^{N_\delta} G_k$  such that  $f_1, f_2 \in \mathcal{F}_d$ ,  $d_{SS}(f_1, f_2) < \delta$ . Fix  $f_k \in SS(f_0)^c \cap \mathcal{F}_n \cap G_k$ ,  $k = 1, \dots, N_\delta$ . We have,  $d_{SS}(f_0, f_k) \geq \epsilon$  for  $k = 1, \dots, N_\delta$ . Hence there exists  $C_k \subset \mathcal{X}$  such that  $\|f_0(\cdot | \mathbf{x}) - f_k(\cdot | \mathbf{x})\|_1 \geq \epsilon/2$  for  $k = 1, \dots, N_\delta$ . Thus if  $f \in G_k$  and  $\mathbf{x} \in C_k$ ,  $\|f_0(\cdot | \mathbf{x}) - f(\cdot | \mathbf{x})\|_1 \geq \|f_0(\cdot | \mathbf{x}) - f_k(\cdot | \mathbf{x})\|_1 - d_{SS}(f_k, f) \geq \epsilon/2 - \delta$ . For each  $k = 1, \dots, N_\delta$ , consider the set  $A_k = \{(\mathbf{x}, y), \mathbf{x} \in C_k, f_k(y | \mathbf{x}) > f_0(y | \mathbf{x})\}$  and  $A_k^\mathbf{x}$  be the  $\mathbf{x}$ -section of  $A_k$ . Observe that for each  $\mathbf{x} \in C_k$ ,  $\int_{C_k} \int_{A_k^\mathbf{x}} f_k(y | \mathbf{x}) q(\mathbf{x}) dy d\mathbf{x} \geq$

$$\begin{aligned} \gamma_k &= \int_{C_k} \int_{A_k^\mathbf{x}} \{f_k(y | \mathbf{x}) - f_0(y | \mathbf{x})\} dy q(\mathbf{x}) d\mathbf{x} + \int_{C_k} \int_{A_k^\mathbf{x}} f_0(y | \mathbf{x}) q(\mathbf{x}) dy d\mathbf{x} \\ &= (1/2) \int_{C_k} \|f_k(\cdot | \mathbf{x}) - f_0(\cdot | \mathbf{x})\|_1 q(\mathbf{x}) d\mathbf{x} + \alpha_k \\ &\geq \epsilon/4 + \alpha_k, \end{aligned}$$

where  $\alpha_k = \int_{C_k} \int_{A_k^\mathbf{x}} f_0(y | \mathbf{x}) q(\mathbf{x}) dy d\mathbf{x}$ . Hence if  $f \in G_k$ ,

$$\left| \int_{C_k} \int_{A_k^\mathbf{x}} \{f_k(y | \mathbf{x}) - f(y | \mathbf{x})\} q(\mathbf{x}) dy d\mathbf{x} \right| \leq d_{SS}(f, f_k) \leq \delta$$

implying  $\int_{C_k} \int_{A_k^\mathbf{x}} f(y | \mathbf{x}) q(\mathbf{x}) dy d\mathbf{x} \geq \alpha_k + \epsilon/4 - \delta$ .

Let

$$B_k = \left\{ \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\} : \frac{1}{n} \sum_{j=1}^n I_{A_k}(\mathbf{x}_j, y_j) \geq \frac{\gamma_k + \alpha_k}{2} \right\}.$$

A straightforward application of Hoeffding's inequality implies

$$\int_{B_k} f_0(y | \mathbf{x}) q(\mathbf{x}) dy d\mathbf{x} \leq \exp\{-n\epsilon^2/32\}.$$

By a similar application as in Ghosal, Ghosh and Ramamoorthi (1999), if  $f \in G_k$ ,

$$\int_{B_k} f(y | \mathbf{x})q(\mathbf{x})dyd\mathbf{x} \geq 1 - \exp\{-2n(\epsilon/8 - \delta)^2\}.$$

If we set

$$\Phi_n = \max_{1 \leq k \leq N_\delta} I_{B_k}\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\},$$

the conditions 3 and 4 are satisfied as  $\Pi_{\mathcal{X}}(\mathcal{F}_n^c) \lesssim O(e^{-n})$  and  $\log N(\mathcal{F}_n, 4\epsilon, d_{SS}) < n\xi$  for any  $\xi > 0$ . An application of Borel-Cantelli Lemma also guarantees 2.  $\square$

**Corollary 6.12.** *The above conditions for consistency in the sup- $L_1$  topology automatically guarantees the same in the  $\nu$ -integrated  $L_1$  topology defined in (5.2).*

**Remark 6.13.** *Verification of condition C1 of Theorem 6.11 is particularly simple. For example, if  $G_0$  is a product of multivariate normals on  $\beta$  and an inverse Gamma prior on  $\sigma^2$ , 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 van der Vaart and van Zanten (2009) that  $f_0 \in KL(\Pi_{\mathcal{X}})$  is still satisfied when we have the additional assumptions C1-C2 together with S1-S3 on the prior  $\Pi_{\mathcal{X}}$ .*

**Remark 6.14.** *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  $B_h^d \sim \text{Exp}(\lambda)$  and  $c_h = (h^{(3d+2)/\eta} \log h)^{-d}$  for any  $0 < \eta < 1$ . Then  $P(A_h > n^{\eta_1/p}) \leq P(B_h^d > n^{\eta_1/p}) \leq e^{-n^{d\eta_1/p}} \leq e^{-n}$  and  $P(A_h > (\log n)^2 n^{-5/2}) \leq \exp\{-n^{-3d} h^{(3d+2)/\eta} \log h\}$ .*

**Remark 6.15.** *If we instead work with the simpler truncated version of the prior for  $A_h$  instead of the one used in the proof of Theorem 6.11, we would only need  $r_n^p n^\eta (\log n)^{p+1} = o(n)$ ,  $P(A_h > r_n) \leq e^{-n}$  and  $P(L > m_n^\eta) \leq e^{-n}$  for some  $0 < \eta < 1$ . This can be achieved by letting  $r_n^p = n^{\eta_1}$  for some  $\eta_1$ ,  $A_h^d \sim \text{Ga}(a, b)$  where  $d\eta_1/p \geq 1$  and choosing  $L$  such that  $P(L > n) \leq \exp\{-n^{1/\eta} \log n\}$ .*

**Remark 6.16.** *The theory of strong posterior consistency can be generalized to an arbitrary monotone differentiable link function  $L : \mathfrak{R} \mapsto [0, 1]$  which is Lipschitz, i.e., there exists a constant  $K > 0$  such that  $|L(\mathbf{x}) - L(\mathbf{x}')| \leq K |\mathbf{x} - \mathbf{x}'|$  for all  $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ .*

**Remark 6.17.** *In applications like brain imaging, the covariates can be thought to lie on a more complicated space such as a compact manifold or a compact metric space e.g., the sphere. As long as the compact manifold or the compact metric space is embedded in  $\mathfrak{R}^p$  for some  $p$ , the results can be extended to the above case by extending the definition of a stochastic process by the usual embedding theorem.*

## 7. Posterior consistency in Gaussian mixture of fixed- $\pi$ dependent processes

### 7.1. Kullback-Leibler property

Once again we approximate  $f_0(y | \mathbf{x})$  by  $\tilde{f}(y | \mathbf{x}) = \int \frac{1}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right) d\tilde{G}_{\mathbf{x}}(\mu, \sigma)$ , so that the first term of 6.2 is arbitrarily small. We construct such an  $\tilde{f}$  in Theorem 7.3 which is analogous to Theorem 6.6. Lemma 7.1 is a variant of Lemma 6.1 which ensures that the second term in (6.2) 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.

$$\left\{ \{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} : \sup_{\mathbf{x} \in \mathcal{X}} \left| \int_{\mathfrak{R} \times \mathfrak{R}^+} \{g(\mu, \sigma) dG_{\mathbf{x}}(\beta, \sigma) - g(\mu, \sigma) dF_{\mathbf{x}}(\mu, \sigma)\} \right| < \epsilon \right\}. \quad (7.1)$$

**Lemma 7.1.** *Assume that  $f_0 \in \mathcal{F}_d$  satisfies  $\int_{\mathcal{X}} \int_{\mathcal{Y}} y^2 f_0(y | \mathbf{x}) dy q(\mathbf{x}) d\mathbf{x} < \infty$ . Suppose  $\tilde{f}(y | \mathbf{x}) = \int \frac{1}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right) d\tilde{G}_{\mathbf{x}}(\mu, \sigma)$ , where  $\exists a > 0$  and  $0 < \underline{\sigma} < \bar{\sigma}$  such that*

$$\tilde{G}_{\mathbf{x}}([-a, a] \times (\underline{\sigma}, \bar{\sigma})) = 1 \quad \forall \mathbf{x} \in \mathcal{X}, \quad (7.2)$$

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 (7.1) such that for any conditional density  $f(y | \mathbf{x}) = \int \frac{1}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right) dG_{\mathbf{x}}(\mu, \sigma)$ ,  $\mathbf{x} \in \mathcal{X}$ , with  $\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} \in W$ ,

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} f_0(y | \mathbf{x}) \log \frac{\tilde{f}(y | \mathbf{x})}{f(y | \mathbf{x})} dy q(\mathbf{x}) d\mathbf{x} < \epsilon. \quad (7.3)$$

The proof of Lemma 7.1 is similar to that of Lemma 6.1 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  $(\mathfrak{R} \times \mathfrak{R}^+, \mathcal{B}(\mathfrak{R} \times \mathfrak{R}^+))$  denoted by  $\mathcal{G}_{\mathcal{X}}^{**}$  satisfying  $\mathbf{x} \mapsto \int_{\mathfrak{R} \times \mathfrak{R}^+} g(\mu, \sigma) dF_{\mathbf{x}}(\mu)$  is a continuous function of  $\mathbf{x}$  for all bounded uniformly continuous functions  $g : \mathfrak{R} \times \mathfrak{R}^+ \rightarrow [0, 1]$ .

**Theorem 7.2.** *Assume the following holds.*

- T1.  $G_0$  is specified by  $\mu_h \sim GP(\mu, c)$ ,  $\sigma_h \sim G_{0, \sigma}$  where  $c$  is chosen so that  $GP(0, c)$  has continuous path realizations and  $\Pi_{\sigma}$  is absolutely continuous w.r.t. Lebesgue measure on  $\mathfrak{R}^+$ .
- T2. For every  $k \geq 2$ ,  $(\pi_1, \dots, \pi_k)$  is absolutely continuous w.r.t. to the Lebesgue measure on  $S_{k-1}$ .
- T3. For any continuous function  $g : \mathcal{X} \mapsto \mathfrak{R}$ ,

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

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

Then for a bounded uniformly continuous function  $g : \mathfrak{R} \times \mathfrak{R}^+ : [0, 1]$  satisfying  $g(\mu, \sigma) \rightarrow 0$  as  $|\mu| \rightarrow \infty, \sigma \rightarrow \infty$ ,

$$\mathcal{P}_{\mathcal{X}} \left\{ \left\{ G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X} \right\} : \sup_{\mathbf{x} \in \mathcal{X}} \left| \int_{\mathfrak{R} \times \mathfrak{R}^+} \{g(\mu, \sigma) dG_{\mathbf{x}}(\beta, \sigma) - g(\mu, \sigma) dF_{\mathbf{x}}(\mu, \sigma)\} \right| < \epsilon \right\} > 0. \quad (7.4)$$

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

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

Define  $\Omega_2 = \{\sup_{\mathbf{x} \in \mathcal{X}} |g(\mu_k(\mathbf{x}), \sigma_k) - \psi(\mathbf{x})| < \epsilon, k = 1, \dots, n_\epsilon\}$ . For a fixed  $\sigma_k$ , there exists a  $\delta$  such that  $\sup_{\mathbf{x} \in \mathcal{X}} |g(\mu_k(\mathbf{x}), \sigma_k) - \psi(\mathbf{x})| < \epsilon$  if  $\sup_{\mathbf{x} \in \mathcal{X}} |\mu_k(\mathbf{x}) - g_{\sigma_k}^{-1} \psi(\mathbf{x})| < \delta$  where  $g_{\sigma_k}^{-1}$  denotes the inverse of  $g(\cdot, \sigma_k)$  for fixed  $\sigma_k$ . Hence there exists a neighborhood  $B_k$  of  $\sigma_k$  such that for  $\sigma_k \in B_k$  and  $\sup_{\mathbf{x} \in \mathcal{X}} |\mu_k(\mathbf{x}) - g_{\sigma_k}^{-1} \psi(\mathbf{x})| < 2\delta$ , we have  $\sup_{\mathbf{x} \in \mathcal{X}} |g(\mu_k(\mathbf{x}), \sigma_k) - \psi(\mathbf{x})| < \epsilon$ . Since for each  $k = 1, \dots, n_\epsilon$ ,  $\mathcal{P}_{\mathcal{X}}\{\sigma_k \in B_k, \sup_{\mathbf{x} \in \mathcal{X}} |\mu_k(\mathbf{x}) - g_{\sigma_k}^{-1} \psi(\mathbf{x})| < 2\delta\} =$

$$\int_{\sigma_k \in B_k} \mathcal{P}_{\mathcal{X}} \left\{ \sup_{\mathbf{x} \in \mathcal{X}} |\mu_k(\mathbf{x}) - g_{\sigma_k}^{-1} \psi(\mathbf{x})| < 2\delta \right\} dG_{0, \sigma}(\sigma_k) > 0,$$

$\mathcal{P}_{\mathcal{X}}(\Omega_2) > 0$ . The conclusion of the theorem follows from the independence of  $\Omega_1$  and  $\Omega_2$ .  $\square$

The following theorem verifies that  $\Pi_{\mathcal{X}}$  has KL property at  $f_0 \in \mathcal{F}_d^*$ . The proof of Theorem 7.3 can be found in Appendix F.

**Theorem 7.3.**  $f_0 \in KL(\Pi_{\mathcal{X}})$  for each  $f_0$  in  $\mathcal{F}_d^*$  if  $\mathcal{P}_{\mathcal{X}}$  satisfies T1-T3.

## 7.2. Strong consistency with the sup- $L_1$ neighborhood

As before we establish sup- $L_1$  consistency of Gaussian mixtures of fixed- $\pi$  dependent processes by verifying the conditions of Theorem 6.8. 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  $\mathbf{x} \in \mathcal{X}$ . From Lemma 4.1 of Tokdar (2006), we obtain for  $\sigma_2 > \sigma_1 > \frac{\sigma_2^2}{2}$  and for each  $\mathbf{x} \in \mathcal{X}$ ,

$$\int_{\mathcal{Y}} |\phi_{\mu_1, \sigma_1}(\mathbf{x}, y) - \phi_{\mu_2, \sigma_2}(\mathbf{x}, y)| dy \leq \left(\frac{2}{\pi}\right)^{1/2} \frac{\|\mu_1 - \mu_2\|_{\infty}}{\sigma_2} + \frac{3(\sigma_2 - \sigma_1)}{\sigma_1}$$

Let  $\mu_h(\mathbf{x}) = \mathbf{x}' \beta_h + \eta_h(\mathbf{x})$ ,  $h = 1, 2, \dots$ ,  $\beta_h \sim G_{\beta}$  where  $G_{\beta}$  is a probability distribution on  $\mathfrak{R}^p$ . Let  $\eta_h \sim GP(0, c)$  independently where  $c(\mathbf{x}, \mathbf{x}') = \tau^2 e^{-A \|\mathbf{x} - \mathbf{x}'\|^2}$ ,

where  $A$  is a distributed with support  $\mathfrak{R}^+$  and  $\tau^2$  is fixed. Assume that  $\sigma_h \sim G_{0,\sigma}$  where  $G_{0,\sigma}$  is a distribution on  $\mathfrak{R}^+$ . Here  $G_{0\mathbf{x}}$  is a distribution on  $\mathfrak{R} \times \mathfrak{R}^+$  induced from the distribution of  $(\mu_h(\mathbf{x}), \sigma_h^2)$ . For any pair  $\mu_1, \mu_2$ ,

$$\|\mu_1 - \mu_2\|_\infty \leq \left(\frac{2}{\pi}\right)^{1/2} \frac{\|\beta_1 - \beta_2\| \sqrt{p} + \|\eta_1 - \eta_2\|_\infty}{\sigma_2}.$$

As before, 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 sequences  $M_n \uparrow \infty, l_n \downarrow 0, r_n \uparrow \infty$  to be determined later and given  $\epsilon > 0$  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

$$\Theta_n = \{ \phi_{\mu,\sigma} : \|\beta\| \leq a_n, \eta \in B_n, l_n \leq \sigma \leq h_n \}. \quad (7.5)$$

In the following Lemma, we provide an upper bound to  $N(\Theta_n, \epsilon, \|\cdot\|_1)$ . The Lemma 7.4 is proved in Appendix G.

**Lemma 7.4.** *There exists constants  $d_1, d_2, K_2$  and  $K_3 > 0$  such that for  $M_n \sqrt{r_n/\delta_n} > 2\epsilon$  and for sufficiently large  $r_n$*

$$\begin{aligned} \log N(\Theta_n, \epsilon, d_{SS}) &\leq K_2 r_n^p \left\{ \log \left( \frac{8\sqrt{2} M_n \sqrt{r_n/\delta_n}}{\epsilon \sqrt{\pi} l_n} \right) \right\}^{p+1} + \log \frac{K_3 M_n}{\epsilon l_n} + \\ &\quad \log \left\{ d_1 \left( \frac{a_n}{l_n} \right)^p + d_2 \log \frac{h_n}{l_n} + 1 \right\}. \end{aligned}$$

Next we summarize the consistency theorem with respect to the sup- $L_1$  topology. The proof of Theorem 7.5 is provided in Appendix H.

**Theorem 7.5.** *Let  $\mu_h(\mathbf{x}) = \mathbf{x}'\beta_h + \eta_h(\mathbf{x}), \beta_h \sim G_\beta$  and  $\eta_h \sim GP(0, c)$ ,  $h = 1, \dots, \infty$  where  $c(\mathbf{x}, \mathbf{x}') = \tau^2 e^{-A\|\mathbf{x}-\mathbf{x}'\|^2}$ ,  $A^{p(1+\eta_2)/\eta_2} \sim Ga(a, b)$  for some  $\eta_2 > 0$ .*

*F1. There exists 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  $G_\beta \{B(0; a_n)\}^c < d_1 e^{-d_2 n}$  and  $G_{0,\sigma} \{[l_n, h_n]\}^c \leq d_3 e^{-d_4 n}$ .*

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

*then  $f_0 \in KL(\Pi_{\mathcal{X}})$  implies that  $\Pi_{\mathcal{X}}$  achieves strong posterior consistency at  $f_0$  with respect to the sup- $L_1$  topology.*

**Remark 7.6.** *Corollary 6.12 and Remark 6.17 also apply to this case with condition F1 of Theorem 7.5 similarly verified as condition T1 of Theorem 6.11.*

**Remark 7.7.** *F2 is satisfied if  $\pi_h$ 's are made to decay more rapidly than the usual Beta(1,  $\alpha$ ) stick-breaking random variables, e.g. if  $\pi_h = \nu_h \prod_{l < h} (1 - \nu_l)$  and if  $\nu_h \sim \text{Beta}(1, \alpha_h)$  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  $\alpha_h$  for the higher indexed weights favors smaller number of components.*

## 8. Discussion

We have provided sufficient conditions to show posterior consistency in estimating the conditional density via probit stick-breaking mixtures of Gaussians and the fixed- $\pi$  dependent processes. 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.

Although we have focused on the case where  $\mathcal{X}$  is a compact subset of  $\mathbb{R}^p$ , we can transform a non-compact space to a compact one by suitable transformation. In doing so, we need to use a non-stationary covariance kernel in the GP transforms. However the computations get somewhat complicated and is beyond the scope of the current article. An interested reader can refer to Tokdar, Zhu and Ghosh (2010) for details.

One alternative possibility is to specify a prior for the joint density  $h(\mathbf{x}, y) = q(\mathbf{x})f(y | \mathbf{x})$ , to induce a prior on the conditional  $f(y | \mathbf{x})$ , where  $q(\mathbf{x})$  denotes the joint density of the covariates. Using such an approach, which was originally proposed by Müller, Erkanli and West (1996) using Dirichlet process mixtures of multivariate Gaussians, one can potentially rely on the theory of large support and posterior consistency for i.i.d. realizations from a multivariate distribution; for example, refer to Wu and Ghosal (2010); Norets and Pelenis (2009). Unfortunately, such an approach has clear disadvantages. When interest focuses on the conditional distribution of  $f(y | \mathbf{x})$  it is very appealing to avoid needing to model the joint density of the predictors,  $q(\mathbf{x})$ , which will be multivariate in typical applications. In addition, standard models for the joint distribution relying on multivariate Dirichlet process mixtures (refer also to Shahbaba and Neal (2009); Park and Dunson (2009)), can have relatively poor performance, because many mixture components may be introduced primarily to provide a good fit to the marginal  $q(\mathbf{x})$ , potentially leading to degradation of performance in estimating  $f(y | \mathbf{x})$  for all  $\mathbf{x} \in \mathcal{X}$ . The MGLR $_{\mathbf{x}}$  and the Gaussian mixture of fixed- $\pi$  dependent processes are examples of priors directly on the conditional densities.

Although, a more reasonable way of evaluating a Bayes procedure is to study the posterior convergence rates, deriving the rates of convergence in our case substantially complicates the analysis and is a topic of future research. Of course our sieve construction can be used to derive the rates, while being more careful in estimating the concentration of the prior around the true density, the rates of decay of the complement of the sieve and calculating the entropy.

### 9. Acknowledgements

The authors would also like to thank the Associate Editor and three referees for their insightful comments in an earlier version of the paper. This work was supported by Award Number R01ES017240 from the National Institute of Environmental Health Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences or the National Institutes of Health.

### Appendix A: Proof of Lemma 6.1

The proof proceeds similarly to that for Theorem 3 in Ghosal, Ghosh and Ramamoorthi (1999). Note that  $\{\tilde{G}_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} \in \mathcal{G}_{\mathcal{X}}^*$ . Let  $B = ([-a, a]^p \times (\underline{\sigma}, \bar{\sigma}))$ . Choose  $k > pa + \bar{\sigma}$  such that

$$\int_{\mathcal{X}} \int_{|y|>k} f_0(y | \mathbf{x}) \left\{ \frac{|y| + pa}{2\sigma^2} \right\}^2 dy q(\mathbf{x}) d\mathbf{x} < \frac{\epsilon}{2}.$$

Take  $V = \{G_{\mathbf{x}} | \mathbf{x} \in \mathcal{X}\} : \inf_{\mathbf{x} \in \mathcal{X}} G_{\mathbf{x}}(B) > \frac{\sigma}{\bar{\sigma}}\}$ . By approximating  $1_B$  by a bounded continuous function, we can show that  $V$  contains a neighborhood  $V'$  of  $\{\tilde{G}_{\mathbf{x}} | \mathbf{x} \in \mathcal{X}\}$  of the type (6.5). For any density  $f \in \mathcal{F}_d$ ,  $f(y | \mathbf{x}) = \int \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)$ ,  $\mathbf{x} \in \mathcal{X}$ , with  $\{G_{\mathbf{x}} | \mathbf{x} \in \mathcal{X}\} \in V'$ ,

$$\begin{aligned} & \int_{\mathcal{X}} \int_{|y|>k} f_0(y | \mathbf{x}) \log \left\{ \frac{\tilde{f}(y | \mathbf{x})}{f(y | \mathbf{x})} \right\} dy q(\mathbf{x}) d\mathbf{x} \\ & \leq \int_{\mathcal{X}} \int_{|y|>k} f_0(y | \mathbf{x}) \log \frac{\int_B \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) d\tilde{G}_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)}{\int_B \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)} dy q(\mathbf{x}) d\mathbf{x} \\ & \leq \int_{\mathcal{X}} \int_{|y|>k} f_0(y | \mathbf{x}) \log \frac{\frac{1}{\sigma} \phi\left(\frac{|y| - pa}{\sigma}\right)}{\frac{1}{\sigma} \phi\left(\frac{|y| + pa}{\sigma}\right) G_{\mathbf{x}}(B)} dy q(\mathbf{x}) d\mathbf{x} \\ & \leq \int_{\mathcal{X}} \int_{|y|>k} f_0(y | \mathbf{x}) \left\{ \frac{|y| + pa}{2\sigma^2} \right\}^2 dy q(\mathbf{x}) d\mathbf{x} < \frac{\epsilon}{2}. \end{aligned}$$

Let  $\inf_{\{y: |y| \leq k\} \times \mathcal{X}} \inf_{(\boldsymbol{\beta}, \sigma) \in B} \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) = c$ . Consider the uniformly equi-continuous family of functions

$$\left\{ g_{y, \mathbf{x}} : g_{y, \mathbf{x}} : \mathbb{R}^p \times \mathbb{R}^+ \rightarrow \mathbb{R}, (\boldsymbol{\beta}, \sigma) \mapsto \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right), (y, \mathbf{x}) \in [-k, k] \times \mathcal{X} \right\}.$$

By the Arzela-Ascoli theorem, given  $\delta > 0$ , there exists finitely many points  $\{(y_i, \mathbf{x}_i) \in [-k, k] \times \mathcal{X}, i = 1, \dots, m\}$  such that for any  $(y, \mathbf{x}) \in [-k, k] \times \mathcal{X}$ ,  $\exists i$  such that

$$\sup_{(\boldsymbol{\beta}, \sigma) \in B} |g_{y, \mathbf{x}}(\boldsymbol{\beta}, \sigma) - g_{y_i, \mathbf{x}_i}(\boldsymbol{\beta}, \sigma)| < c\delta.$$

$$\text{Let } g_{y_i, \mathbf{x}_i} * G_{\mathbf{x}} = \int \frac{1}{\sigma} \int_{\mathbf{x} \in \mathcal{X}} \phi\left(\frac{y_i - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma).$$

$$E = \left\{ \{G_{\mathbf{x}} \mid \mathbf{x} \in \mathcal{X}\} : \sup_{\mathbf{x} \in \mathcal{X}} |g_{y_i, \mathbf{x}_i} * G_{\mathbf{x}} - g_{y_i, \mathbf{x}_i} * \tilde{G}_{\mathbf{x}}| < c\delta, i = 1, 2, \dots, m \right\}.$$

It holds that  $E$  is a neighborhood of  $\{\tilde{G}_{\mathbf{x}} \mid \mathbf{x} \in \mathcal{X}\}$  formed by finite intersections of sets of the type (6.5) and for  $\{G_{\mathbf{x}} \mid \mathbf{x} \in \mathcal{X}\} \in E$  and  $(y, \mathbf{x}) \in [-k, k] \times \mathcal{X}$ ,

$$\left| \frac{\int \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) d\tilde{G}_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)}{\int \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)} - 1 \right| < \frac{3\delta}{1 - 3\delta}.$$

for  $\delta < \frac{1}{3}$ . Thus given any  $\epsilon > 0$ , there exists a neighborhood  $E$  of  $\{\tilde{G}_{\mathbf{x}} \mid \mathbf{x} \in \mathcal{X}\}$  such that for  $\{G_{\mathbf{x}} \mid \mathbf{x} \in \mathcal{X}\} \in E$  with  $f(y \mid \mathbf{x}) = \int \frac{1}{\sigma} \phi\left(\frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right) dG_{\mathbf{x}}(\boldsymbol{\beta}, \sigma)$ ,

$$\int_{\mathcal{X}} \int_{y: |y| \leq k} f_0(y \mid \mathbf{x}) \log \left\{ \frac{\tilde{f}(y \mid \mathbf{x})}{f(y \mid \mathbf{x})} \right\} dy q(\mathbf{x}) d\mathbf{x} < \frac{\epsilon}{2}. \quad (\text{A.1})$$

Taking  $W = V' \cap E$  and since  $W$  is a finite intersection of neighborhoods of  $\tilde{G}_{\mathbf{x}}$  of the type (6.5), the result follows immediately.  $\square$

## Appendix B: A useful lemma

**Lemma B.1.**  $\{\pi_h(\mathbf{x}), h = 1, \dots, \infty\}$  constructed as in (6.1) satisfies S1 and S2 if  $c_h$  is chosen so that  $GP(0, c_h)$  has continuous path realizations and for any continuous function  $g : \mathcal{X} \mapsto \mathfrak{R}$ ,

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

for any  $\epsilon > 0$ ,  $h \geq 1$ .

*Proof.* Let  $\{A_i, i = 1, \dots, k\}$  be a measurable partition of  $\mathfrak{R}^p \times \mathfrak{R}^+$ . Without loss of generality, let  $0 < F_{\mathbf{x}}(A_i) < 1, i = 1, \dots, k \forall \mathbf{x} \in \mathcal{X}$ . We want to show that for any  $\epsilon_i > 0, i = 1, \dots, k$

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

Construct continuous functions  $g_i : \mathcal{X} \mapsto \mathfrak{R}, 0 < g_i(\mathbf{x}) < 1 \forall \mathbf{x} \in \mathcal{X}, i = 1, \dots, k-1$  such that

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

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

$$\mathcal{P}_{\mathcal{X}} \left\{ \sup_{\mathbf{x} \in \mathcal{X}} |\Phi\{\alpha_i(\mathbf{x})\} - g_i(\mathbf{x})| < \epsilon_i \right\} > 0 \quad (\text{B.2})$$

and for  $i = k$ ,

$$\mathcal{P}_{\mathcal{X}} \left\{ \sup_{\mathbf{x} \in \mathcal{X}} |\Phi\{\alpha_k(\mathbf{x})\} - 1| < \epsilon_k \right\} = \mathcal{P}_{\mathcal{X}} \left\{ \inf_{\mathbf{x} \in \mathcal{X}} \alpha_k(\mathbf{x}) > \Phi^{-1}(1 - \epsilon_k) \right\}. \quad (\text{B.3})$$

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  $p_i(\mathbf{x}) = \Phi\{\alpha_i(\mathbf{x})\}, g_i(\mathbf{x}) = F_{\mathbf{x}}(A_i), i = 1, \dots, k-1$  and  $g_k(\mathbf{x}) = 1$ . Let  $\mathbf{z} = (z_1, \dots, z_p)'$ ,  $f_i : S_{k-1} \rightarrow \mathfrak{R}, \mathbf{z} \mapsto z_i \prod_{l < i} (1 - z_l), i = 2, \dots, k$  and  $f_1(\mathbf{z}) = z_1$ . Let  $\mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), \dots, p_k(\mathbf{x}))$  and  $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \dots, g_k(\mathbf{x}))$ . Then we need to show that

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

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

$$\begin{aligned} \|f_i(\mathbf{p}) - f_i(\mathbf{g})\|_{\infty} &= \left\| p_i \left\{ 1 - \sum_{l < i} f_l(\mathbf{p}) \right\} - g_i \left\{ 1 - \sum_{l < i} f_l(\mathbf{g}) \right\} \right\|_{\infty} \\ &\leq (i-1) \|p_i - g_i\|_{\infty} + \sum_{l < i} \|f_l(\mathbf{p}) - f_l(\mathbf{g})\|_{\infty}. \end{aligned}$$

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

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

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

**Appendix C: Proof of Theorem 6.2**

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  $g : \mathbb{R}^p \times \mathbb{R}^+ \times \mathcal{X} \rightarrow [0, 1]$  and  $\epsilon > 0$ ,

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

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

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

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

where  $\pi_h$ 's are specified by 6.1 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  $\{A_{k,n}, k = 1, \dots, n\}$  be sequence of partitions of  $C$  with increasing refinement as  $n$  increases. Assume  $\max_{1 \leq k \leq n} \text{diam}(A_{k,n}) \rightarrow 0$  as  $n \uparrow \infty$ . Fix  $(\tilde{\boldsymbol{\beta}}_{k,n}, \tilde{\sigma}_{k,n}) \in A_{k,n}, k = 1, \dots, n$ . Then by DCT as  $n \rightarrow \infty$ ,

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

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

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

Consider the set

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

By Lemma B.1 which is proved in Appendix B,  $\mathcal{P}_{\mathcal{X}}(\Omega_1) > 0$ . It holds that  $\exists \Omega$  with  $\mathcal{P}_{\mathcal{X}}(\Omega) = 1$ , such that for each  $\omega = \{\pi_h, h = 1, \dots, \infty\} \in \Omega$ ,  $g_n(\mathbf{x}) = \sum_{h=1}^n \pi_h(\mathbf{x}) \rightarrow 1$  as  $n \rightarrow \infty$  for each  $\mathbf{x}$  in  $\mathcal{X}$ . Note that this convergence is uniform since,  $g_n(\cdot), n \geq 1$  are continuous functions defined on a compact set monotonically increasing to a continuous function identically equal to 1. Hence for each  $\omega = \{\pi_h, h = 1, \dots, \infty\} \in \Omega$ ,  $g_n(\mathbf{x}) \rightarrow 1$  uniformly in  $\mathbf{x}$ . By Egoroff's theorem, there exists a measurable subset  $\Omega_2$  of  $\Omega_1$  with  $\mathcal{P}_{\mathcal{X}}(\Omega_2) > 0$  such that within this subset  $g_n(\mathbf{x}) \rightarrow 1$  uniformly in  $\mathbf{x}$  and uniformly in  $\omega$  in  $\Omega_2$ . Thus there exists a positive integer  $n_\epsilon \geq n_1$  not depending on  $\mathbf{x}$  and  $\omega$ , such that  $\sum_{h=n_\epsilon+1}^{\infty} \pi_h(\mathbf{x}) < \epsilon$  on  $\Omega_2$ . Moreover, one can find a  $K > 0$  independent of  $\mathbf{x}$  such that  $g(\beta, \sigma, \mathbf{x}) < \epsilon$  if  $\|\beta\| > K$  and  $\sigma > K$ . Let  $A_1 = \{(\beta, \sigma) : \|\beta\| > K, \sigma > K\}$ . Let  $\Omega_3 = \Omega_2 \cap \{(\beta_{n_1+1}, \sigma_{n_1+1}) \in A_1, \dots, (\beta_{n_\epsilon-1}, \sigma_{n_\epsilon-1}) \in A_1\}$ . For  $\omega \in \Omega_3$ ,

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

and

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

There exists sets  $B_k, k = 1, \dots, n_1$  depending on  $n_1$  but independent of  $\mathbf{x}$  such that if  $(\beta_k, \sigma_k) \in B_k$ ,  $\left| g(\beta_k, \sigma_k, \mathbf{x}) - g(\tilde{\beta}_{k,n_1}, \tilde{\sigma}_{k,n_1}, \mathbf{x}) \right| < \epsilon$ . So for  $\omega \in \Omega_4 = \Omega_3 \cap \{(\beta_1, \sigma_1) \in B_1, \dots, (\beta_{n_1}, \sigma_{n_1}) \in B_{n_1}\}$ ,

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

Now since  $\mathcal{P}_{\mathcal{X}}(\Omega_2) > 0$  and the sets  $\{(\beta_{n_1+1}, \sigma_{n_1+1}) \in A_1, \dots, (\beta_{n_\epsilon-1}, \sigma_{n_\epsilon-1}) \in A_1\}$  and  $\{(\beta_1, \sigma_1) \in B_1, \dots, (\beta_{n_1}, \sigma_{n_1}) \in B_{n_1}\}$  are independent from  $\Omega_2$  and have positive probability, it follows that  $\mathcal{P}_{\mathcal{X}}(\Omega_4) > 0$ .  $\square$

#### Appendix D: Proof of Theorem 6.6

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 Tokdar

(2006). The  $\tilde{f}$  in (6.2) will be constructed so as to satisfy the assumptions of Lemma 6.1 and such that  $\int_{\mathcal{X}} \int_{\mathcal{Y}} f_0(y | \mathbf{x}) \log \frac{f_0(y|\mathbf{x})}{\tilde{f}(y|\mathbf{x})} dy q(\mathbf{x}) d\mathbf{x} < \frac{\epsilon}{2}$  for any  $\epsilon > 0$ . Define a sequence of conditional densities  $f_n(y | \mathbf{x}) = \int \frac{1}{\sigma} \phi(\frac{y-\mathbf{x}'\boldsymbol{\beta}}{\sigma}) d\tilde{G}_{n,\mathbf{x}}(\boldsymbol{\beta}, \sigma)$ ,  $n \geq 1$  where for  $\sigma_n = n^{-\eta}$ ,

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

Define

$$f_n(y | \mathbf{x}) = \frac{\int_{-nx_1}^{nx_1} \frac{1}{\sigma_n} \phi(\frac{y-t}{\sigma_n}) f_0(t | \mathbf{x}) dt}{\int_{-nx_1}^{nx_1} f_0(t | \mathbf{x}) dt}. \quad (\text{D.2})$$

Proceeding as in Theorem 3.2 of Tokdar (2006), an application of DCT using the conditions A1-A5 yields

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} f_0(y | \mathbf{x}) \log \frac{f_0(y | \mathbf{x})}{f_n(y | \mathbf{x})} dy q(\mathbf{x}) d\mathbf{x} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Therefore one can simply choose  $\tilde{f} = f_{n_0}$  for sufficiently large  $n_0$ .  $f_{n_0}$  satisfies the assumptions of Lemma 6.1 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  $\mathbf{x} \rightarrow G_{n_0,\mathbf{x}}(A)$  is continuous. Hence there exists a finite intersection  $W$  of neighborhoods of  $\{G_{n_0,\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$  the type (6.5) such that for any  $\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} \in W$ , the second term of (6.2) is arbitrarily small. The conclusion of the theorem follows immediately from Corollary 6.4.  $\square$

### Appendix E: Another useful lemma

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

*Proof.* Denote by  $f$  the corresponding density functions.

$$\begin{aligned} P(A_1 + A_2 \leq t_0) &= \int_0^{t_0} f_{A_1}(u) P(A_2 \leq t_0 - u) \leq C_2 \int_0^{t_0} f_{A_1}(u) P(B_2 \leq t_0 - u) \\ &= C_2 P(A_1 + B_2 \leq t_0) = C_2 \int_0^{t_0} f_{B_2}(u) P(A_1 \leq t_0 - u) \\ &\leq C_1 C_2 \int_0^{t_0} f_{B_2}(u) P(B_1 \leq t_0 - u) = C_1 C_2 P(B_1 + B_2 \leq t_0). \end{aligned}$$

$\square$

### Appendix F: Proof of Theorem 7.3

$\tilde{f}$  in (6.2) will be constructed so as to satisfy the assumptions of Lemma 7.1 and such that  $\int_{\mathcal{X}} \int_{\mathcal{Y}} f_0(y | \mathbf{x}) \log \frac{f_0(y|\mathbf{x})}{\tilde{f}(y|\mathbf{x})} dy q(\mathbf{x}) d\mathbf{x} < \frac{\epsilon}{2}$  for any  $\epsilon > 0$ . Define

a sequence of conditional densities  $f_n(y | \mathbf{x}) = \int \frac{1}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right) d\tilde{G}_{n,\mathbf{x}}(\mu, \sigma)$ ,  $n \geq 1$  where for  $\sigma_n = n^{-\eta}$ ,

$$dG_{n,\mathbf{x}}(\mu, \sigma) = \frac{\int_{-n}^n f_0(\mu | \mathbf{x}) \delta_{\sigma_n}(\sigma)}{\int_{-n}^n f_0(\mu | \mathbf{x})}. \quad (\text{F.1})$$

As before define the approximator

$$f_n(y | \mathbf{x}) = \frac{\int_{-n}^n \frac{1}{\sigma_n} \phi\left(\frac{y-t}{\sigma_n}\right) f_0(t | \mathbf{x}) dt}{\int_{-n}^n f_0(t | \mathbf{x}) dt}. \quad (\text{F.2})$$

$\tilde{f}$  will be chosen to be  $f_{n_0}$  for some large  $n_0$ .  $f_{n_0}$  satisfies the assumptions of Lemma 7.1 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  $\mathbf{x} \rightarrow \int_{\mathbb{R} \times \mathbb{R}^+} g(\mu, \sigma) dG_{n_0,\mathbf{x}}(\mu, \sigma)$  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 (7.1) such that for any  $\{G_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\} \in W$ , the second term of (6.2) is arbitrarily small. The conclusion of the theorem follows immediately from a variant of Corollary 6.4 applied to neighborhoods of the type (7.1).  $\square$

#### Appendix G: Proof of Lemma 7.4

We have  $\Theta_n \subset \{\phi_{\mu,\sigma} | \beta \in (-a_n\sqrt{p}, a_n\sqrt{p}]^p, \eta \in B_n, h_n \leq \sigma \leq l_n\}$ . Let  $\kappa < \min(\frac{\epsilon}{6}, 1)$  and  $\sigma_m = l_n(1 + \kappa)^m$ ,  $m \geq 0$ . Let  $m_0$  be the smallest integer such that  $\sigma_{m_0} = l_n(1 + \kappa)^{m_0} > h$ . This implies  $m_0 \leq (1 + \kappa)^{-1} \log \frac{h_n}{l_n} + 1$ . By the choice of  $\sigma_m$ ,  $m \geq 1$ ,  $\frac{3(\sigma_m - \sigma_{m-1})}{\sigma_{m-1}} < \frac{\epsilon}{2}$ . Let  $N_j = \lceil (\frac{128}{\pi})^{1/2} \frac{a_n p \sqrt{p}}{\sigma_{j-1} \epsilon} \rceil$ . For each  $1 \leq j \leq m_0$ , construct a  $\frac{\epsilon \sqrt{\pi} \sigma_{j-1}}{4\sqrt{2}}$ -covering  $\{A_{kj}, k = 1, \dots, M_j\}$  of  $B_n$  with

$$M_j = N \left( \frac{\epsilon \sqrt{\pi} \sigma_{j-1}}{4\sqrt{2}}, B_n, \|\cdot\|_{\infty} \right) \leq \exp \left[ K_2 r_n^p \left\{ \log \left( \frac{8\sqrt{2} M_n \sqrt{r_n / \delta_n}}{\epsilon \sqrt{\pi} \sigma_{j-1}} \right) \right\}^{p+1} \log \frac{K_3 M_n}{\epsilon l_n} \right]$$

for some constants  $K_2, K_3 > 0$ . For  $1 \leq i \leq N_j$ ,  $1 \leq k \leq M_j$  &  $1 \leq j \leq m_0$ , define

$$E_{ikj} = \left( -a'_n + \frac{2a'_n(i-1)}{N_j}, -a'_n + \frac{2a'_n i}{N_j} \right]^p \times A_{kj} \times (\sigma_{j-1}, \sigma_j] \quad (\text{G.1})$$

where  $a'_n = a_n \sqrt{p}$ . We have for  $(\beta, \eta, \sigma), (\beta', \eta', \sigma') \in E_{ikj}$  and for each  $\mathbf{x} \in \mathcal{X}$ ,

$$\begin{aligned} \|\phi_{\mu,\sigma}(\mathbf{x}, \cdot) - \phi_{\mu',\sigma'}(\mathbf{x}, \cdot)\|_1 &\leq \left( \frac{2}{\pi} \right)^{1/2} \frac{\|\beta - \beta'\| \sqrt{p} + \|\eta_1 - \eta_2\|_{\infty}}{\sigma_2} + \frac{\epsilon}{2} \\ &\leq \left( \frac{2}{\pi} \right)^{1/2} \left( \frac{2a_n p \sqrt{p}}{\sigma_{j-1} N_j} + \frac{\epsilon \sqrt{\pi}}{4\sqrt{2}} \right) + \frac{\epsilon}{2} \leq \epsilon. \end{aligned}$$

Thus

$$\begin{aligned}
 N(\Theta_n, \epsilon, d_{SS}) &\leq \sum_{j=1}^{m_0} \left\{ \left( \frac{128}{\pi} \right)^{1/2} \frac{a_n p \sqrt{p}}{\sigma_{j-1} \epsilon} + 1 \right\}^p \times \\
 &\quad \exp \left\{ K_2 r_n^p \log \left( \frac{8\sqrt{2} M_n \sqrt{r_n / \delta_n}}{\epsilon \sqrt{\pi} \sigma_{j-1}} \right)^{p+1} \log \frac{K_3 M_n}{\epsilon l_n} \right\} \\
 &\leq \exp \left[ K_2 r_n^p \left\{ \log \left( \frac{8\sqrt{2} M_n \sqrt{r_n / \delta_n}}{\epsilon \sqrt{\pi} l_n} \right)^{p+1} \right\} \log \frac{K_3 M_n}{\epsilon l_n} \right] \times \\
 &\quad \left\{ d_1 \left( \frac{a_n}{l_n} \right)^p + d_2 \log \frac{h_n}{l_n} + 1 \right\}.
 \end{aligned}$$

□

### Appendix H: Proof of Theorem 7.5

The proof follows similar to that of Theorem 6.11 by verifying conditions 2, 3 and 4 of Theorem 6.8. Consider the sequence of sieves defined by

$$\mathcal{F}_n = \left\{ f : f(y | \mathbf{x}) = \sum_{h=1}^{\infty} \pi_h \frac{1}{\sigma_h} \phi \left( \frac{y - \mu_h(\mathbf{x})}{\sigma_h} \right), \{\phi_{\mu_h, \sigma_h}\}_{h=1}^{m_n} \in \Theta_n, \right. \\
 \left. \sup_{\mathbf{x} \in \mathcal{X}} \sum_{h \geq m_n+1} \pi_h \leq \epsilon \right\}.$$

We will show that given any  $\xi > 0$ , there exists a  $c_1, c_2 > 0$  such that  $\Pi_{\mathcal{X}}(\mathcal{F}_n^c) \leq c_1 e^{-nc_2}$  and  $\log(\delta, \mathcal{F}_n, d_{SS}) < n\xi$ . For  $f_1, f_2 \in \mathcal{F}_n$ , we have

$$\begin{aligned}
 \|f_1(\cdot | \mathbf{x}) - f_2(\cdot | \mathbf{x})\|_1 &\leq \int_{\mathcal{Y}} \sum_{h=1}^{\infty} \left| \pi_h^{(1)} \phi_{\mu_h^{(1)}, \sigma_h^{(1)}}(\mathbf{x}, y) - \pi_h^{(2)} \phi_{\mu_h^{(2)}, \sigma_h^{(2)}}(\mathbf{x}, y) \right| dy \\
 &\leq \sum_{h=1}^{m_n} \pi_h^{(1)} \int_{\mathcal{Y}} \left| \phi_{\mu_h^{(1)}, \sigma_h^{(1)}}(\mathbf{x}, y) - \phi_{\mu_h^{(2)}, \sigma_h^{(2)}}(\mathbf{x}, y) \right| dy \\
 &\quad + \sum_{h=1}^{m_n} \left| \pi_h^{(1)} - \pi_h^{(2)} \right| + 2\epsilon.
 \end{aligned}$$

Let  $\Theta_{\pi, n} = \{\pi^{m_n} = (\pi_1, \pi_2, \dots, \pi_{m_n}) : \nu_h, h = 1, \dots, m_n \in [0, 1]\}$ . Fix  $\pi_1^{m_n}, \pi_2^{m_n} \in \Theta_{\pi, n}$ . It is easy to see that if we can make  $|\nu_{h,1} - \nu_{h,2}| < \frac{\epsilon}{m_n^2}, h = 1, \dots, m_n$ , we would have  $\sum_{h=1}^{m_n} \left| \pi_h^{(1)} - \pi_h^{(2)} \right| < \epsilon$ . Since  $\nu_{h,1}, \nu_{h,2} \in [0, 1]$ , the number of balls required to cover  $\Theta_{\pi, n}$  so that  $\sum_{h=1}^{m_n} \left| \pi_h^{(1)} - \pi_h^{(2)} \right| < \epsilon$  is  $K_4(m_n^2/\epsilon)^{m_n}$  for some

constant  $K_4 > 0$ . Hence

$$\begin{aligned} \log N(\mathcal{F}_n, 4\epsilon, d_{SS}) &\leq K_2 m_n r_n^p \left\{ \log \left( \frac{8\sqrt{2}M_n \sqrt{r_n/\delta_n}}{\epsilon \sqrt{\pi} l_n} \right) \right\}^{p+1} + m_n \log \frac{K_4 m_n^2}{\epsilon} \\ &\quad + m_n \log \frac{K_3 M_n}{\epsilon l_n} + m_n \log \left\{ d_1 \left( \frac{a_n}{l_n} \right)^p + d_2 \log \frac{h_n}{l_n} + 1 \right\} \end{aligned} \quad (\text{H.1})$$

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  $P(\Theta_n^c) \leq \{P(\|\beta\| > a_n) + P(\sigma \in [l_n, h_n]^c) + P(\eta \in B_n^c)\}$ . It follows from the proof of Theorem 3.1 of [van der Vaart and van Zanten \(2009\)](#) 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 [van der Vaart and van Zanten \(2009\)](#) indicates that  $P(A > r_n) \lesssim \exp\{-r_n^{p(1+\eta_2)/\eta_2}\}$ . Hence with  $M_n = O(n^{1/2})$ ,  $m_n = O\{n/(\log n)^{p+1}\}^{1/(1+\eta_2)}$  and  $r_n^p = O\{n^{\eta_2/(1+\eta_2)}\}$ ,  $P(\Theta_n^c) \lesssim e^{-n}$  and

$$P\left(\sum_{h=m_n}^{\infty} \pi_h > \epsilon\right) \lesssim \exp\{-m_n^{1+\eta_2}(\log m_n)^{(p+1)}\} \lesssim e^{-n}. \quad (\text{H.2})$$

Also, the first term in the right hand side of (H.1) 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 (H.1) can be made to grow at  $o(n)$ . Henceforth verification of conditions 2, 3 and 4 of Theorem 6.8 follows similar to the proof of Theorem 6.11.  $\square$

## References

- ADLER, R. (1990). *An introduction to continuity, extrema, and related topics for general Gaussian processes* **12**. Institute of Mathematical Statistics.
- ALBERT, J. and CHIB, S. (2001). Sequential ordinal modeling with applications to survival data. *Biometrics* **57** 829–836.
- BARRIENTOS, F., JARA, A. and QUINTANA, F. (2011). On the support of MacEacherns dependent Dirichlet processes.
- BARRON, A., SCHERVISH, M. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics* **27** 536–561.
- BHATTACHARYA, A. and DUNSON, D. (2010). Strong consistency of nonparametric Bayes density estimation on complex Watson kernels. *Duke University, DSS discussion series*.
- CHUNG, Y. and DUNSON, D. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association* **104** 1646–1660.

- DE IORIO, M., MUELLER, P., ROSNER, G. and MACEACHERN, S. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association* **99** 205–215.
- DE IORIO, M., JOHNSON, W., MÜLLER, P. and ROSNER, G. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics* **65** 762–771.
- DUNSON, D. and PARK, J. (2008). Kernel stick-breaking processes. *Biometrika* **95** 307–323.
- DUNSON, D., PILLAI, N. and PARK, J. (2007). Bayesian density regression. *Journal of the Royal Statistical Society, Series B* **69** 163–183.
- FERGUSON, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1** 209–230.
- FERGUSON, T. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics* **2** 615–629.
- GELFAND, A., KOTTAS, A. and MACEACHERN, S. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100** 1021–1035.
- GHOSAL, S., GHOSH, J. and RAMAMOORTHI, R. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics* **27** 143–158.
- GHOSAL, S. and VAN DER VAART, A. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics* **29** 1233–1263.
- GHOSAL, S. and VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics* **35** 697–723.
- GRIFFIN, J. and STEEL, M. (2006). Order-based dependent Dirichlet processes. *Journal of The American Statistical Association* **101** 179–194.
- GRIFFIN, J. and STEEL, M. (2008). Bayesian nonparametric modelling with the dirichlet process regression smoother. *Statistica Sinica*. (to appear).
- JARA, A., LESAFFRE, E., DE IORIO, M. and QUINTANA, F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *The Annals of Applied Statistics* **4** 2126–2149.
- MACEACHERN, S. (1999). Dependent nonparametric processes. In *Proceedings of the Section on Bayesian Statistical Science* 50–55.
- MÜLLER, P., ERKANLI, A. and WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83** 67–79.
- NORETS, A. and PELENIS, J. (2009). Bayesian modeling of joint and conditional distributions. *Unpublished manuscript, Princeton Univ.*
- NORETS, A. and PELENIS, J. (2010). Posterior consistency in conditional distribution estimation by covariate dependent mixtures. *Unpublished manuscript, Princeton Univ.*
- PAPASPILIOPOULOS, O. and ROBERTS, G. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95** 169.
- PARK, B. and DUNSON, D. (2009). Bayesian generalized product partition model. *Statistica Sinica*. (to appear).

- PATI, D. and DUNSON, D. (2009). Bayesian nonparametric regression with varying residual density. *Unpublished paper*.
- RODRIGUEZ, A. and DUNSON, D. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis* **6** 145–178.
- SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26.
- SHAHBABA, B. and NEAL, R. (2009). Nonlinear models using dirichlet process mixtures. *Journal of Machine Learning Research* **10** 1829–1850.
- TANG, Y. and GHOSAL, S. (2007a). A consistent nonparametric Bayesian procedure for estimating autoregressive conditional densities. *Computational Statistics & Data Analysis* **51** 4424–4437.
- TANG, Y. and GHOSAL, S. (2007b). Posterior consistency of Dirichlet mixtures for estimating a transition density. *Journal of Statistical Planning and Inference* **137** 1711–1726.
- TOKDAR, S. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics* **67** 90–110.
- TOKDAR, S. and GHOSH, J. (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference* **137** 34–42.
- TOKDAR, S., ZHU, Y. and GHOSH, J. (2010). Bayesian Density Regression with Logistic Gaussian Process and Subspace Projection. *Bayesian Analysis* **5** 1–26.
- VAN DER VAART, A. and VAN ZANTEN, J. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections* **3** 200–222.
- VAN DER VAART, A. and VAN ZANTEN, J. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics* **37** 2655–2675.
- WALKER, S. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation* **36** 45–54.
- WU, Y. and GHOSAL, S. (2010). L1-Consistency of Dirichlet Mixtures in Multivariate Bayesian Density Estimation. *Journal of Multivariate Analysis*. (to appear).
- YOON, J. (2009). Bayesian analysis of conditional density functions: a limited information approach. *Unpublished manuscript, Claremont Mckenna College*.