Econometric Theory, **30**, 2014, 606–646.
doi:10.1017/S026646661300042X

POSTERIOR CONSISTENCY IN CONDITIONAL DENSITY ESTIMATION BY COVARIATE DEPENDENT MIXTURES

ANDRIY NORETS
University of Illinois at Urbana-Champaign

JUSTINAS PELENIS
Institute for Advanced Studies, Vienna

This paper considers Bayesian nonparametric estimation of conditional densities by countable mixtures of location-scale densities with covariate dependent mixing probabilities. The mixing probabilities are modeled in two ways. First, we consider finite covariate dependent mixture models, in which the mixing probabilities are proportional to a product of a constant and a kernel and a prior on the number of mixture components is specified. Second, we consider kernel stick-breaking processes for modeling the mixing probabilities. We show that the posterior in these two models is weakly and strongly consistent for a large class of data-generating processes. A simulation study conducted in the paper demonstrates that the models can perform well in small samples.

1. INTRODUCTION

The estimation of conditional distributions is an important problem in empirical economics. It is often desirable to estimate not only the effect of covariates on the average of outcomes but also how the whole distribution of outcomes depends on covariates. Relevant classical semi- and nonparametric methods, such as quantile regression, kernel smoothing, and sieves, are widely used in econometrics. Yatchew (1998), Koenker and Hallock (2001), DiNardo and Tobias (2001), Ichimura and Todd (2007), and Chen (2007) provide surveys of methodological and applied work. Typical applications include estimation of how distributions of wages, prices, and costs depend on covariates. In time series settings, nonparametric estimation of conditional densities is useful for forecasting; see the literature survey in Fan (2005).

The use of Bayesian nonparametric models is less common, especially in methodological econometric research. However, Bayesian nonparametric methods have a number of attractive properties. First, they never result in logical

We are grateful to Ulrich Müller for helpful discussions. We thank the co-editor and anonymous referees for useful suggestions. All remaining errors are ours. Address corresponding to Andriy Norets, Department of Economics, University of Illinois, 1407 W. Gregory Drive, Urbana, IL 61801; e-mail: anorets@illinois.edu.

inconsistencies such as crossing quantiles in quantile regression or negative density estimates in higher-order kernel smoothing. Second, in the Bayesian framework, it is straightforward to incorporate the uncertainty about parameter/density estimates in forecasting and, more generally, decision-making problems. Third, prior information can be explicitly added to an estimation procedure. Finally, Bayesian nonparametric methods have been proved to possess excellent frequentist properties in several important problems. For example, Rousseau (2010) and Kruijer, Rousseau, and van der Vaart (2010) show that in univariate density estimation, Bayesian models based on mixtures of distributions automatically adapt to the smoothness of the target density and deliver minimax convergence rates up to a logarithmic factor. They also demonstrate that there is no need to select sample-size-dependent tuning parameters such as bandwidth to achieve the optimal convergence rates. See also van der Vaart and van Zanten (2009) for similar results for priors based on Gaussian processes.

The econometrics literature on Bayesian nonparametric conditional density estimation includes papers by Geweke and Keane (2007), Villani, Kohn, and Giordani (2009), Li, Villani, and Kohn (2010), Tran, Nott, and Kohn (2012), and Villani, Kohn, and Nott (2012), among others. Typical applications in these papers are estimation of the distribution of earnings and firms' leverage ratios and forecasting stock returns and inflation. These authors develop estimation methods, conduct Monte Carlo experiments, and provide assessment of out-of-sample performance for several different model specifications. Many other specifications have also been suggested in statistics in references provided below. However, there is little theoretical guidance on what specifications are preferable or at least guaranteed to work well in large samples. A widely accepted minimal requirement for large-sample behavior of Bayesian nonparametric models is posterior consistency (see Ghosh and Ramamoorthi, 2003, for a textbook treatment). Posterior consistency means that in a frequentist thought experiment with a fixed (possibly infinite-dimensional) parameter of a data-generating process (DGP), the posterior concentrates around this fixed parameter as the sample size increases. The benefits of posterior consistency from the Bayesian perspective are at least twofold. First, it means that the prior is not dogmatic and can be overwhelmed by the data. Second, it ensures that Bayesians with different priors agree when the sample is sufficiently large. In this paper, we demonstrate posterior consistency for several nonparametric models for conditional densities and, thus, provide a validation for their use in applications.

There are two alternative approaches to modeling conditional densities in the Bayesian framework. First, the conditional distributions of interest can be obtained as a byproduct of the joint distribution estimation. Second, the conditional distribution can be modeled directly and the marginal distribution of the covariates can be left unspecified. Bayesian nonparametric modeling of densities involves specifying a flexible prior on the space of densities. The theory of posterior consistency for (unconditional) density estimation is well developed. However, if only

conditional density is of interest, modeling the marginal distribution of covariates is unnecessary. Also, it is not clear how to select covariates, which is useful in applications, when the joint distribution is estimated. While there are many proposed methods for direct conditional density estimation, their consistency properties are largely unknown. We address this gap in the literature by demonstrating consistency for Bayesian nonparametric procedures based on countable mixtures of location-scale densities with covariate dependent mixing probabilities. The mixing probabilities are modeled in two ways. First, we consider finite covariate dependent mixture models, in which the mixing probabilities are proportional to a product of a constant and a kernel and a prior on the number of mixture components is specified. Second, we consider the kernel stick-breaking processes of Dunson and Park (2008) for modeling the mixing probabilities. We show that the posterior in these two models is weakly and strongly consistent for a large class of DGPs. Below, we provide a more detailed overview of the literature and our contribution.

There are several important classes of priors that are used in the Bayesian nonparametric literature. One approach to nonparametric density estimation is based on Gaussian process priors; see, for example, Tokdar and Ghosh (2007), Tokdar (2007), van der Vaart and van Zanten (2008), Liang, Carlin, and Gelfand (2009), and Tokdar, Zhu and Ghosh (2010). These priors are not considered in our paper. Priors based on mixtures of distributions play an important role in the applied and theoretical literature on Bayesian nonparametric density estimation. A commonly used prior for the mixing distribution is the Dirichlet process prior introduced by Ferguson (1973). Markov Chain Monte Carlo (MCMC) estimation methods for these models were developed by Escobar (1994) and Escobar and West (1995), who used a Polya urn representation of the Dirichlet process from Blackwell and MacQueen (1973) (see Dey, Muller, and Sinha, 1998, for a more extensive list of references and applications). An alternative approach to modeling mixing distributions is to consider finite mixture models and define a prior on the number of mixture components (references on finite mixture models can be found in a comprehensive book by McLachlan and Peel, 2000).

A general weak posterior consistency theorem for density estimation was established by Schwartz (1965). Barron (1988), Barron, Schervish, and Wasserman (1999), and Ghosal et al. (1999) developed a theory of strong posterior consistency. The latter authors demonstrated that the theory applies to Dirichlet process mixtures of normals, which is often used in practice. Tokdar (2006) relaxed some of their sufficient conditions in the context of the Dirichlet process mixture of normals. An alternative approach to consistency was introduced by Walker (2004). Ghosal and Tang (2006) used this approach to obtain posterior consistency for Markov processes. Zeevi and Meir (1997), Genovese and Wasserman (2000), Roeder and Wasserman (1997), and Li and Barron (1999) also obtained approximation and classical and Bayesian consistency results for mixture models. Posterior convergence rates for mixture models were obtained by Ghosal, Ghosh,

and van der Vaart (2000) and Kruijer et al. (2010), among others. Wu and Ghosal (2010) and Norets and Pelenis (2012) considered consistency in estimation of multivariate densities.

Muller, Erkanli, and West (1996), Roeder and Wasserman (1997), Norets and Pelenis (2012), and Taddy and Kottas (2010) suggested obtaining conditional densities of interest from joint distribution estimation. MacEachern (1999), De Iorio, Muller, Rosner, and MacEachern (2004), Griffin and Steel (2006), Dunson and Park (2008), and Chung and Dunson (2009), among others, developed dependent Dirichlet processes in which conditional distribution is modeled as a mixture with covariate dependent mixing distribution and possibly covariate dependent means and variances of the mixed distributions. There are alternative approaches to modeling conditional distributions directly that are based on finite covariate dependent mixtures known in the literature as mixtures of experts and smoothly mixing regressions (Jacobs, Jordan, Nowlan, and Hinton, 1991; Jordan and Xu, 1995; Peng, Jacobs, and Tanner, 1996; Wood, Jiang, and Tanner, 2002; Geweke and Keane, 2007; Villani et al., 2009; and Norets, 2010).

Posterior consistency results for direct conditional density estimation are scarce. Norets (2010) shows that large nonparametric classes of conditional densities can be approximated in the Kullback-Leibler distance by three different specifications of finite mixtures of normal densities: (i) Only means of the mixed normals depend flexibly on covariates; (ii) only mixing probabilities depend flexibly on covariates; and (iii) only mixing probabilities modeled by multinomial logit model depend on covariates. Schwartz's (1965) theory suggests that these Kullback-Leibler approximation results imply posterior consistency in a weak topology norm. Pati, Dunson, and Tokdar (2013) specify dependent Dirichlet processes that are similar to specifications (i) and (ii) of Norets (2010) and demonstrate weak and strong posterior consistency. They use Gaussian processes to specify flexible priors for mixing probabilities (for specification (ii)) and means of normals (for specification (i)).

Relative to Norets (2010) and Pati et al. (2013), our contribution is fivefold. First, we generalize Kullback-Leibler approximation results from Norets to finite mixture specifications in which mixing probabilities are proportional to a general kernel multiplied by a constant. We will call such a mixture specification kernel mixture (KM). Second, we prove weak and strong posterior consistency for kernel mixtures combined with a prior on the number of mixture components. Third, we show that the kernel stick-breaking processes introduced by Dunson and Park (2008) can approximate kernel mixtures. Fourth, we obtain weak and strong posterior consistency results for the kernel stick-breaking mixtures. Fifth, our weak and strong posterior consistency results hold for mixtures of general location-scale densities.

While approximation and weak posterior consistency results for kernel mixtures could be anticipated from the results of Norets (2010), the approximation and consistency results for kernel stick-breaking mixtures seem to be novel. We show that it is not necessary to use fully flexible in covariates components in the

stick-breaking process as in Pati et al. (2013), and it is sufficient to use kernels instead, which are fixed, known functions that depend on finite dimensional location and scale parameters.

The regularity conditions on the DGP that we assume in proving weak and strong posterior consistency are very mild. Assumptions about the prior for the location and scale parameters of the mixed densities employed in showing strong posterior consistency are similar under both types of mixing. Standard normal priors for locations and inverse gamma for squared scales satisfy the assumptions. Although the parameters entering the mixing probabilities under two types of mixing are the same, the priors on these parameters might have to be different in the two models if strong posterior consistency is desired. For kernel mixtures there are no restrictions on the prior for constants multiplying the kernels. For stick-breaking mixtures these constants are assumed to have a prior that puts more mass on values close to 1. The only restriction on the prior for locations of the mixing probability kernels is that its support has to cover the space for covariates.

The organization of the paper is as follows: Section 2 defines weak and strong posterior consistency for conditional densities and presents general theoretical results that are used later in the paper. Posterior consistency results for kernel mixtures are given in Section 3. Section 4 covers kernel stick-breaking mixtures. Section 5 discusses generalizations of models defined in Sections 3–4. The finite sample performance of the models is evaluated in simulation exercises in Section 6. Section 7 concludes.

2. THE NOTION OF POSTERIOR CONSISTENCY FOR CONDITIONAL DENSITIES

Consider a product space $Y \times X$, $Y \subset R$ and $X \subset R^{d_x}$. Let $\mathcal{F} = \left\{ f : Y \times X \to [0, \infty), \int_Y f(y|x)dy = 1 \right\}$ be the set of all conditional densities on Y with respect to the Lebesgue measure. Let us denote the data-generating density of covariates x with respect to some generic measure ν by $f_0^x(x)$ and the data-generating conditional density of interest by $f_0 \in \mathcal{F}$. The joint probability measure implied by f_0 and $f_0^x(x)$ is denoted by F_0.

To define a notion of posterior consistency, we need to define neighborhoods on the space of conditional densities. The previous literature on Bayesian nonparametric density estimation employed weak and strong topologies on spaces of densities with respect to some common dominating measure. Quite general weak and strong posterior consistency theorems were established (Schwartz, 1965; Barron, 1988; Barron et al., 1999; Ghosal, Ghosh, and Ramamoorthi, 1999; and Walker, 2004). It is possible to use these results if we define the distances between conditional densities as the corresponding distances between the joint densities, where the density of the covariates is equal to the data-generating density $f_0^x(x)$. For example, a distance between conditional densities $f_1, f_2 \in \mathcal{F}$ that generates strong neighborhoods is defined by the total variation distance

between the joint distributions,

$$\int \left| f_1 f_0^x - f_2 f_0^x \right| = \int \left| f_1(y|x) f_0^x(x) - f_2(y|x) f_0^x(x) \right| dy dv(x).$$

A distance that generates weak neighborhoods for conditional densities can be defined similarly (an explicit expression for the distance generating weak topology can be found in Billingsley, 1999). Equivalently, one can define a weak neighborhood of $f_0 \in \mathcal{F}$ as a set containing a set of the form

$$U = \left\{ f \in \mathcal{F} : \left| \int g_i f f_0^x - \int g_i f_0 f_0^x \right| < \epsilon, i = 1, 2, \ldots, k \right\},$$

where g_i's are bounded continuous functions on $Y \times X$.

For $\epsilon > 0$ define a Kullback-Leibler neighborhood of f_0 as

$$K_\epsilon(f_0) = \left\{ f \in \mathcal{F} : \int \log \frac{f_0(y|x)}{f(y|x)} dF_0(y,x) = \int \log \frac{f_0(y|x) f_0^x(x)}{f(y|x) f_0^x(x)} dF_0(y,x) < \epsilon \right\}.$$

Similarly defined integrated total variation and Kullback-Leibler distances for conditional densities were considered in Ghosal and Tang (2006) and Norets and Pelenis (2012).

Since we are interested only in conditional distributions, we specify a prior on $\mathcal{F}$. The corresponding posterior given data $(X_T, Y_T) = (x_1, y_1, \ldots, x_T, y_T)$ is denoted by $\Pi(\cdot|X_T, Y_T)$. In order to apply posterior consistency theorems formulated for joint densities, we can think of a prior Π on $\mathcal{F}$ as a prior on the space of joint densities on $Y \times X$ that puts probability 1 on f_0^x. The posterior of the conditional density does not involve f_0^x; f_0^x plays a role only in the proof of posterior consistency.

The following weak posterior consistency theorem is an immediate implication of Schwartz's theorem.

THEOREM 2.1. *If $\Pi(K_\epsilon(f_0)) > 0$ for any $\epsilon > 0$, then the corresponding posterior is weakly consistent at f_0: For any weak neighborhood U of f_0,*

$$\Pi(U|Y_T, X_T) \to 1, \quad \text{a.s. } F_0^\infty.$$

The proof of the theorem is exactly the same as the proof of Schwartz's theorem and its implications (see Ghosh and Ramamoorthi (2003) for a textbook treatment).

To show strong posterior consistency we use a theorem from Ghosal et al. (1999). To state the theorem we need a notion of the L_1-metric entropy. Let $A \subset \mathcal{F}$. For $\delta > 0$, the L_1-metric entropy $J(\delta, A)$ is defined as the logarithm of the minimum of all k such that there exist $f_1, \ldots, f_k$ in $\mathcal{F}$ with the property $A \subset \cup_{i=1}^k \{ f : \int |f - f_i| f_0^x < \delta \}$.

THEOREM 2.2. *Suppose $\Pi(K_\epsilon(f_0)) > 0$ for any $\epsilon > 0$. Let $U = \{ f : \int |f - f_0| f_0^x < \epsilon \}$. If for given $\epsilon > 0$ there is a $\delta < \epsilon/4$, $c_1, c_2 > 0$, $\beta < \epsilon^2/8$ and $\mathcal{F}_n \subset \mathcal{F}$ such that for all n large enough;*

(i) $\Pi(\mathcal{F}_n^c) < c_1 \exp\{-c_2 n\}$ *and*
(ii) $J(\delta, \mathcal{F}_n) < \beta n$,

then $\Pi(U|Y_T, X_T) \to 1$, *almost surely* (*a.s.*) F_0^∞.

The proof of the theorem is exactly the same as the proof of Theorem 2 in Ghosal et al. (1999). In the following sections we use these weak and strong posterior consistency theorems to demonstrate consistency for countable covariate dependent location-scale mixtures.

3. KERNEL MIXTURES WITH A VARIABLE NUMBER OF COMPONENTS

Consider the model for a conditional density,

$$p(y|x,\theta,m) = \frac{\sum_{j=1}^m \alpha_j K\big(-Q_j||x-q_j||^2\big)\phi\big(y,\mu_j,\sigma_j\big)}{\sum_{i=1}^m \alpha_i K\big(-Q_i||x-q_i||^2\big)}, \tag{3.1}$$

where $\phi(y,\mu,\sigma)$ is a fixed symmetric density with location μ and scale σ evaluated at y and $K(\cdot)$ is a fixed positive function, for example, $K(\cdot) = \exp(\cdot)$. The prior on the space of conditional densities is defined by a prior distribution for a positive integer m (the number of mixture components) and $\theta = \big\{Q_j,\mu_j,\sigma_j,q_j,\alpha_j\big\}_{j=1}^\infty \in \Theta = (R_+ \times Y \times R_+ \times X \times (0,1))^\infty$, where $Q_j \in R_+$, $\mu_j \in Y$, $\sigma_j \in R_+$, $q_j \in X$, and $\alpha_j \in (0,1)$. Also, let $\theta_{1:m} = \{Q_j,\mu_j,\sigma_j,q_j,\alpha_j\}_{j=1}^m$ and note that $p(y|x,\theta,m) = p(y|x,\theta_{1:m},m)$. In a slight abuse of notation, $\Pi(\cdot)$ and $\Pi(\cdot|X_T,Y_T)$ will denote the prior and the posterior on the space of conditional densities and, equivalently, on $\Theta \times \{1,2,\ldots,\infty\}$.

3.1. Weak Consistency

We impose the following assumption on the DGP.

Assumption 3.1.

(i) $X = [0,1]^{d_x}$ (the arguments would go through for a bounded X).
(ii) $f_0(y|x)$ is continuous in (y,x) a.s. F_0.
(iii) There exists $r > 0$ such that

$$\int \log \frac{f_0(y|x)}{\inf_{||z-y||\le r,\ ||t-x||\le r} f_0(z|t)} F_0(dy,dx) < \infty. \tag{3.2}$$

The condition in (3.2) requires logged relative changes in $f_0(y|x)$ to be finite on average. The condition also implies that $f_0(y|x)$ is positive for any $x \in X$ and $y \in R$. The condition can be modified to accommodate bounded support of y; see Norets (2010) (this generalization is not pursued here to simplify the notation). Norets shows that Laplace and Student's t-distributions with covariate dependent parameters as well as nonparametrically specified DGPs satisfy this assumption.

The assumption of the bounded support for covariates seems difficult to relax. In the following Kullback-Leibler distance approximation result (Theorem 3.1), we need an integrable upper bound on the logarithm of the ratio of the data-generating density and the model density. The boundedness of covariates plays an important role in obtaining such a bound. One way to apply our theoretical results to the data with unbounded covariates is to transform the covariates. In this case, the condition in (3.2) is admittedly stronger but still could be satisfied; for example, it holds when the true conditional density is normal with the mean equal to a uniformly bounded function of covariates. Another way to handle unbounded covariates is to estimate the conditional density on a bounded subset of the support of the covariates.

We also make the following assumption about the location-scale density ϕ.

Assumption 3.2.

(i) $\phi(y, \mu, \sigma) = \sigma^{-1}\psi((y-\mu)/\sigma)$, where $\psi(z)$ is a bounded, continuous, symmetric around zero, and monotone decreasing in $|z|$ probability density.

(ii) For any μ and $\sigma > 0$, $\log\phi(y, \mu, \sigma)$ is integrable with respect to F_0.

A standard normal density satisfies this assumption as long as the second moments of y are finite. A Laplace density also satisfies this assumption if the first moments of y are finite. The condition seems to imply that to estimate $f_0(y|x)$ by mixtures, one needs to mix densities with tails that are not too thin relative to $f_0(y|x)$.

We also make the following assumption about the kernel $K(\cdot)$.

Assumption 3.3. The kernel $K(\cdot)$ is positive, bounded above, continuous, nondecreasing, and has a bounded derivative on $(-\infty, 0]$. The upper bound can be set to 1 and, thus, $1 \geq K(z) > 0$ for $z \in (-\infty, 0]$. Also, we assume $n^{d_x/2}K(-2n)/K(-n) \to 0$ as $n \to \infty$.

An exponential kernel $K(z) = \exp(z)$ satisfies the assumption. The following theorem is a generalization of Proposition 4.1 in Norets (2010).

THEOREM 3.1. *If Assumptions 3.1–3.3 hold, then for any $\epsilon > 0$ there exist m and $\theta_{1:m} = \{Q_j, \mu_j, \sigma_j, q_j, \alpha_j\}_{j=1}^m$ such that*

$$\int \log \frac{f_0(y|x)}{p(y|x, \theta_{1:m}, m)} dF_0(y, x) < \epsilon.$$

The theorem is proved in Appendix A. The intuition behind the proof is as follows. For a fixed x, the conditional density can be approximated by a finite location-scale mixture. The mixing probabilities in the approximation depend continuously on x. These continuous mixing probabilities can be approximated by step functions (sums of products of indicator functions and constants). The indicator functions in turn can be approximated by $K(\cdot)$, which gives rise to an expression in (3.1) after a normalization. The following corollary shows that the approximation stays good in a sufficiently small neighborhood of $\theta_{1:m}$.

COROLLARY 3.1. *Suppose Assumptions 3.1–3.3 hold. Then, for a given $\epsilon > 0$ there are m and an open neighborhood Θ^m such that for any $\theta_{1:m} \in \Theta^m$,*

$$\int \log \frac{f_0(y|x)}{p(y|x,\theta_{1:m},m)} dF_0(y,x) < \epsilon.$$

Proof. By Theorem 3.1, there exist m and $\tilde{\theta}_{1:m}$ such that

$$\int \log \frac{f_0(y|x)}{p(y|x,\tilde{\theta}_{1:m},m)} dF_0(y,x) < \epsilon/2.$$

For any $\theta_{1:m}$,

$$\int \log \frac{f_0(y|x)}{p(y|x,\theta_{1:m},m)} dF_0(y,x) = \int \log \frac{f_0(y|x)}{p(y|x,\tilde{\theta}_{1:m},m)} dF_0(y,x) + \int \log \frac{p(y|x,\tilde{\theta}_{1:m},m)}{p(y|x,\theta_{1:m},m)} dF_0(y,x).$$

The first part of the right-hand side (r.h.s) of this equality is bounded above by $\epsilon/2$. It suffices to show that the second part is continuous in $\theta_{1:m}$ at $\tilde{\theta}_{1:m}$. Let $\theta^n_{1:m}$ be a sequence of parameter values converging to some $\tilde{\theta}_{1:m}$ as $n \to \infty$. For every y, $p(y|x,\tilde{\theta}_{1:m},m)/p(y|x,\theta^n_{1:m},m) \to 1$. The result will follow from the dominated convergence theorem if there is an integrable (with respect to F_0) upper bound on $|\log p(y|x,\theta^n_{1:m},m)|$. Since $\theta^n_{1:m} \to \tilde{\theta}_{1:m}$, $\mu^n_j \in (\underline{\mu},\overline{\mu})$ and $\sigma^n_j \in (\underline{\sigma},\overline{\sigma})$ for some finite $\underline{\mu}$, $\overline{\mu}$, $\underline{\sigma} > 0$, and $\overline{\sigma}$ for all sufficiently large n. From Assumption 3.2,

$$\frac{\psi(0)}{\underline{\sigma}} \geq p\left(y|x,\theta^n_{1:m}\right) \geq \frac{1_{(-\infty,\underline{\mu})}(y)\psi(\frac{y-\overline{\mu}}{\underline{\sigma}}) + 1_{(\overline{\mu},\infty)}(y)\psi(\frac{y-\underline{\mu}}{\underline{\sigma}}) + 1_{[\underline{\mu},\overline{\mu}]}(y)\psi(\frac{\overline{\mu}-\underline{\mu}}{\underline{\sigma}})}{\overline{\sigma}}. \tag{3.3}$$

The upper bound in (3.3) is a constant, and the logarithm of the lower bound is integrable by part (ii) of Assumption 3.2. ■

The corollary combined with a prior that puts positive mass on open neighborhoods essentially states that the Kullback-Leibler (KL) property holds: The prior probabilities of the Kullback-Leibler neighborhoods of the data-generating density $f_0(y|x)f^x_0(x)$ have positive prior probability, where the prior on the density of x puts probability one on f^x_0 and the prior for conditional densities is defined by Π introduced above. By Theorem 2.1, the KL property immediately implies the following weak posterior consistency theorem.

THEOREM 3.2. *Suppose*

(i) Assumptions 3.1–3.3 hold.

(ii) For any m, $\theta_{1:m}$ and an open neighborhood of $\theta_{1:m}$, Θ^m, $\Pi(\tilde{\theta}_{1:m} \in \Theta^m, m) > 0$.

Then for any weak neighborhood U of $f_0(y|x)$,

$$\Pi(U|Y_T, X_T) \to 1, \text{ a.s. } F_0^\infty.$$

3.2. Strong Consistency

A natural way to define a sieve $\mathcal{F}_n$ on $\mathcal{F}$ for application of Theorem 2.2, for which bounds on prior probabilities $\Pi(\mathcal{F}_n^c)$ can be easily calculated, is to consider densities $p(y|x,\theta,m)$ where m and θ are restricted in some way. To obtain a finite value for the L_1-metric entropy, one at least has to restrict components of θ to a bounded set. Thus, let us define

$$\mathcal{F}_n = \{p(y|x,\theta,m) : |\mu_j| \le \overline{\mu}_n,\ Q_j \le \overline{Q}_n,\ \underline{\sigma}_n < \sigma_j < \overline{\sigma}_n,\ 1 \le j \le m,\ m \le m_n\}.$$

We calculate a bound on $J(\delta, \mathcal{F}_n)$ in the following proposition.

PROPOSITION 3.1. *Suppose Assumptions 3.2 and 3.3 hold. Then*

$$J(\delta, \mathcal{F}_n) \le m_n \left(\log \left[b_0 \frac{\overline{\mu}_n}{\underline{\sigma}_n} + b_1 \log \frac{\overline{\sigma}_n}{\underline{\sigma}_n} + 1 \right] + b_2 + b_3 \log \overline{Q}_n + b_4 \log K\left(-\overline{Q}_n d_x\right) \right), \tag{3.4}$$

where b_0, b_1, b_2, b_3, and b_4 depend on δ but not on m_n, $\overline{Q}_n$, $\overline{\mu}_n$, $\overline{\sigma}_n$, and $\underline{\sigma}_n$.

A proof is provided in Appendix A. In addition to addressing the case of covariate dependent mixing probabilities, the proposition shows that the entropy bounds derived in Ghosal et al. (1999) and Tokdar (2006) for mixtures of normal densities hold for mixtures of general location-scale densities. The next theorem formulates sufficient conditions for strong posterior consistency.

THEOREM 3.3. *Suppose*

(i) A priori (μ_j, σ_j, Q_j) are independent and identically distributed (i.i.d.) across j and independent from other parameters of the model.

(ii) For any $\epsilon > 0$, there exist $\delta < \epsilon/4$, $\beta < \epsilon^2/8$, positive constants c_1 and c_2, and sequences m_n, $\overline{Q}_n$, $\overline{\mu}_n$, $\overline{\sigma}_n \uparrow \infty$, and $\underline{\sigma}_n \downarrow 0$ with $\overline{\sigma}_n > \underline{\sigma}_n$ such that

$$m_n \left[\Pi(|\mu_j| > \overline{\mu}_n) + \Pi\left(\underline{\sigma}_n > \sigma_j\right) + \Pi\left(\sigma_j > \overline{\sigma}_n\right) + \Pi(Q_j > \overline{Q}_n) \right] + \Pi(m > m_n) \le c_1 e^{-c_2 n}, \tag{3.5}$$

$$m_n \left(\log \left[b_0 \frac{\overline{\mu}_n}{\underline{\sigma}_n} + b_1 \log \frac{\overline{\sigma}_n}{\underline{\sigma}_n} + 1 \right] + b_2 + b_3 \log \overline{Q}_n + b_4 \log K\left(-\overline{Q}_n d_x\right) \right) < n\beta, \tag{3.6}$$

where constants $(b_0, \ldots, b_4)$ are defined in Proposition 3.1.

(iii) The conditions of Theorem 3.2 hold.

Then the posterior is strongly consistent at f_0.

Theorem 3.3 is a direct consequence of Theorem 2.2. Possible choices of prior distributions and sieve parameters that satisfy the conditions of the theorem are presented in the following example.

Example 3.1
Consider $K(z) = \exp(z)$. Let $\overline{\mu}_n = \sqrt{n}$, $\underline{\sigma}_n = 1/\sqrt{n}$, $\overline{\sigma}_n = e^n$, and $\overline{Q}_n = \sqrt{n}$. Then condition (3.6) is satisfied for $m_n = c\sqrt{n}$, where $c > 0$ is a sufficiently small constant. Next let us choose prior distributions for (μ_j, σ_j, Q_j) so that condition (3.5) holds. For a normal prior on μ_j, $\Pi(|\mu_j| > \overline{\mu}_n) < c_1 e^{-c_2 n}$ for some c_1 and c_2. For an inverse gamma prior on σ_j we will show that $\Pi(\underline{\sigma}_n > \sigma_j) + \Pi(\sigma_j > \overline{\sigma}_n) < c_1 e^{-c_2 n}$ for n large enough and some c_1 and c_2. For n large enough,

$$\begin{aligned}\Pi(\underline{\sigma}_n^2 > \sigma_j^2) + \Pi(\sigma_j^2 > \overline{\sigma}_n^2) &= const \cdot \left(\int_0^{1/n} x^{-\alpha-1} e^{-\beta/x} dx + \int_{e^{2n}}^{\infty} x^{-\alpha-1} e^{-\beta/x} dx \right) \\ &\leq const \cdot \left(\int_0^{1/n} (1/n)^{-\alpha-1} e^{-\beta/(1/n)} dx + \int_{e^{2n}}^{\infty} x^{-\alpha-1} dx \right) \\ &= const \cdot \left(n^{\alpha} e^{-\beta n} + e^{-2\alpha n}/\alpha \right) < c_1 e^{-n c_2},\end{aligned}$$

as desired. Let $m = \lfloor \tilde{m} \rfloor$ and choose a Weibull prior with shape parameter $k \geq 2$ for $\tilde{m}$ and Q_j; then (3.5) is satisfied. Alternative choices of prior distributions and sequences are possible as well.

4. KERNEL STICK-BREAKING MIXTURES

For a location-scale mixture model to have a large support, the mixing distribution has to have at least countably infinite support. In the previous section we defined countable mixtures by specifying a prior on the number of mixture components that has support on positive integers. Estimation of such models by reversible jump MCMC methods is feasible (Green, 1995); however, it could be complicated. A popular alternative for countable mixtures is Dirichlet process prior mixtures. A stick-breaking representation of the Dirichlet process introduced by Sethuraman (1994) proved to be especially convenient for specifying countable covariate dependent mixtures. In this section, we consider the kernel stick-breaking (KSB) mixture introduced by Dunson and Park (2008),

$$p(y|x,\theta) = \sum_{j=1}^{\infty} \pi_j(x) \phi\left(\frac{y-\mu_j}{\sigma_j}\right), \tag{4.1}$$

$$\pi_j(x) = \alpha_j K\left(-Q_j||x-q_j||^2\right) \prod_{l=1}^{j-1} \left\{1 - \alpha_l K\left(-Q_l||x-q_l||^2\right)\right\},$$

where θ, K, and ϕ were defined in Section 3. Even though mixing probabilities $\pi_j(x)$ look quite different from the mixing probabilities of KM in (3.1) we show in the following proposition that KSB mixtures can approximate KMs.

PROPOSITION 4.1. *(i) For any m, $\theta^{KM} \in \Theta$, and $\epsilon > 0$ there exist $\theta^{KSB} \in \Theta$ and n such that*

$$\int \log \frac{p(y|x,\theta^{KM},m)}{p(y|x,\theta^{KSB}_{1:n})} dF_0(y,x) < \epsilon, \tag{4.2}$$

where $p(y|x,\theta_{KM},m)$ is defined in (3.1) and $p(y|x,\theta^{KSB}_{1:n})$ is a truncated version of (4.1),

$$p\left(y|x,\theta^{KSB}_{1:n}\right) = \sum_{j=1}^{n} \pi_j(x)\phi\left(\frac{y-\mu_j}{\sigma_j}\right).$$

(ii) Under Assumptions 3.1–3.3, (4.2) holds on an open neighborhood of $\theta^{KSB}_{1:n}$.

The proof of the proposition is in Appendix A. Using this approximation result, we obtain weak and strong consistency in the following subsections.

4.1. Weak Consistency

To show that a KSB mixture is weakly consistent, we will prove that the KL property holds.

PROPOSITION 4.2. *Suppose Assumptions 3.1–3.3 hold and for any n, $\theta_{1:n}$, and an open neighborhood of $\theta_{1:n}$, Θ^n, $\Pi(\tilde{\theta}_{1:n} \in \Theta^n) > 0$. Then for $p(y|x,\theta)$ defined in (4.1) and any $\epsilon > 0$,*

$$\Pi\left(\theta : \int \log \frac{f_0(y|x)}{p(y|x,\theta)} dF_0(y,x) < \epsilon\right) > 0.$$

Proof. By Theorem 3.1 there exist m and $\theta^{KM} \in \Theta$ such that

$$\int \log\left(f_0(y|x)/p\left(y|x,\theta^{KM},m\right)\right) dF_0(y,x) < \epsilon/2.$$

By Proposition 4.1 there exist n, $\theta^{KSB}_{1:n}$, and an open neighborhood of $\theta^{KSB}_{1:n}$, Θ^n, such that for any $\tilde{\theta}^{KSB}_{1:n} \in \Theta^n$,

$$\int \log\left(p\left(y|x,\theta^{KM},m\right)/p\left(y|x,\tilde{\theta}^{KSB}_{1:n}\right)\right) dF_0(y,x) < \epsilon/2.$$

Let $\tilde{\theta}^{KSB} = \left(\tilde{\theta}^{KSB}_{1:n}, \tilde{\theta}^{KSB}_{n+1:\infty}\right) \in \Theta$, where $\tilde{\theta}^{KSB}_{1:n} \in \Theta^n$ and $\tilde{\theta}^{KSB}_{n+1:\infty}$ is an unrestricted continuation of $\tilde{\theta}^{KSB}_{1:n}$. Since $p\left(y|x,\tilde{\theta}^{KSB}\right) \geq p\left(y|x,\theta^{KSB}_{1:n}\right)$,

$$\begin{aligned}\int \log \frac{f_0(y|x)}{p(y|x,\tilde{\theta}^{KSB})} dF_0(y,x) \leq & \int \log \frac{f_0(y|x)}{p(y|x,\theta^{KM},m)} dF_0(y,x) \\ & + \int \log \frac{p(y|x,\theta^{KM},m)}{p(y|x,\theta^{KSB}_{1:n})} dF_0(y,x) < \epsilon.\end{aligned}$$

By the proposition, assumption $\Pi\left(\tilde{\theta}^{KSB}_{1:n} \in \Theta^n\right) > 0$ and the result follows. ■

By Theorem 2.1 the KL property implies the following weak posterior consistency theorem.

THEOREM 4.1. *Under the assumptions of Proposition 4.2, for any weak neighborhood U of $f_0(y|x)$,*

$\Pi(U|Y_T, X_T) \to 1, \quad$ *a.s.* F_0^∞.

4.2. Strong Consistency

To apply Theorem 2.2, we define sieves as follows. For a given $\delta > 0$ and a sequence m_n, let

$$\mathcal{F}_n = \left\{ p(y|x,\theta) : |\mu_j| \le \overline{\mu}_n,\ Q_j \le \overline{Q}_n, \qquad \underline{\sigma}_n < \sigma_j < \overline{\sigma}_n, \qquad j=1,\dots,m_n, \qquad \sup_{x\in X} \sum_{j=m_n+1}^{\infty} \pi_j(x) \le \delta \right\}.$$

The restriction on the mixing probabilities in the sieve definition is similar to the one used by Pati et al. (2013). We calculate a bound on the metric entropy of $\mathcal{F}_n$ in the following proposition.

PROPOSITION 4.3. *Suppose Assumptions 3.2 and 3.3 hold. Then*

$$J(4\delta, \mathcal{F}_n) \le m_n \left(\log \left[b_0 \frac{\overline{\mu}_n}{\underline{\sigma}_n} + b_1 \log \frac{\overline{\sigma}_n}{\underline{\sigma}_n} + 1 \right] + b_2 + b_3 \log \overline{Q}_n + b_4 \log m_n \right), \tag{4.3}$$

where b_0, b_1, b_2, b_3, and b_4 depend on δ but not on n, m_n, $\overline{Q}_n$, $\overline{\mu}_n$, $\overline{\sigma}_n$, and $\underline{\sigma}_n$.

A proof is given in Appendix A.

The next theorem formulates sufficient conditions for strong consistency.

THEOREM 4.2. *Suppose*

(i) A priori $(\alpha_j, \mu_j, \sigma_j, Q_j)$ are i.i.d. across j.

(ii) For any $\epsilon > 0$, there exist $\delta < \epsilon/16$, $\beta < \epsilon^2/8$, constants $c_1, c_2 > 0$, and sequences m_n, $\overline{Q}_n$, $\overline{\mu}_n$, $\overline{\sigma}_n \uparrow \infty$, and $\underline{\sigma}_n \downarrow 0$ with $\overline{\sigma}_n > \underline{\sigma}_n$ such that

$$m_n \left[\Pi(|\mu_j| > \overline{\mu}_n) + \Pi(\underline{\sigma}_n > \sigma_j) + \Pi(\sigma_j > \overline{\sigma}_n) + \Pi(Q_j > \overline{Q}_n) \right] \tag{4.4}$$
$$+ \Pi \left(\sup_{x\in X} \sum_{j=m_n+1}^{\infty} \pi_j(x) > \delta \right) \le c_1 e^{-c_2 n},$$

$$m_n \left(\log \left[b_0 \frac{\overline{\mu}_n}{\underline{\sigma}_n} + b_1 \log \frac{\overline{\sigma}_n}{\underline{\sigma}_n} + 1 \right] + b_2 + b_3 \log \overline{Q}_n + b_4 \log m_n \right) < n\beta, \tag{4.5}$$

where b_0, b_1, b_2, b_3, and b_4 are defined by Proposition 4.3.

(iii) The conditions of Theorem 4.1 hold.

Then the posterior is strongly consistent at f_0.

Theorem 4.2 is a direct consequence of Theorem 2.2 and Proposition 4.3. The difficulty in verifying the sufficient conditions of the theorem arises in finding a prior distribution and sieve parameters that satisfy the requirements that

$$\Pi\left(\sup_{x\in X}\sum_{j=m_n+1}^{\infty}\pi_j(x)>\delta\right)<c_1e^{-nc_2}$$

and $m_n\log\overline{Q}_n<n\beta$ for n large enough, as this requires delicate handling of mixing weights and prior distributions. Observe that $\sum_{j=m_n+1}^{\infty}\pi_j(x)=\prod_{j=1}^{m_n}(1-\alpha_jK(-Q_j||x-q_j||^2))$ and thus

$$\Pi\left(\sup_{x\in X}\sum_{j=m_n+1}^{\infty}\pi_j(x)>\delta\right)\le\Pi\left(\prod_{j=1}^{m_n}(1-\alpha_jK_j)>\delta\right), \tag{4.6}$$

where $K_j=K(-Q_jd_x)\le K(-Q_j||x-q_j||^2))$. The following lemma describes priors for α_j and Q_j that imply an exponential bound on the r.h.s of (4.6).

LEMMA 4.1. *If prior distributions of α_j first-order stochastically dominate $Beta(1,\gamma)$ and $K_j=K(-Q_jd_x)$ first-order stochastically dominates $Beta(\gamma+1,1)$ for any $\gamma>0$, then*

$$\Pi_\theta\left(\prod_{j=1}^{m_n}(1-\alpha_jK_j)>\delta\right)<e^{-0.5m_n\log m_n}.$$

The lemma is proved in Appendix A. With the result of the lemma, we are ready to present an example of priors that satisfy the conditions of Theorem 4.2.

Example 4.1

Suppose priors for μ and σ and sequences $\overline{\mu}_n$, $\underline{\sigma}_n$, and $\overline{\sigma}_n$ are the same as in Example 3.1 (normal and inverse gamma priors). Then for $m_n=cn/\log n$ and $\overline{Q}_n=n^r$, where c and r are constants, condition (4.5) is satisfied for c sufficiently small.

By Lemma 4.1 condition (4.4) is satisfied if the prior distributions for α_j first-order stochastically dominate $Beta(1,\gamma)$ and $K(-Q_jd_x)$ first-order stochastically dominate $Beta(\gamma+1,1)$ for any $\gamma>0$ (note that for $m_n=cn/\log n$, $\exp(-0.5m_n\log m_n)\le\exp(-0.25cn)$ for large enough n).

Explicit priors for Q_j and α_j satisfying the sufficient conditions can be constructed for particular choices of $K(\cdot)$. For example, for $K(\cdot)=\exp(\cdot)$, $\alpha_j\sim Beta(1,\gamma)$ and $Q_j\sim Exponential((\gamma+1)d_x)$, which is equivalent to $K_j=\exp(-Q_jd_x)\sim Beta(\gamma+1,1)$, satisfy the conditions of Lemma 4.1. Also, $\Pi(Q_j>n^r)\le c_1e^{-nc_2}$ for $r\ge1$.

5. COVARIATE DEPENDENT LOCATIONS

It has been suggested in the literature (Geweke and Keane (2007), Villani et al. (2009)) that covariate dependent mixture models in which locations also depend

on covariates perform well in applications. It is not surprising that weak and strong posterior consistency can be established for such models, as they are generalizations of the models we considered above. Specifically, let $z : X \to Z \subset R^{d_z}$ denote a transformation of the original covariate x. For example, $z(x)$ can be x itself or include polynomials of x. Kernel mixtures and kernel stick-breaking mixtures with covariate dependent locations can be defined by (3.1) and (4.1) with μ_j replaced by $\beta_j' z(x)$, where $\beta_j \in R^{d_z}$ for each j. We make the following assumption on the space Z and the function z to extend the consistency results to this setup.

Assumption 5.1.

(i) $Z = [0,1]^{d_z}$,

(ii) $z(x)_1 = 1$ (the first coordinate) for any $x \in X$.

Under this assumption, each location-scale density can still have a constant location, and all the theoretical results for models with constant locations μ_j presented in Sections 3–4 hold for models with locations $\beta_j' z(x)$ with minor modifications. Theorem A.1 in Appendix A provides the details. The theorem implies that the priors from Examples 3.1 and 4.1, in which a normal prior for μ_j is replaced by independent normal priors on components of β_j, guarantee strong posterior consistency for models with covariate dependent locations.

6. FINITE SAMPLE PERFORMANCE

In this section we assess the finite sample performance of a Bayesian conditional density estimator based on a kernel stick-breaking mixture prior. We do not consider an estimator based on kernel mixtures from Section 3 because similar models have been extensively studied in the literature; see, for example, Geweke and Keane (2007), Villani et al. (2009), and Norets and Pelenis (2012). First, we discuss the model setup and prior specification. Second, we present a graphical illustration of the estimator performance and a comparison with a kernel smoothing estimator for simulated data sets of different sizes. Third, we conduct Monte Carlo studies comparing the KSB estimator and a kernel smoothing estimator for two DGPs from the previous literature.

The model we use in the simulation exercises is a special case of the models discussed in Section 5 with locations linear in covariates,

$$p(y|x,\theta) = \sum_{j=1}^{\infty} \alpha_j K\left(-Q_j||x-q_j||^2\right) \prod_{l=1}^{j-1} \left\{1 - \alpha_l K\left(-Q_l||x-q_l||^2\right)\right\} \times \phi\left(\frac{y-\beta_{j,0}-\beta_{j,1}'x}{\sigma_j}\right), \tag{6.1}$$

where $K(\cdot) = \exp(\cdot)$ and ϕ is a normal density.

The prior distribution we use is based on Example 4.1: $\beta_j \sim N(\mu_\beta, H_\beta^{-1})$, $\sigma_j^2 \sim InvGamma(\nu, b_\sigma)$, $\alpha_j \sim Beta(a,b)$, $q_j \sim U(0,1)$, $Q_j \sim Exponential(\tau)$ i.i.d. across j, where $\{\mu_\beta, H_\beta, \nu, b_\sigma, a, b, \tau\}$ are fixed hyperparameters. In actual applications, the values of hyperparameters can be selected to reflect the researcher's beliefs about the density. We use data dependent values,

$$\mu_\beta = (\overline{y}, 0)', \; H_\beta = \text{diag}\left(0.5/\hat{\sigma}_y^2, 1\right), \; \nu = 3, \; b_\sigma = 2/\hat{\sigma}_y^2, \; a = 1, \; b = 0.5, \; \tau = 1.5, \tag{6.2}$$

where $\overline{y}$ and $\hat{\sigma}_y^2$ are the sample mean and variance. Parameters $\{\beta_j\}_{j=1}^{\infty}$ control the impact of covariates on the response for each location-scale density, and values of (μ_β, H_β) are chosen so that observed values of y are plausible. Parameters $\{\sigma_j^2\}_{j=1}^{\infty}$ control the variance in each mixture component and, thus, should reflect the range of observable y at possible covariate values. The values of hyperparameters (ν, b_σ) are chosen so that the observed variances of the response variable are plausible. Parameters $\{\alpha_j, Q_j\}_{j=1}^{\infty}$ control the expected number of mixture components and borrowing of information across covariates. Values of (a,b) that imply a prior for α_j concentrating near 1 lead to mixtures with a smaller number of components. Higher values of τ imply the prior for Q_j concentrating near 0. Smaller Q_j in turn implies a smaller number of mixture components and more information sharing across covariate values. Sufficient conditions for the strong consistency are satisfied if $a = 1$ (a standard choice in the literature) and $1 + b \leq \tau / d_x$ (see Example 4.1, Lemma 4.1, and Section 5). It appear that large values of τ lead to considerable oversmoothing. Thus, we suggest using low values of τ.

To estimate the model, we develop an MCMC algorithm based on slice sampling (Neal, 2003; Walker, 2007) and retrospective sampling (Papaspiliopoulos and Roberts, 2008; Papaspiliopoulos, 2008). The algorithm is described in detail in Appendix B. To check the correctness of the algorithm design and implementation we used the joint distribution tests from Geweke (2004) and related tests from Cook, Gelman, and Rubin (2006).

The first of the two DGPs we consider is taken from Section 5 of Dunson and Park (2008): $x_i \sim U[0,1]$ and the true conditional density is

$$f_0(y_i|x_i) = e^{-2x_i} N\left(y_i; x_i, 0.1^2\right) + \left(1 - e^{-2x_i}\right) N\left(y_i; x_i^4, 0.2^2\right). \tag{6.3}$$

According to Dunson and Park (p. 315), this DGP is a "challenging example as the shape of the conditional density changes rapidly, with limited sample size in any particular local region." Furthermore, the same setup for the simulation exercise enables a comparison with the results in Dunson, Pillai, and Park (2007) and Dunson and Park.

Figure 1 presents the DGP (6.3) conditional densities and posterior means of the estimated conditional densities. Each column in the figure shows densities conditional on a particular value of covariate $x \in \{0.25, 0.5, 0.75\}$. The rows correspond to different sample sizes of the simulated data, $N \in \{200, 500, 2,000\}$.

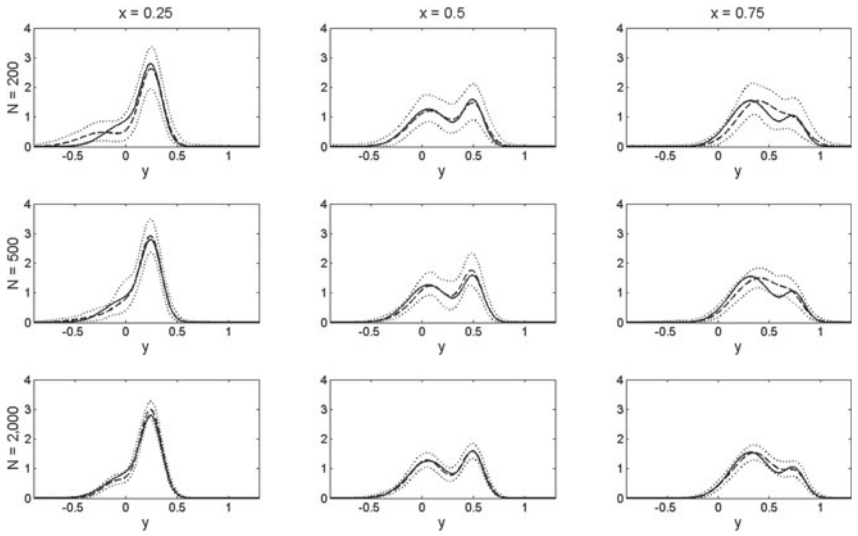

FIGURE 1. Estimated conditional densities for different covariate values and sample sizes. The solid lines are the true values, the dashed lines are the posterior means, and the dotted lines are pointwise 99% equal-tailed credible intervals.

For estimation, we perform 400, 000 MCMC iterations, of which the first 100, 000 are discarded for burn-in. For plots, we use every 20th of the remaining iterations. The separated partial means test for the first and second moments of the conditional density draws suggests that the MCMC chains converge. The numerical standard errors (NSEs) of conditional density estimates are less than 0.02 (the average of NSEs over all (y, x) is 0.002). The fit is comparable to the results obtained by Dunson and Park (2008) for a slightly different model. Note that for larger sample sizes the width of posterior credible intervals is smaller, which is expected from our posterior consistency results.

To assess the sensitivity of the estimation results with respect to prior specification, we repeated the estimation exercise with various modifications of prior hyperparameters. In summary, prior hyperparameters $(\mu_\beta, H_\beta, \nu, b_\sigma)$ do not seem to affect the results as long as they imply a plausible range of response variables. At the same time, hyperparameters (b, τ) can have a considerable effect on the estimation results. Smaller values of τ (and, thus, by Example 4.1, b) seem to deliver better results, as they lead to stronger dependence of mixing weights on covariates, which allows the model to accommodate sudden changes in conditional densities in the DGP. The details of prior sensitivity analysis are delegated to Appendix C.

Additionally, we compare the KSB model with the nonparametric kernel smoothing method of Hall, Racine, and Li (2004) implemented by Hayfield and Racine (2008) in the publicly available R package **np**. Figure 2 shows that the estimation results for the DGP (6.3) from both approaches are pretty close.

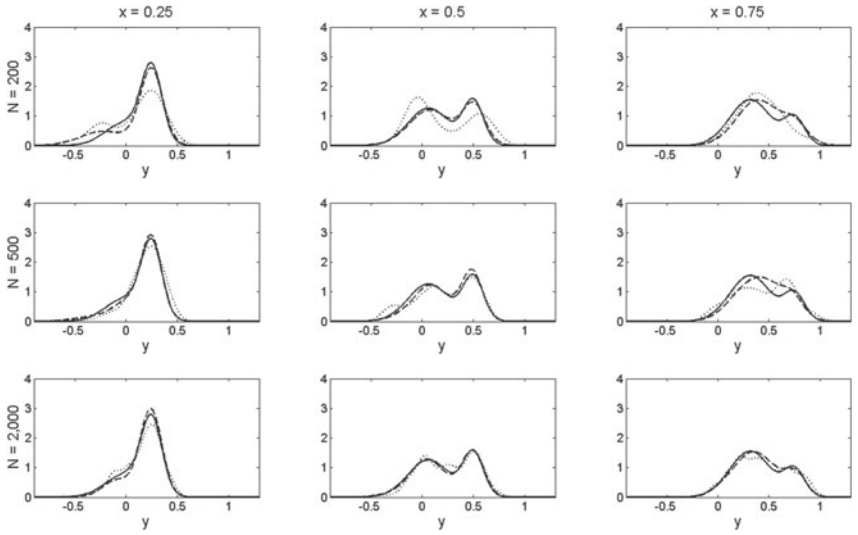

FIGURE 2. Estimated conditional densities for different covariate values and different sample sizes. The solid lines are the true values, the dashed lines are the posterior means, and the dotted lines are the kernel estimate of conditional densities $p(y|x)$.

Furthermore, we conduct a Monte Carlo study to compare the two estimators. For the DGP defined in equation (6.3), we simulated 100 samples of size $N = 500$. For each sample, the performance of an estimator is evaluated by the root mean squared error (RMSE) and the mean absolute error (MAE),

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_y}\sum_{j=1}^{N_x}\left(\hat{f}(y_i|x_j) - f_0(y_i|x_j)\right)^2}{N_y N_x}},$$

$$MAE = \frac{\sum_{i=1}^{N_y}\sum_{j=1}^{N_x}\left|\hat{f}(y_i|x_j) - f_0(y_i|x_j)\right|}{N_y N_x},$$

where (y_i, x_j) are evenly distributed grid points with $y_i \in \{-0.88, -0.84, \ldots, 1.08\}$ and $x_j \in \{0.01, 0.03, \ldots, 0.99\}$. Table 1 provides the averages and the sample standard deviations of the RMSE and the MAE for the KSB and the kernel smoothing methods and their ratios.

In this particular example, the KSB model outperforms the nonparametric kernel smoothing method of Hall et al. (2004) by RMSE and MAE criteria.

We also conduct a Monte Carlo study for the DGP from Hall, Wolff, and Yao (1999),

$$y_i = 2\sin(\pi x_i) + \epsilon_i, \tag{6.4}$$

where x_i and ϵ_i are i.i.d random variables with a density $1 - |x|$ on $[-1, 1]$. Again, we evaluate the relative performance by RMSE and MAE on evenly distributed

TABLE 1. Monte Carlo study for DGP (6.3)

Method	RMSE	MAE
KSB	0.148(0.019)	0.085(0.011)
np	0.195(0.017)	0.109(0.010)
Ratio KSB/**np**	0.758(0.092)	0.784(0.101)

TABLE 2. Monte Carlo study for DGP (6.4)

Method	RMSE	MAE
KSB	0.149(0.010)	0.074(0.005)
np	0.119(0.017)	0.051(0.005)
Ratio KSB/**np**	1.270(0.176)	1.478(0.149)

grid points with $y_i \in \{-2.94, -2.82, \ldots, 2.94\}$ and $x_j \in \{-0.98, -0.94, \ldots, 0.98\}$ for 100 random samples of size $N = 500$. The results are summarized in Table 2.

As can be seen from Tables 1 and 2, the kernel smoothing estimator outperforms the KSB estimator for the DGP in (6.4) by a margin similar to the one by which the latter outperforms the former for the DGP in (6.3). These results suggest that these two approaches have comparable small sample performance.

7. DISCUSSION

The regularity conditions on the DGP assumed in proving weak and strong posterior consistency are very mild. The conditions require that the tails of the mixed location-scale density not be too thin relative to the data-generating density. They also require the local changes in the logged data-generating density to be integrable.

Weak posterior consistency is proved under no special requirements on the prior for parameters beyond conditions on the support (0 has to be in the support of the scale parameters, and the support of location parameters has to be unbounded).

Assumptions about the prior for the location and scale parameters of the mixed densities employed in showing strong posterior consistency are similar under both types of mixing. They are in the spirit of the assumptions employed in previous work on the estimation of unconditional densities. Examples of priors that satisfy the assumptions include the normal priors for locations and inverse gamma for squared scales commonly used in practice.

Although the parameters entering the mixing probabilities under the two types of mixing are the same, the mixing probabilities are constructed differently. This seems to require different priors for attaining strong posterior consistency under the two types of mixing. For kernel mixtures with a variable number of components, there are no restrictions on the constants multiplying the kernels. For stick-breaking mixtures, these constants are assumed to have a prior that puts more mass on values of the constants that are close to 1 (see Lemma 4.1). The inverse of the scales of the mixing probability kernels may have thicker tails under

stick-breaking mixtures. The prior for locations of the mixing probability kernels is not restricted under both types of mixing, which is not surprising given that the space for covariates is assumed to be bounded.

It would be desirable to derive posterior convergence rates to get more insight into covariate dependent mixture models. The techniques for obtaining bounds on posterior convergence rates are similar to those for obtaining strong posterior consistency (Ghosal et al., 2000). However, bounds on convergence rates are mostly of interest if they are tight (for example, if they are close to minimax rates for certain classes of DGPs). As we mention in the Introduction, the results of this type were obtained for unconditional density estimation. In order to obtain such results for our models, one needs to improve the Kullback-Leibler approximation results from Section 3. We leave this problem to future research.

REFERENCES

Barron, A. (1988) The Exponential Convergence of Posterior Probabilities with Implications for Bayes Estimators of Density Functions, Working paper, University of Illinois.

Barron, A., M.J. Schervish, & L. Wasserman (1999) The consistency of posterior distributions in nonparametric problems. *Annals of Statistics* 27, 536–561.

Billingsley, P. (1999) *Convergence of Probability Measures*. Wiley-Interscience.

Blackwell, D. & J.B. MacQueen (1973) Ferguson distributions via polya urn schemes. *Annals of Statistics* 1, 353–355.

Chen, X. (2007) Large sample sieve estimation of semi-nonparametric models. In J. Heckman & E. Leamer (eds), *Handbook of Econometrics*, vol. 6 of *Handbook of Econometrics*, ch. 76. Elsevier.

Chung, Y. & D.B. Dunson (2009) Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association* 104, 1646–1660.

Cook, S.R., A. Gelman, & D.B. Rubin (2006) Validation of software for bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics* 15, 675–692.

De Iorio, M., P. Muller, G.L. Rosner, & S.N. MacEachern (2004) An ANOVA model for dependent random measures. *Journal of the American Statistical Association* 99, 205–215.

Dey, D., P. Muller, & D. Sinha (eds.) (1998) *Practical Nonparametric and Semiparametric Bayesian Statistics*, Lecture Notes in Statistics, Vol. 133. Springer.

DiNardo, J. & J.L. Tobias (2001) Nonparametric density and regression estimation. *Journal of Economic Perspectives* 15, 11–28.

Dunson, D.B. & J.-H. Park (2008) Kernel stick-breaking processes. *Biometrika* 95, 307–323.

Dunson, D.B., N. Pillai, & J.-H. Park (2007) Bayesian density regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 69, 163–183.

Escobar, M. & M. West (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90, 577–588.

Escobar, M.D. (1994) Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association* 89, 268–277.

Fan, J. (2005) A selective overview of nonparametric methods in financial econometrics. *Statistical Science* 20, 317–337.

Ferguson, T.S. (1973) A Bayesian analysis of some nonparametric problems, *Annals of Statistics* 1, 209–230.

Genovese, C.R. & L. Wasserman (2000) Rates of convergence for the gaussian mixture sieve. *Annals of Statistics* 28, 1105–1127.

Geweke, J. (2004) Getting it right: joint distribution tests of posterior simulators. *Journal of the American Statistical Association* 99, 799–804.

Geweke, J. (2005) *Contemporary Bayesian Econometrics and Statistics*. Wiley-Interscience.
Geweke, J. & M. Keane (2007) Smoothly mixing regressions. *Journal of Econometrics* 138, 252–290.
Ghosal, S., J.K. Ghosh, & R.V. Ramamoorthi (1999) Posterior consistency of dirichlet mixtures in density estimation. *Annals of Statistics* 27, 143–158.
Ghosal, S., J.K. Ghosh, & A.W. ven der Vaart (2000) Convergence rates of posterior distributions. *Annals of Statistics* 28, 500–531.
Ghosal, S. & Y. Tang (2006) Bayesian consistency for Markov processes. *Sankhya* 68, 227–239.
Ghosh, J. & R. Ramamoorthi (2003) *Bayesian Nonparametrics*. 1st ed. Springer.
Green, P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 711–732.
Griffin, J.E. & M.F.J. Steel (2006) Order-based dependent Dirichlet processes. *Journal of the American Statistical Association* 101, 179–194.
Hall, P., J. Racine, & Q. Li (2004) Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* 99, 1015–1026.
Hall, P., R.C.L. Wolff, & Q. Yao (1999) Methods for estimating a conditional distribution function. *Journal of the American Statistical Association* 94, 154–163.
Hayfield, T. & J.S. Racine (2008) Nonparametric econometrics: The np package. *Journal of Statistical Software* 27, 1–32.
Ichimura, H. & P.E. Todd (2007) Implementing nonparametric and semiparametric estimators. In *Handbook of Econometrics*, vol. 6, Part B, pp. 5369–5468. Elsevier.
Jacobs, R.A., M.I. Jordan, S.J. Nowlan, & G.E. Hinton (1991) Adaptive mixtures of local experts. *Neural Computation* 3, 79–87.
Jambunathan, M.V. (1954) Some properties of beta and gamma distributions. *Annals of Mathematical Statistics* 25, 401–405.
Jordan, M. & L. Xu (1995) Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks* 8, 1409–1431.
Koenker, R. & K.F. Hallock (2001) Quantile regression. *Journal of Economic Perspectives* 15, 143–156.
Kruijer, W., J. Rousseau, & A. van der Vaart (2010) Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics* 4, 1225–1257.
Li, F., M. Villani, & R. Kohn (2010) Flexible modeling of conditional distributions using smooth mixtures of asymmetric student t densities. *Journal of Statistical Planning and Inference* 140, 3638–3654.
Li, J.Q. & A.R. Barron (1999) Mixture density estimation. In *Advances in Neural Information Processing Systems 12*, pp. 279–285. MIT Press.
Liang, S., B.P. Carlin, & A.E. Gelfand (2009) Analysis of Minnesota colon and rectum cancer point patterns with spatial and nonspatial covariate information. *Annals of Applied Statistics* 3, 943–962.
MacEachern, S.N. (1999) Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*.
McLachlan, G. & D. Peel (2000) *Finite Mixture Models*. Wiley.
Muller, P., A. Erkanli, & M. West (1996) Bayesian curve fitting using multivariate normal mixtures. *Biometrika* 83, 67–79.
Neal, R.M. (2003) Slice sampling. *Annals of Statistics* 31, 705–767, with discussions and a rejoinder by the author.
Norets, A. (2010) Approximation of conditional densities by smooth mixtures of regressions. *Annals of Statistics* 38, 1733–1766.
Norets, A. & J. Pelenis (2012) Bayesian modeling of joint and conditional distributions. *Journal of Econometrics* 168, 332–346.
Papaspiliopoulos, O. (2008) A note on Posterior sampling from Dirichlet mixture models. Preprint.
Papaspiliopoulos, O. & G.O. Roberts (2008) Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* 95, 169–186.
Pati, D., D.B. Dunson, & S.T. Tokdar (2013) Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis* 116, 456–472.

Peng, F., R.A. Jacobs, & M.A. Tanner (1996) Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association* 91, 953–960.

Roeder, K. & L. Wasserman (1997) Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* 92, 894–902.

Rousseau, J. (2010) Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Annals of Statistics* 38, 146–180.

Schwartz, L. (1965) On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 4, 10–26.

Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statistica Sinica* 4, 639–650.

Taddy, M.A. & A. Kottas (2010) A Bayesian nonparametric approach to inference for quantile regression. *Journal of Business and Economic Statistics* 28, 357–369.

Tokdar, S. (2007) Towards a faster implementation of density estimation with logistic Gaussian process priors. *Journal of Computational and Graphical Statistics* 16, 633–655.

Tokdar, S., Y. Zhu, & J. Ghosh (2010) Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Analysis* 5, 319–344.

Tokdar, S.T. (2006) Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhya : The Indian Journal of Statistics* 67, 99–100.

Tokdar, S.T. & J.K. Ghosh (2007) Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference* 137, 34–42.

Tran, M.-N., D.J. Nott, & R. Kohn (2012) Simultaneous variable selection and component selection for regression density estimation with mixtures of heteroscedastic experts. *Electronic Journal of Statistics* 6, 1170–1199.

van der Vaart, A.W. & J.H. van Zanten (2008) Rates of contraction of posterior distributions based on Gaussian process priors. *Annals of Statistics* 36, 1435–1463.

van der Vaart, A.W. & J.H. van Zanten (2009) Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Annals of Statistics* 37, 2655–2675.

Villani, M., R. Kohn, & P. Giordani (2009) Regression density estimation using smooth adaptive Gaussian mixtures. *Journal of Econometrics* 153, 155–173.

Villani, M., R. Kohn, & D.J. Nott (2012) Generalized smooth finite mixtures. *Journal of Econometrics* 171, 121–133.

Walker, S.G. (2004) Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics* 32, 2028–2043.

Walker, S.G. (2007) Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation* 36, 45–54.

Wood, S., W. Jiang, & M. Tanner (2002) Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika* 89, 513–528.

Wu, Y. & S. Ghosal (2010) The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *Journal of Multivariate Analysis* 101, 2411–2419.

Yatchew, A. (1998) Nonparametric regression techniques in economics. *Journal of Economic Literature* 36, 669–721.

Zeevi, A.J. & R. Meir (1997) Density estimation through convex combinations of densities: approximation and estimation bounds. *Neural Networks* 10, 99–109.

APPENDIX A: Proofs

Proof (Theorem 3.1).

The theorem can be proved by exhibiting a sequence of m and $\theta_{1:m}$ such that

$$\int \log \frac{f_0(y|x)}{p(y|x,\theta,m)} dF_0(y,x) \to 0.$$

Since d_{KL} is always nonnegative,

$$0 \le \int \log \frac{f_0(y|x)}{p(y|x,\theta_{1:m},m)} F_0(dy,dx) \le \int \log \max\left\{1, \frac{f_0(y|x)}{p(y|x,\theta_{1:m},m)}\right\} F_0(dy,dx).$$

Thus, it suffices to show that the last integral in the inequality above converges to zero as m increases. The dominated convergence theorem (DCT) is used for that. First, we demonstrate the pointwise convergence of the integrand to zero a.s. F_0. Then, we present an integrable upper bound on the integrand required by the DCT. To define m and $\theta_{1:m}$, we first define partitions of Y and X.

Let A_j^m, $j = 0, 1, \ldots, m_y$, be a partition of Y consisting of adjacent half-open half-closed intervals $A_1^m, \ldots, A_{m_y}^m$ with length h_m and the rest of the space A_0^m. As m increases, the fine part of the partition becomes finer, $h_m \to 0$, and $m_y \to \infty$. Also, it covers larger and larger parts of Y: For any $y \in Y$ there exists M_0 such that

$$\forall m \ge M_0, \qquad C_{\delta_m}(y) \cap A_0^m = \emptyset, \tag{A.1}$$

where $C_{\delta_m}(y)$ is an interval with center y and half-length $\delta_m \to 0$. It is always possible to construct such a partition. For example, if $Y = (-\infty, \infty)$ let $A_0^m = (-\infty, -\log m_y] \cup [\log m_y, \infty)$, $A_j^m = [-\log m_y + 2(j-1)\log m_y/m_y, -\log m_y + 2j \log m_y/m_y)$ for $j \ne 0$, and $h_m = 2\log m_y/m_y$.

Let B_i^m, $i = 1, \ldots, m_x$ be equal-size half-open half-closed hypercubes forming a partition of $X = [0,1]^{d_x}$. Note $m = (m_y + 1) \cdot m_x$. The partition becomes finer as m increases, $\lambda(B_i^m) = m_x^{-1} \to 0$, where λ is the Lebesgue measure. Let q_i^m denote the center of B_i^m.

Taking into account that $\sum_{j=0}^{m_y} F_0(A_j^m|q_i^m) = 1$, define m and $\theta_{1:m}$ as

$$p(y|x,\theta,m) = \frac{\sum_{i=1}^{m_x} \left[\sum_{j=1}^{m_y} F_0(A_j^m|q_i^m)\phi(y,\mu_j^m,\sigma_m) + F_0(A_0^m|q_i^m)\phi(y,0,\sigma_0)\right] K(-Q^m||x-q_i^m||^2)}{\sum_{i=1}^{m_x} K\left(-Q^m||x-q_i^m||^2\right)},$$

where σ_0 is fixed, σ_m converges to zero as m increases, and μ_j^m is the center of A_j^m. One can always construct a partition A_j^m so that

$$\delta_m \to 0,\ \sigma_m/\delta_m \to 0,\ h_m/\sigma_m \to 0; \tag{A.2}$$

for example, in the example from two paragraphs above, let $\sigma_m = h_m^{0.5}$ and $\delta_m = h_m^{0.25}$.

Also, under Assumption 3.3 it is always possible to define a positive diverging to infinity sequence Q^m and a sequence s_m (the squared diagonal of B_i^m) satisfying

$$\frac{K(-2Q^m s_m)}{K(-Q^m s_m) s_m^{d_x/2}} \to 0, \qquad s_m = d_x \lambda(B_i^m)^{2/d_x} \to 0. \tag{A.3}$$

For example, one can set $Q^m = s_m^{-2}$. This condition specifies that Q^m should increase fast relative to how fine the partition of X becomes.

Define $I_1^m(x,s_m) = \{i : ||q_i^m - x||^2 \le 2s_m\}$ and $I_2^m(x,s_m) = \{i : ||q_i^m - x||^2 > 2s_m\}$. Since s_m is the squared diagonal of B_i^m, there exists $i \in I_1^m(x,s_m)$ such that

$$K\left(-Q^m||x-q_i^m||^2\right) \ge K\left(-Q^m s_m\right). \tag{A.4}$$

For all $i \in I_2^m(x, s_m)$,

$$K\left(-Q^m||x-q_i^m||^2\right) \le K\left(-2Q^m s_m\right). \tag{A.5}$$

Note that

$$\begin{aligned}
&\frac{\sum_{i\in I_1^m(x,s_m)} K(-Q^m||x-q_i^m||^2)}{\sum_{i=1}^{m_x} K(-Q^m||x-q_i^m||^2)} \\
&\quad \ge 1 - \frac{\sum_{i\in I_2^m(x,s_m)} K(-Q^m||x-q_i^m||^2)}{\sum_{i\in I_1^m(x,s_m)} K(-Q^m||x-q_i^m||^2)} \\
&\quad \ge 1 - \frac{\mathrm{card}(I_2^m(x,s_m))K(-2Q^m s_m)}{K(-Q^m s_m)} \ge 1 - d_x^{d_x/2}\frac{K(-2Q^m s_m)}{K(-Q^m s_m)s_m^{d_x/2}},
\end{aligned} \tag{A.6}$$

where the second inequality follows from (A.4) and (A.5). The last inequality follows from $\mathrm{card}(I_2^m(x,s_m)) \le m_x = d_x^{d_x/2}s_m^{-d_x/2}$.

For $i \in I_1^m(x,s_m)$ and $A_j^m \subset C_{\delta_m}(y)$,

$$F\left(A_j^m|x_i^m\right) \ge \lambda(A_j^m) \inf_{z\in C_{\delta_m}(y),\ ||t-x||^2\le 2s_m} f(z|t). \tag{A.7}$$

Inequalities (A.6), (A.7), and Lemma A.3 imply that $p(y|x,\theta,m)$ exceeds

$$\begin{aligned}
&\sum_{j:A_j^m\subset C_{\delta_m}(y)} \sum_{i\in I_1^m(x,s_m)} F\left(A_j^m|q_i^m\right)\frac{K(-Q^m||x-q_i^m||^2)}{\sum_l K(-Q^m||x-q_l^m||^2)}\phi\left(y,\mu_j^m,\sigma_m\right) \\
&\qquad \ge \inf_{z\in C_{\delta_m}(y),\ ||t-x||^2\le 2s_m} f(z|t) \\
&\cdot\left[1-\frac{6\psi(0)h_m}{\sigma_m} - 2\int_{\delta_m/\sigma_m}^{\infty}\psi(\mu)d\mu\right]\cdot\left[1-d_x^{d_x/2}\frac{K(-Q^m s_m)}{K(-Q^m s_m/2^2)s_m^{d_x/2}}\right].
\end{aligned} \tag{A.8}$$

By (A.2) and (A.3), given some $\epsilon_1 > 0$, there exists M_1 such that for $m \ge M_1$ the product in the last line of (A.8) is bounded below by $(1-\epsilon_1)$.

If $f_0(y|x)$ is continuous at (y,x) and $f_0(y|x) > 0$, there exists M_2 such that for $m \ge M_2$, $[f_0(y|x)/\inf_{z\in C_{\delta_m}(y),\ ||t-x||^2\le 2s_m} f_0(z|t)] \le (1+\epsilon_1)$ since $\delta_m, s_m \to 0$. For any $m \ge \max\{M_1, M_2\}$

$$\begin{aligned}
1 &\le \max\left\{1, \frac{f(y|x)}{p(y|x,\theta,m)}\right\} \\
&\le \max\left\{1, \frac{f_0(y|x)}{\inf_{z\in C_{\delta_m}(y),\ ||t-x||^2\le 2s_m} f_0(z|t)(1-\epsilon_1)}\right\} \le \frac{1+\epsilon_1}{1-\epsilon_1}.
\end{aligned}$$

Thus, $\log\max\{1, f_0(y|x)/p(y|x,\theta,m)\} \to 0$ a.s. F as long as $f(y|x)$ is continuous in (y,x) a.s. F_0 ($f_0(y|x)$ is always positive a.s. F_0).

Let us derive an integrable upper bound for the DCT,

$$p(y|x,\theta,m) \geq \left[1 - d_x^{d_x/2}\frac{K(-2Q^m s_m)}{K(-Q^m s_m)s_m^{d_x/2}}\right] \cdot \left([1 - 1_{A_0^m}(y)] \cdot \inf_{||z-y||\leq r,\, ||t-x||\leq r} f_0(z|t) \cdot \sum_{j: A_j^m \subset C_r(y)\cap (A_0^m)^c} \lambda(A_j^m)\phi(y,\mu_j^m,\sigma_m) + 1_{A_0^m}(y) \cdot \inf_{||z-y||\leq r,\, ||t-x||\leq r} f_0(z|t) \cdot \lambda(C_r(y)\cap A_0^m)\phi(y,0,\sigma_0) \right). \tag{A.9}$$

For any m larger then some M_3, the Riemann sum in (A.9) is bounded below by $1/4$ (by Lemma A.3) and

$$\left[1 - d_x^{d_x/2}\frac{K(-2Q^m s_m)}{K(-Q^m s_m)s_m^{d_x/2}}\right] \geq 1/2$$

(by (A.3)).

Choose σ_0 so that for $y \in A_0^m$, $1 > 1/4 \geq \lambda(C_r(y)\cap A_0^m)\phi(y,0,\sigma_0) \geq r\phi(y,0,\sigma_0)$, for example, $\sigma_0 = 8r\,\psi(0)$. Then

$$\begin{aligned}
&\log\max\left\{1, \frac{f_0(y|x)}{p(y|x,\theta,m)}\right\} \\
&\leq \log\max\left\{1, \frac{f_0(y|x)}{\inf_{||z-y||\leq r,\, ||t-x||\leq r} f_0(z|t)\cdot\phi(y,0,\sigma_0)\cdot(r/2)}\right\} \\
&= \log\frac{1}{\phi(y,0,\sigma_0)(r/2)}\max\left\{\phi(y,0,\sigma_0)(r/2), \frac{f_0(y|x)}{\inf_{||z-y||\leq r,\, ||t-x||\leq r} f_0(z|t)}\right\} \\
&\leq -\log(\phi(y,0,\sigma_0)(r/2)) + \log\frac{f_0(y|x)}{\inf_{||z-y||\leq r,\, ||t-x||\leq r} f_0(z|t)}.
\end{aligned} \tag{A.10}$$

The first expression in (A.10) is integrable by Assumption 3.2 part (ii). The second expression in (A.10) is integrable by Assumption 3.1 (iii). Thus the proposition is proved. ■

Proof (Proposition 3.1).
The proof extends ideas from Theorem 6 in Ghosal et al. (1999) and Lemma 4.1 in Tokdar (2006) to general location scale densities and covariate dependent mixing weights.

A generic element of $\mathcal{F}_n$ is a mixture with m_n components (a mixture with the number of components smaller than m_n is a special case with some α_j's equal to zero). It is proved in Lemma A.1 below that for $p(y|x,\theta_{m_n}^i, m_n) \in \mathcal{F}_n$, $\theta_{m_n}^i = \{Q_j^i, \mu_j^i, \sigma_j^i, q_j^i, \alpha_j^i\}_{j=1}^{m_n}$, $i = 1,2$, and any $x \in X$,

$$\begin{aligned}
&\int |p\left(y|x,\theta_{m_n}^1, m_n\right) - p\left(y|x,\theta_{m_n}^2, m_n\right)|dy \\
&\leq 2\max_{j=1,\ldots,m_n}\left(\psi(0)\frac{|\mu_j^1 - \mu_j^2|}{\sigma_j^1} + \frac{|\sigma_j^1 - \sigma_j^2|}{\min(\sigma_j^1,\sigma_j^2)}\right)
\end{aligned} \tag{A.11}$$

$$+\frac{2}{K(-\overline{Q}_n d_x)}\sum_{j=1}^{m_n}|\widetilde{\alpha}_j^1-\widetilde{\alpha}_j^2|,$$

$$+\frac{2\overline{K}' d_x}{K(-\overline{Q}_n d_x)}\max_{j=1,\ldots,m_n}|Q_j^2-Q_j^1|$$

$$+\frac{4\overline{K}' d_x \overline{Q}_n}{K(-\overline{Q}_n d_x)}\max_{j=1,\ldots,m_n}\max_{l=1,\ldots,d_x}|q_{j,l}^2-q_{j,l}^1|,$$

where $\overline{K}'$ is a finite fixed bound on the derivative of K (Assumption 3.3) and $\widetilde{\alpha}_j^i = \alpha_j^i/\sum_{l=1}^{m_n}\alpha_l^i$.

The outline of the argument below is as follows. We define grids $G_{\mu\sigma}=\{(\mu_i,\sigma_i), i=1,\ldots,N_{\mu\sigma}\}$, $G_Q=\{Q_i, i=1,\ldots,N_Q\}$, $G_q=\{q_i, i=1,\ldots,N_q\}$, and $G_\alpha=\{\alpha_i=(\alpha_{i1},\ldots,\alpha_{im_n}), i=1,\ldots,N_\alpha\}$ on sets $[-\overline{\mu}_n,\overline{\mu}_n]\times[\underline{\sigma}_n,\overline{\sigma}_n]$, $[0,\overline{Q}_n]$, $[0,1]^{d_x}$, and $[0,1]^{m_n}$ correspondingly. Then, we show that for any $\theta_{m_n}^1$ with $p(y|x,\theta_{m_n}^1,m_n)\in\mathcal{F}_n$ there exists $\theta_{m_n}^2\in\left[G_{\mu\sigma}\times G_Q\times G_q\right]^{m_n}\times G_\alpha$ such that $||p(y|x,\theta_{m_n}^1,m_n)-p(y|x,\theta_{m_n}^2,m_n)||_1\le\delta$. Thus, $J(\delta,\mathcal{F}_n)\le m_n\log(N_{\mu\sigma}N_Q N_q)+\log N_\alpha$. Plugging values of $(N_{\mu\sigma},N_Q,N_q,N_\alpha)$ into this inequality will deliver the claim of the proposition.

Consider $G_{\mu\sigma}$ first. Let $\zeta=\min(\delta/12,1)$. Define $\sigma_h=\underline{\sigma}_n(1+\zeta)^h$, $h\ge 0$. Let H be the smallest integer such that $\sigma_H=\underline{\sigma}_n(1+\zeta)^H\ge\overline{\sigma}_n$. This implies that $H\le\frac{1}{\log(1+\zeta)}\log(\frac{\overline{\sigma}_n}{\underline{\sigma}_n})+1$ and for any $h\ge 1$, $2\frac{\sigma_h-\sigma_{h-1}}{\sigma_{h-1}}\le\frac{\delta}{6}$. Let $N_j=\left\lceil\frac{24\psi(0)}{\delta}\frac{\overline{\mu}_n}{\sigma_{j-1}}\right\rceil$. For $1\le i\le N_j$ and $1\le j\le H$, define

$$E_{ij}^{\mu\sigma}=\left(-\overline{\mu}_n+\frac{2\overline{\mu}_n(i-1)}{N_j}\ ,-\overline{\mu}_n+\frac{2\overline{\mu}_n i}{N_j}\right]\times\left(\sigma_{j-1},\sigma_j\right].$$

For any (μ^1,σ^1) and (μ^2,σ^2) in $E_{ij}^{\mu\sigma}$,

$$\left(2\psi(0)\frac{|\mu^1-\mu^2|}{\sigma^1}+2\frac{|\sigma^1-\sigma^2|}{\min(\sigma^1,\sigma^2)}\right)\le\frac{\delta}{3}.$$

Thus, when $G_{\mu\sigma}$ consists of centers of sets $E_{ij}^{\mu\sigma}$, the first bound in (A.11) can be made no larger than $\delta/3$ with $(\mu_1^2,\sigma_1^2,\ldots,\mu_{m_n}^2,\sigma_{m_n}^2)\in[G_{\mu\sigma}]^{m_n}$. The number of points in $G_{\mu\sigma}$, $N_{\mu\sigma}=\sum_{j=1}^H N_j$, can be bounded as

$$\begin{aligned}N_{\mu\sigma}&\le\sum_{j=1}^{H}\left(\frac{24\psi(0)}{\delta}\frac{\overline{\mu}_n}{\sigma_j}+1\right)=\frac{24\psi(0)}{\delta}\frac{\overline{\mu}_n}{\underline{\sigma}_n}\sum_{j=1}^{H}(1+\zeta)^{-j}+H\\&\le\frac{24\psi(0)}{\delta}\frac{\overline{\mu}_n}{\underline{\sigma}_n}\frac{1}{\zeta}+\frac{1}{\log(1+\zeta)}\log(\frac{\overline{\sigma}_n}{\underline{\sigma}_n})+1\\&=c_0\frac{\overline{\mu}_n}{\underline{\sigma}_n}+c_1\log\frac{\overline{\sigma}_n}{\underline{\sigma}_n}+1,\end{aligned}\tag{A.12}$$

where c_0,c_1 depend on δ, but not on n.

Next, consider G_α. Since only the normalized values of α_j's appear on the r.h.s. of (A.11), G_α can include only points that belong to the (m_n-1)-dimensional simplex.

Thus, we can take G_α from Lemma 1 in Ghosal et al. (1999). It follows immediately from this lemma that the second bound in (A.11) can be made no larger than $\delta/3$ with $(\alpha_1^2, \ldots, \alpha_{m_n}^2) \in G_\alpha$ and N_α satisfying

$$\log N_\alpha \leq m_n \left(1 + \log \frac{1 + \delta K(-\overline{Q}_n d_x)/6}{\delta K(-\overline{Q}_n d_x)/6}\right) \leq m_n \left(c_2 + c_3 \log K(-\overline{Q}_n d_x)\right),$$

where c_2, c_3 depend on δ, but not n.

Define G_Q to be a uniform grid on $[0, \overline{Q}_n]$, with $Q_i = (2i-1)\delta K(-\overline{Q}_n d_x)/(12\overline{K}' d_x)$, $i = 1, \ldots, N_Q$,

$$N_Q = \left\lceil \frac{6\overline{K}' d_x \overline{Q}_n}{\delta K(-\overline{Q}_n d_x)} \right\rceil.$$

Since for any $Q_j^1 \in [0, \overline{Q}_n]$ there exists $Q_i \in G_Q$ such that $|Q_j^1 - Q_i| \leq \delta K(-\overline{Q}_n d_x)/(12\overline{K}' d_x)$, the third bound in (A.11) can be made no larger than $\delta/6$ with $(Q_1^2, \ldots, Q_{m_n}^2) \in [G_Q]^{m_n}$.

Define G_q to be a uniform grid on $[0,1]^{d_x}$,

$$G_q = \left\{ r_l = (2l-1)\frac{\delta K(-\overline{Q}_n d_x)}{24\overline{K}' d_x \overline{Q}_n}, l = 1, \ldots, (N_q)^{1/d_x} \right\}^{d_x},$$

$$N_q = \left\lceil \frac{12\overline{K}' d_x \overline{Q}_n}{\delta K(-\overline{Q}_n d_x)} \right\rceil^{d_x}.$$

Since for any $q_{j,l}^1 \in [0,1]$ there exists r_i such that $|q_{j,l}^1 - r_i| \leq \delta K(-\overline{Q}_n d_x)/(24\overline{K}' d_x \overline{Q}_n)$, the last bound in (A.11) can be made no larger than $\delta/6$ with $(q_1^2, \ldots, q_{m_n}^2) \in [G_q]^{m_n}$.

Obtained bounds for $N_{\mu\sigma}$, N_α, N_Q, and N_q imply

$$\begin{aligned} J(\delta, \mathcal{F}_n) &\leq m_n \log(N_{\mu\sigma} N_Q N_q) + \log N_\alpha \\ &\leq m_n \left(\log \left[b_0 \frac{\overline{\mu}_n}{\underline{\sigma}_n} + b_1 \log \frac{\overline{\sigma}_n}{\underline{\sigma}_n} + 1 \right] + b_2 + b_3 \log \overline{Q}_n + b_4 \log K(-\overline{Q}_n d_x) \right), \end{aligned}$$

where b_0, b_1, b_2, b_3, b_4 do not depend on n. ■

LEMMA A.1. *Inequality (A.11) holds.*

Proof. For notational simplicity let

$$\pi_j^i(x) = \frac{\alpha_j^i K(-Q_j^i ||x - q_j^i||^2)}{\sum_{l=1}^{m_n} \alpha_l^i K(-Q_l^i ||x - q_l^i||^2)}.$$

Then for any given $x \in X$,

$$\begin{aligned}
&\int |f_1(y|x) - f_2(y|x)|dy \\
&= \int \left| \sum_{j=1}^{m_n} \pi_j^1(x) \frac{1}{\sigma_j^1} \psi\left(\frac{y-\mu_j^1}{\sigma_j^1}\right) - \sum_{j=1}^{m_n} \pi_j^2(x) \frac{1}{\sigma_j^2} \psi\left(\frac{y-\mu_j^2}{\sigma_j^2}\right) \right| dy \\
&= \int | \sum_{j=1}^{m_n} \pi_j^1(x)\psi_j^1(y) - \pi_j^2(x)\psi_j^2(y) + \pi_j^1(x)\psi_j^2(y) - \pi_j^1(x)\psi_j^2(y)|dy \\
&\le \int \sum_{j=1}^{m_n} \pi_j^1(x)|\psi_j^1(y) - \psi_j^2(y)|dy + \int \sum_{j=1}^{m_n} |\pi_j^1(x) - \pi_j^2(x)|\psi_j^2(y)dy \\
&= \sum_{j=1}^{m_n} \pi_j^1(x) \int |\psi_j^1(y) - \psi_j^2(y)|dy + \sum_{j=1}^{m_n} |\pi_j^1(x) - \pi_j^2(x)|, \qquad \textbf{(A.13)}
\end{aligned}$$

where $\psi_j^i(y) = (\sigma_j^i)^{-1}\psi((y-\mu_j^i)/\sigma_j^i)$. We will construct bounds for $\int |\psi_j^1(y) - \psi_j^2(y)|dy$ and $\sum_{j=1}^{m_n} |\pi_j^1(x) - \pi_j^2(x)|$ separately. First, let's find an upper bound for

$$\begin{aligned}
&\int |\psi_j^1(y) - \psi_j^2(y)|dy \\
&= \int \left| \frac{1}{\sigma_j^1}\psi\left(\frac{y-\mu_j^1}{\sigma_j^1}\right) - \frac{1}{\sigma_j^2}\psi\left(\frac{y-\mu_j^2}{\sigma_j^2}\right) + \frac{1}{\sigma_j^1}\psi\left(\frac{y-\mu_j^2}{\sigma_j^1}\right) - \frac{1}{\sigma_j^1}\psi\left(\frac{y-\mu_j^2}{\sigma_j^1}\right) \right| dy \\
&\le \int \left| \frac{1}{\sigma_j^1}\psi\left(\frac{y-\mu_j^1}{\sigma_j^1}\right) - \frac{1}{\sigma_j^1}\psi\left(\frac{y-\mu_j^2}{\sigma_j^1}\right) \right| dy + \int \left| \frac{1}{\sigma_j^1}\psi\left(\frac{y-\mu_j^2}{\sigma_j^1}\right) - \frac{1}{\sigma_j^2}\psi\left(\frac{y-\mu_j^2}{\sigma_j^2}\right) \right| dy.
\end{aligned}$$

Note that

$$\begin{aligned}
&\int \left| \frac{1}{\sigma_j^1}\psi\left(\frac{y-\mu_j^1}{\sigma_j^1}\right) - \frac{1}{\sigma_j^1}\psi\left(\frac{y-\mu_j^2}{\sigma_j^1}\right) \right| dy = 2\int_{-\frac{|\mu_j^1-\mu_j^2|}{2}}^{\frac{|\mu_j^1-\mu_j^2|}{2}} \frac{1}{\sigma_j^1}\psi\left(\frac{y}{\sigma_j^1}\right) dy \\
&\le 2\int_{-\frac{|\mu_j^1-\mu_j^2|}{2}}^{\frac{|\mu_j^1-\mu_j^2|}{2}} \frac{1}{\sigma_j^1}\psi(0)dy = 2\psi(0)\frac{|\mu_j^1-\mu_j^2|}{\sigma_j^1}. \qquad \textbf{(A.14)}
\end{aligned}$$

Without loss of generality assume that $\sigma_j^1 > \sigma_j^2$, then

$$\begin{aligned}
&\int \left| \frac{1}{\sigma_j^1}\psi\left(\frac{y-\mu_j^2}{\sigma_j^1}\right) - \frac{1}{\sigma_j^2}\psi\left(\frac{y-\mu_j^2}{\sigma_j^2}\right) \right| dy \\
&= 4\int_0^{+\infty} \max\left(0, \frac{1}{\sigma_j^2}\psi\left(\frac{y}{\sigma_j^2}\right) - \frac{1}{\sigma_j^1}\psi\left(\frac{y}{\sigma_j^1}\right)\right) dy \\
&\le 4\int_0^{+\infty} \max\left(0, \frac{1}{\sigma_j^2}\psi\left(\frac{y}{\sigma_j^1}\right) - \frac{1}{\sigma_j^1}\psi\left(\frac{y}{\sigma_j^1}\right)\right) dy
\end{aligned}$$

$$= 4\int_0^{+\infty} \left(\frac{1}{\sigma_j^2} - \frac{1}{\sigma_j^1}\right) \psi\left(\frac{y}{\sigma_j^1}\right) dy$$
$$= 4\frac{\sigma_j^1 - \sigma_j^2}{\sigma_j^2} \int_0^{+\infty} \frac{1}{\sigma_j^1} \psi\left(\frac{y}{\sigma_j^1}\right) dy \le 4\frac{\sigma_j^1 - \sigma_j^2}{\sigma_j^2} \frac{1}{2} = 2\frac{\sigma_j^1 - \sigma_j^2}{\sigma_j^2}.$$

Combining the two pieces we find that

$$\sum_{j=1}^{m_n} \pi_j^1(x) \int |\psi_j^1(y) - \psi_j^2(y)| dy \le \sum_{j=1}^{m_n} \pi_j^1(x) \left(2\psi(0)\frac{|\mu_j^1 - \mu_j^2|}{\sigma_j^1} + 2\frac{|\sigma_j^1 - \sigma_j^2|}{\min(\sigma_j^2, \sigma_j^1)}\right). \tag{A.15}$$

The next step is to find an upper bound for $\sum_{j=1}^{m_n} |\pi_j^1(x) - \pi_j^2(x)|$. We introduce additional notation $K_j^i(x) = K(-Q_j^i||x - q_j^i||^2)$ and $A_i(x) = \sum_{j=1}^{m_n} \widetilde{\alpha}_j^i K_j^i(x)$. Then for any $x \in X$,

$$\sum_{j=1}^{m_n} |\pi_j^1(x) - \pi_j^2(x)| = \sum_{j=1}^{m_n} \left| \frac{\widetilde{\alpha}_j^1 K_j^1(x)}{\sum_{i=1}^{m_n} \widetilde{\alpha}_i^1 K_i^1(x)} - \frac{\widetilde{\alpha}_j^2 K_j^2(x)}{\sum_{i=1}^{m_n} \widetilde{\alpha}_i^2 K_i^2(x)} \right|$$
$$= \sum_{j=1}^{m_n} \left| \frac{\widetilde{\alpha}_j^1 K_j^1(x)}{A_1(x)} - \frac{\widetilde{\alpha}_j^2 K_j^2(x)}{A_2(x)} \right|$$
$$= \frac{1}{A_1(x)A_2(x)} \sum_{j=1}^{m_n} \left| \widetilde{\alpha}_j^1 K_j^1(x) A_2(x) - \widetilde{\alpha}_j^2 K_j^2(x) A_1(x) \right.$$
$$\left. + \widetilde{\alpha}_j^2 K_j^2(x) A_2(x) - \widetilde{\alpha}_j^2 K_j^2(x) A_2(x) \right|$$
$$\le \frac{\sum_{j=1}^{m_n} |\widetilde{\alpha}_j^1 K_j^1(x) - \widetilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)} + \frac{\sum_{j=1}^{m_n} \widetilde{\alpha}_j^2 K_j^2(x) |A_2(x) - A_1(x)|}{A_1(x)A_2(x)}$$
$$= \frac{\sum_{j=1}^{m_n} |\widetilde{\alpha}_j^1 K_j^1(x) - \widetilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)} + \frac{|A_2(x) - A_1(x)|}{A_1(x)}$$
$$= \frac{\sum_{j=1}^{m_n} |\widetilde{\alpha}_j^1 K_j^1(x) - \widetilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)} + \frac{|\sum_{j=1}^{m_n} \widetilde{\alpha}_j^1 K_j^1(x) - \widetilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)}$$
$$\le 2\frac{\sum_{j=1}^{m_n} |\widetilde{\alpha}_j^1 K_j^1(x) - \widetilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)}$$
$$= 2\frac{\sum_{j=1}^{m_n} |\widetilde{\alpha}_j^1 K_j^1(x) - \widetilde{\alpha}_j^2 K_j^2(x) + \widetilde{\alpha}_j^1 K_j^2(x) - \widetilde{\alpha}_j^1 K_j^2(x)|}{\sum_{j=1}^{m_n} \widetilde{\alpha}_j^1 K_j^1(x)}$$
$$\le 2\left[\frac{\sum_{j=1}^{m_n} \widetilde{\alpha}_j^1 |K_j^1(x) - K_j^2(x)|}{\sum_{j=1}^{m_n} \widetilde{\alpha}_j^1 K_j^1(x)} + \frac{\sum_{j=1}^{m_n} |\widetilde{\alpha}_j^1 - \widetilde{\alpha}_j^2| K_j^2(x)}{\sum_{j=1}^{m_n} \widetilde{\alpha}_j^1 K_j^1(x)}\right]$$
$$\le 2\frac{1}{K(-\overline{Q}_n d_X)} \left[\max_{j=1,\dots,m_n} |K_j^1(x) - K_j^2(x)| + \sum_{j=1}^{m_n} |\widetilde{\alpha}_j^1 - \widetilde{\alpha}_j^2|\right]. \tag{A.16}$$

By Assumption 3.3, the derivative K' is bounded above, let $K' < \overline{K}'$ for some $\overline{K}' < \infty$, then

$$\begin{aligned}
&\left|K\left(-Q_j^1||x-q_j^1||^2\right)-K\left(-Q_j^2||x-q_j^2||^2\right)\right| \\
&\quad\le \left|K\left(-Q_j^1||x-q_j^1||^2\right)-K\left(-Q_j^2||x-q_j^1||^2\right)\right| \\
&\qquad+\left|K\left(-Q_j^2||x-q_j^1||^2\right)-K\left(-Q_j^2||x-q_j^2||^2\right)\right| \\
&\quad\le \overline{K}'\left(||x-q_j^1||^2\right)|Q_j^1-Q_j^2|+\overline{K}'\overline{Q}_n\sum_{l=1}^{d_x}2\left(|q_{j,l}^2-q_{j,l}^1|\right) \\
&\quad\le \overline{K}' d_x\left|Q_j^2-Q_j^1\right|+2\overline{K}' d_x\overline{Q}_n\max_{l=1,\ldots,d_x}\left|q_{j,l}^2-q_{j,l}^1\right|. \qquad \textbf{(A.17)}
\end{aligned}$$

■

Proof (Proposition 4.1).

(i) Let the parameters associated with KM be $\theta^{KM}=\{\alpha_j,Q_j,q_j,\mu_j,\sigma_j\}_{j=1}^m$. For $\delta\in(0,1)$ and a large integer M to be determined later, let the parameters for the KSB mixture be

$$\theta_{1:m\cdot M}^{KSB}=\{\alpha_j\delta,Q_j,q_j,\mu_j,\sigma_j\}_{j=1}^m\times\cdots\times\{\alpha_j\delta,Q_j,q_j,\mu_j,\sigma_j\}_{j=1}^m,$$

so that $\theta_{1:m\cdot M}^{KSB}$ is given by M repetitions of θ^{KM} (except α_j's are multiplied by δ). For brevity let $K_j(x)=K(-Q_j||x-q_j||^2)$. Then

$$\begin{aligned}
p(y|x,\theta_{1:m\cdot M}^{KSB}) &= \sum_{j=1}^{m\cdot M}\alpha_j\delta K_j(x)\prod_{l<j}\{1-\alpha_l\delta K_l(x)\}\phi(y,\mu_j,\sigma_j) \\
&= \sum_{h=1}^{M}\left(\sum_{j=1}^{m}\phi(y,\mu_j,\sigma_j)\alpha_j\delta K_j(x)\prod_{l<j}(1-\alpha_l\delta K_l(x))\right)\left[\prod_{i=1}^{m}(1-\alpha_i\delta K_i(x))\right]^{h-1} \\
&= \left(\sum_{j=1}^{m}\phi(y,\mu_j,\sigma_j)\alpha_j\delta K_j(x)\prod_{l<j}(1-\alpha_l\delta K_l(x))\right)\sum_{h=1}^{M}\left[\prod_{i=1}^{m}(1-\alpha_i\delta K_i(x))\right]^{h-1} \\
&= \frac{\sum_{j=1}^m\phi(y,\mu_j,\sigma_j)\alpha_j\delta K_j(x)\prod_{l<j}(1-\alpha_l\delta K_l(x))}{1-\prod_{i=1}^m(1-\alpha_i\delta K_i(x))}\left(1-\left[\prod_{i=1}^{m}(1-\alpha_i\delta K_i(x))\right]^M\right) \\
&= \frac{\sum_{j=1}^m\phi(y,\mu_j,\sigma_j)\alpha_j\delta K_j(x)\prod_{l<j}(1-\alpha_l\delta K_l(x))}{\sum_{j=1}^m\alpha_j\delta K_j(x)\prod_{l<j}(1-\alpha_l\delta K_l(x))}\left(1-\left[\prod_{i=1}^{m}(1-\alpha_i\delta K_i(x))\right]^M\right) \\
&> \frac{\sum_{j=1}^m\phi(y,\mu_j,\sigma_j)\alpha_j\delta K_j(x)\prod_{l=1}^m(1-\alpha_l\delta K_l(x))}{\sum_{j=1}^m\alpha_j\delta K_j(x)}\left(1-\left[\prod_{i=1}^{m}(1-\alpha_i\delta K_i(x))\right]^M\right) \\
&> \frac{\sum_{j=1}^m\phi(y,\mu_j,\sigma_j)\alpha_j\delta K_j(x)}{\sum_{j=1}^m\alpha_j\delta K_j(x)}\left([1-\delta\max_{j=1,\ldots,m}\alpha_j]^m\right)\left(1-\left[\prod_{i=1}^{m}(1-\alpha_i\delta K_i(x))\right]^M\right)
\end{aligned}$$

$$= p(y|x,\theta^{KM},m)\left([1-\delta\max_{j=1,\ldots,m}\alpha_j]^m\right)\left(1-\left[\prod_{i=1}^{m}(1-\alpha_i\delta K_i(x))\right]^M\right),$$

where the equality in the fifth line follows by induction and we used the fact that $K(\cdot)\le 1$.

Let $\delta < (1-\exp(-\epsilon/(2m)))/\max_{j=1,\ldots,m}\alpha_j$, then $[1-\delta\max_{j=1,\ldots,m}\alpha_j]^m > \exp\{-\epsilon/2\}$. There exists j such that $\alpha_j > 1/m$ and by Assumption 3.3 $K_j(x) > K(-\overline{Q}d_x)$ for any $x\in X$, where $\overline{Q}=\max_{j=1,\ldots,m}Q_j$. Therefore,

$$\prod_{i=1}^{m}(1-\alpha_i\delta K_i(x)) < 1-\frac{\delta K(-\overline{Q}d_x)}{m}.$$

For $M > \frac{\log(1-e^{-\epsilon/2})}{\log\left(1-\frac{\delta K(-\overline{Q}d_x)}{m}\right)}$ the following is true:

$$\left(1-\left[\prod_{i=1}^{m}(1-\alpha_i\delta K_i(x))\right]^M\right) > 1-\left(1-\frac{\delta K(-\overline{Q}d_x)}{m}\right)^M > \exp\{-\epsilon/2\}.$$

Thus, $\log\left(p\left(y\left|x,\theta^{KM},m\right.\right)/p\left(y\left|x,\theta^{KSB}_{1:m\cdot M}\right.\right)\right) < \epsilon$ and the proposition claim (i) follows.

(ii) By part (i) of the proposition, (4.2) holds for $\epsilon/2$ and some $\tilde{\theta}^{KSB}_{1:n}$. The rest of the proof is identical to the proof of Corollary 3.1 (one only needs to replace $\theta_{1:m}$ and $\tilde{\theta}_{1:m}$ with $\theta^{KSB}_{1:n}$ and $\tilde{\theta}^{KSB}_{1:n}$ correspondingly). ■

Proof (Proposition 4.3).
The proof is similar to the proof of Proposition 3.1, and it uses the same notation.
It is shown in Lemma A.2 below that for $p(y|x,\theta^i)\in\mathcal{F}_n$, $i=1,2$, and any $x\in X$,

$$\int\left|p\left(y|x,\theta^1\right)-p\left(y|x,\theta^2\right)\right|dy \le 2\delta \qquad \textbf{(A.18)}$$
$$+2\max_{j=1,\ldots,m_n}\left(\psi(0)\frac{|\mu_j^1-\mu_j^2|}{\sigma_j^1}+\frac{|\sigma_j^1-\sigma_j^2|}{\min(\sigma_j^1,\sigma_j^2)}\right)$$
$$+m_n^2\max_{j=1,\ldots,m_n}\left|\alpha_j^1-\alpha_j^2\right|,$$
$$+m_n^2\overline{K}'d_x\max_{j=1,\ldots,m_n}\left|Q_j^1-Q_j^2\right|$$
$$+m_n^2 2\overline{K}'d_x\overline{Q}_n\max_{j=1,\ldots,m_n}\max_{l=1,\ldots,d_x}\left|q_{j,l}^1-q_{j,l}^2\right|,$$

where $\overline{K}'$ is a finite fixed bound on the derivative of K (Assumption 3.3).

Thus, we can set up $G_{\mu\sigma}$ and $N_{\mu\sigma}$ exactly as in the proof of Proposition 3.1.

Define G_α to be a uniform grid on $[0,1]^{m_n}$,

$$G_\alpha=\left\{\kappa_l=(2l-1)\frac{\delta}{3m_n^2},\, l=1,\ldots,(N_\alpha)^{1/m_n}\right\}^{m_n},$$

$$N_\alpha=\left\lceil\frac{3m_n^2}{2\delta}\right\rceil^{m_n}.$$

Since for any $\alpha_j^1 \in [0,1]$ there exists κ_l such that $|\alpha_j^1 - \kappa_l| \le \delta/(3m_n^2)$, the second bound in (A.18) can be made no larger than $\delta/3$ with $(\alpha_1^2, \ldots, \alpha_{m_n}^2) \in G_\alpha$.

Define G_Q to be a uniform grid on $[0, \overline{Q}_n]$, with $Q_i = (2i-1)\delta/(3\overline{K}' d_x m_n^2)$, $i = 1, \ldots, N_Q$,

$$N_Q = \left\lceil \frac{3\overline{K}' d_x m_n^2}{2\delta} \right\rceil.$$

Since for any $Q_j^1 \in [0, \overline{Q}_n]$ there exists $Q_i \in G_Q$ such that $|Q_j^1 - Q_i| \le \delta/(3\overline{K}' d_x m_n^2)$, the third bound in (A.18) can be made no larger than $\delta/3$ with $(Q_1^2, \ldots, Q_{m_n}^2) \in [G_Q]^{m_n}$.

Define G_q to be a uniform grid on $[0,1]^{d_x}$,

$$G_q = \left\{ r_l = (2l-1)\frac{\delta}{6\overline{K}' d_x \overline{Q}_n m_n^2}, l = 1, \ldots, (N_q)^{1/d_x} \right\}^{d_x}$$

$$N_q = \left\lceil \frac{6\overline{K}' d_x \overline{Q}_n m_n^2}{2\delta} \right\rceil^{d_x}.$$

Since for any $q_{j,l}^1 \in [0,1]$ there exists r_i such that $|q_{j,l}^1 - r_i| \le \delta/(6\overline{K}' d_x \overline{Q}_n m_n^2)$, the last bound in (A.18) can be made no larger than $\delta/3$ with $(q_1^2, \ldots, q_{m_n}^2) \in [G_q]^{m_n}$.

Obtained bounds for $N_{\mu\sigma}$, N_α, N_Q, and N_q imply

$$\begin{aligned} J(4\delta, \mathcal{F}_n) &\le m_n \log(N_{\mu\sigma} N_Q N_q) + \log N_\alpha \\ &\le m_n \left(\log \left[b_0 \frac{\overline{\mu}_n}{\underline{\sigma}_n} + b_1 \log \frac{\overline{\sigma}_n}{\underline{\sigma}_n} + 1 \right] + b_2 + b_3 \log \overline{Q}_n + b_4 \log m_n \right). \end{aligned}$$

■

LEMMA A.2. *Inequality* (A.18) *holds.*

Proof. For $f_1, f_2 \in \mathcal{F}_n$,

$$\begin{aligned} &\int_Y \sum_{j=1}^{\infty} \left| \pi_j^1(x) \phi\left(y; \mu_j^1, \sigma_j^1\right) - \pi_j^2(x) \phi\left(y; \mu_j^2, \sigma_j^2\right) \right| dy \\ &\quad \le \int_Y \sum_{j=1}^{m_n} \pi_j^1(x) \left| \phi\left(y; \mu_j^1, \sigma_j^1\right) - \phi\left(y; \mu_j^2, \sigma_j^2\right) \right| dy \\ &\qquad + \int_Y \sum_{j=1}^{m_n} \left| \pi_j^1(x) - \pi_j^2(x) \right| \phi\left(y; \mu_j^2, \sigma_j^2\right) dy \\ &\qquad + \sum_{j=m_n+1}^{\infty} |\pi_j^1(x) - \pi_j^2(x)| \\ &\quad \le \sum_{j=1}^{m_n} \pi_j^1(x) \int_Y \left| \phi\left(y; \mu_j^1, \sigma_j^1\right) - \phi\left(y; \mu_j^2, \sigma_j^2\right) \right| dy \end{aligned}$$

$$
\begin{aligned}
&+\sum_{j=1}^{m_n}\left|\left|\pi_j^1-\pi_j^2\right|\right|_1+\sup_{x\in X}\sum_{j=m_n+1}^{\infty}\left|\pi_j^1(x)\right|+\left|\pi_j^2(x)\right| \\
&\le \sum_{j=1}^{m_n}\pi_j^1(x)\int_Y\left|\phi\left(y;\mu_j^1,\sigma_j^1\right)-\phi\left(y;\mu_j^2,\sigma_j^2\right)\right|dy \\
&+\sum_{j=1}^{m_n}\left|\left|\pi_j^1-\pi_j^2\right|\right|_1+2\delta,
\end{aligned}
$$

where the last inequality is true by construction of $\mathcal{F}_n$ as $\sup_{x\in X}\sum_{j=m_n+1}^{\infty}|\pi_j^i(x)|\le\delta$ for $i=1,2$. As shown in Lemma A.1, the first expression on the r.h.s. of the last inequality is bounded by the first bound in (A.18),

$$
\begin{aligned}
|\pi_j^1(x)-\pi_j^2(x)| &= \left|\alpha_j^1K_j^1(x)\prod_{i<j}\left(1-\alpha_i^1K_i^1(x)\right)-\alpha_j^2K_j^2(x)\prod_{i<j}\left(1-\alpha_i^2K_i^2(x)\right)\right| \\
&\le \left|\alpha_j^1K_j^1(x)-\alpha_j^2K_j^2(x)\right|\prod_{i<j}\left(1-\alpha_i^1K_i^1(x)\right) \\
&\quad+\alpha_j^2K_j^2(x)\left|\prod_{i<j}\left(1-\alpha_i^1K_i^1(x)\right)-\prod_{i<j}\left(1-\alpha_i^2K_i^2(x)\right)\right| \\
&\le \left|\alpha_j^1K_j^1(x)-\alpha_j^2K_j^2(x)\right|+\left|\prod_{i<j}\left(1-\alpha_i^1K_i^1(x)\right)-\prod_{i<j}\left(1-\alpha_i^2K_i^2(x)\right)\right| \\
&\le \sum_{i=1}^{j}\left|\alpha_i^1K_i^1(x)-\alpha_i^2K_i^2(x)\right| \\
&\le \sum_{i=1}^{j}\left|\alpha_i^1-\alpha_i^2\right|+\left|K_i^1(x)-K_i^2(x)\right|.
\end{aligned}
$$

Using the bound on $|K_i^1(x)-K_i^2(x)|$ from (A.17) in Lemma A.1 and noting that $\sum_{j=1}^{m_n}j\le m_n^2$ complete the proof. ■

Proof (Lemma 4.1).

First, we note that if the prior distribution of α_j first-order stochastically dominates $Beta(1,\gamma)$ and if the prior distribution of $K_j=K(-Q_jd_x)$ first-order stochastically dominates $Beta(\gamma+1,1)$, then $\alpha_j\cdot K_j$ first-order stochastically dominates $Beta(1,\gamma+1)$. This is true by Theorem 1 of Jambunathan (1954), which states that if $a_1\sim Beta(1,\gamma)$ and if $a_2\sim Beta(\gamma+1,1)$, then $a_1\cdot a_2\sim Beta(1,\gamma+1)$.

Second, another auxiliary result that will be used in the proof of the lemma is that if $c\sim Gamma(m,1/\gamma)$, then $\Pr(c<x)<e^{-0.5m\log m}$ for m large enough. For positive integer m,

$$
\begin{aligned}
\Pr(c<x) &= \frac{\int_0^x\gamma^m t^{m-1}e^{-\gamma t}dt}{(m-1)!}=\frac{\int_0^{\gamma x}t^{m-1}e^{-t}dt}{(m-1)!}<(\gamma x)^m/m! \\
&= \frac{(\gamma x)^m}{\exp\{m\log m-m+O(\log(m))\}} \quad \text{(by Sterling formula)}
\end{aligned}
$$

$$
\begin{aligned}
&= \exp\{-m\log m + m + m\log(\gamma x) - O(\log(m))\} \\
&= \exp(-0.5m\log m)\frac{\exp(m\log(\gamma x)+m+O(\log(m)))}{\exp(0.5m\log m)} \\
&< \exp(-0.5m\log m)
\end{aligned}
$$

when m is sufficiently large.

Using these two auxiliary results, note that if α_j and K_j first-order stochastically dominate $Beta(1,\gamma)$ and $Beta(\gamma+1,1)$, then for $a_1 \overset{i.i.d.}{\sim} Beta(1,\gamma)$, $a_2 \overset{i.i.d.}{\sim} Beta(\gamma+1,1)$, $b_j \overset{i.i.d.}{\sim} Beta(1,\gamma+1)$, and $c \sim Gamma(m_n, 1/(\gamma+1))$,

$$
\begin{aligned}
&\Pi\left(\prod_{j=1}^{m_n}(1-\alpha_j K_j) > \delta\right) \\
&= \int \Pi\left(\alpha_1 K_1 < 1 - \frac{\delta}{\prod_{j\neq 1}(1-\alpha_j K_j)} | \alpha_j, K_j, j\neq 1\right) d\Pi(\alpha_j, K_j, j\neq 1) \\
&\leq \int \Pi\left(a_1 a_2 < 1 - \frac{\delta}{\prod_{j\neq 1}(1-\alpha_j K_j)} | \alpha_j, K_j, j\neq 1\right) d\Pi(\alpha_j, K_j, j\neq 1) \\
&\leq \int \Pi\left(b_1 < 1 - \frac{\delta}{\prod_{j\neq 1}(1-\alpha_j K_j)} | \alpha_j, K_j, j\neq 1\right) d\Pi(\alpha_j, K_j, j\neq 1) \\
&= \Pi\left((1-b_1)\prod_{j\neq 1}(1-\alpha_j K_j) > \delta\right) \quad \text{(repeat for } b_2,\ldots,b_{m_n}\text{)} \\
&\leq \Pi\left(\prod_{j=1}^{m_n}(1-b_j) > \delta\right) = \Pi\left(\sum_{j=1}^{m_n} -\log(1-b_j) < -\log(\delta)\right) \\
&= \Pi(c < -\log(\delta)) < e^{-0.5m_n\log m_n}.
\end{aligned}
$$

■

LEMMA A.3. *Let $A_1,\ldots,A_m$ be a partition of an interval on R such that $\lambda(A_j)\leq h$ and $\mu_j \in A_j$. Assume $C_\delta(y) = [y-\delta, y+\delta] \subset \cup A_j$ is an interval with center y and length δ. Then*

$$
\sum_{j=1}^{m}\lambda(A_j\cap C_\delta(y))\sigma^{-1}\psi((y-\mu_j)/\sigma) \geq 1 - \frac{4h\psi(0)}{\sigma} - 2\int_{\delta/\sigma}^{\infty}\psi(\mu)d\mu.
$$

If $C_\delta(y) = [y-\delta, y]$ or $C_\delta(y) = [y, y+\delta]$, the lower bound in the above expression should be divided by 2.

Proof. Let $J = \{j : A_j \cap C_\delta(y) \subset [y-\delta, y]\}$. For any $j\in J$ and $\mu \in A_j\cap C_\delta(y)$, $\mu - h \leq \mu_j$ as $\lambda(A_j) < h$ and $\mu_j \in A_j$, which implies $\phi(y,\mu_j,\sigma) \geq \phi(y,\mu-h,\sigma)$. Therefore,

$$
\sum_{j\in J}\lambda(A_j\cap C_\delta(y))\phi(y,\mu_j,\sigma) \geq \int_{\cup_{j\in J}[A_j\cap C_\delta(y)]}\phi(y,\mu-h,\sigma)d\mu. \tag{A.19}
$$

Note next that

$$\begin{aligned}
&\int_{\cup_{j\in J}[A_j\cap C_\delta(y)]}\phi(y,\mu-h,\sigma)d\mu\\
&\quad\geq\int_{y-\delta}^{y-h}\phi(y,\mu-h,\sigma)d\mu=\int_{y-\delta-h}^{y-2h}\phi(y,\mu,\sigma)d\mu\\
&\quad\geq\int_{y-\delta}^{y}\phi(y,\mu,\sigma)d\mu-\int_{y-2h}^{y}\phi(y,\mu,\sigma)d\mu\\
&\quad\geq\int_{y-\delta}^{y}\phi(y,\mu,\sigma)d\mu-\frac{2h\psi(0)}{\sigma}.
\end{aligned}$$

By symmetry the same results can be obtained for $J=\{j: A_j\cap C_\delta(y)\subset[y,y+\delta]\}$. Thus

$$\sum_{j=1}^{m}\lambda(A_j\cap C_\delta(y))\phi(y,\mu_j,\sigma)\geq\int_{y-\delta}^{y+\delta}\phi(y,\mu,\sigma)d\mu-2\frac{2h\psi(0)}{\sigma}.$$

A change of variables delivers the claim of the lemma. ■

THEOREM A.1. *The theorem summarizes modifications of the theoretical results from Sections 3–4 to models with covariate dependent locations $\beta_j'z(x)$ introduced in Section 5.*

Suppose Assumption 5.1 holds. Replace the definition of parameter vector

$$\theta=\{Q_j,\mu_j,\sigma_j,q_j,\alpha_j\}_{j=1}^{\infty}\in\Theta=(R_+\times Y\times R_+\times X\times(0,1))^{\infty}\quad\text{by}$$

$$\theta=\{Q_j,\beta_j,\sigma_j,q_j,\alpha_j\}_{j=1}^{\infty}\in\Theta=\left(R_+\times R^{d_z}\times R_+\times X\times(0,1)\right)^{\infty}$$

(make the same change in $\theta_{1:m}$, θ^{KM}, θ^{KSB}, and $\theta_{1:n}^{KSB}$).

(i) Theorem 3.1 and Corollary 3.1 hold for the model with locations $\beta_j'z(x)$. Thus, the weak posterior consistency for kernel mixtures (Theorem 3.2) also holds.

(ii) Propositions 4.1 and 4.2 and, thus, the weak posterior consistency for kernel stick-breaking mixtures (Theorem 4.1) hold for the model with locations $\beta_j'z(x)$.

(iii) For the models with locations $\beta_j'z(x)$, replace inequality $|\mu_j|\leq\overline{\mu}_n$ in the sieve definitions by $|\beta_{j,l}|\leq\overline{\beta}_n$. Then, the entropy bounds in Propositions 3.1 and 4.3 are changed as follows: The term

$$\log\left[b_0\frac{\overline{\mu}_n}{\underline{\sigma}_n}+b_1\log\frac{\overline{\sigma}_n}{\underline{\sigma}_n}+1\right]$$

in the bounds is replaced by

$$d_z\log\left[b_0\frac{\overline{\beta}_n}{\underline{\sigma}_n}+b_1\log\frac{\overline{\sigma}_n}{\underline{\sigma}_n}+1\right].$$

(iv) Replace $\Pi(|\mu_j|>\overline{\mu}_n)$ in (3.4) *and* (4.4) *with $\sum_{l=1}^{d_z}\Pi(|\beta_{j,l}|>\overline{\beta}_n)$ and make the changes in the sieve definitions and the entropy bounds* (3.6) *and* (4.5) *as described in part (iii) above. Then, strong posterior consistency (Theorems 3.3 and 4.2) holds.*

Proof. (i) Theorem 3.1 is obtained by setting $\beta_{j,l} = 0$ for all j and all $l = 2, \ldots, d_z$.

The proof of Corollary 3.1 can be modified in the following way. Let $|\beta^n_{j,l}| \le \overline{\beta}_l$ and let $\overline{\beta} = \max_{z\in Z} ||z||_\infty \sum_{l=1}^{d_z} \overline{\beta}_l$. Note that $\overline{\beta} < \infty$ since $\max_{z\in Z} ||z||_\infty = 1$ by Assumption 5.1. Then equation (3.3) is true by setting $\underline{\mu} = -\overline{\beta}$ and $\overline{\mu} = \overline{\beta}$ and hence Corollary 3.1 holds.

(ii) Propositions 4.1 and 4.2 remain true without any changes.

(iii) Equivalents of Propositions 3.1 and 4.3 can be proved as follows. The bounds in (3.4) and (4.3) can be adapted to the current setup by replacing $|\mu^1_j - \mu^2_j|$ with $\sum_{l=1}^{d_z} |\beta^1_{j,l} - \beta^2_{j,l}|$. Thus, we only need to replace $G_{\mu\sigma}$ and $N_{\mu\sigma}$ with suitable $G_{\beta\sigma}$ and $N_{\beta\sigma}$. Using the notation from the definition of $E^{\mu\sigma}_{ij}$ in the proof of Proposition 3.1, we define

$$E^{\beta\sigma}_{i_1,\ldots,i_{d_z},j} = \prod_{l=1}^{d_z} \left(-\overline{\beta}_n + \frac{2\overline{\beta}_n(i_l - 1)}{N_j} \, , -\overline{\beta}_n + \frac{2\overline{\beta}_n i_l}{N_j} \right] \times \left(\sigma_{j-1}, \sigma_j\right],$$

where $N_j = \left\lceil \frac{24 d_z \psi(0)}{\delta} \frac{\overline{\beta}_n}{\sigma_{j-1}} \right\rceil$, $1 \le i_k \le N_j$, and $1 \le j \le H$. If $(\beta_1, \sigma_1), (\beta_2, \sigma_2) \in E^{\beta\sigma}_{i_1,\ldots,i_{d_z},j}$, then

$$2\psi(0) \frac{\sum_{l=1}^{d_z} |\beta_{1,l} - \beta_{2,l}|}{\sigma_1} + 2 \frac{|\sigma_1 - \sigma_2|}{\min(\sigma_1, \sigma_2)} \le \frac{\delta}{3}. \tag{A.20}$$

Thus, when $G_{\beta\sigma}$ consists of centers of sets $E^{\beta\sigma}_{i_1,\ldots,i_{d_z},j}$, an analog of the first bound in (A.11) can be made no larger than $\delta/3$ with $(\beta^2_1, \sigma^2_1, \ldots, \beta^2_{m_n}, \sigma^2_{m_n}) \in [G_{\beta\sigma}]^{m_n}$. The number of points in $G_{\beta\sigma}$, $N_{\beta\sigma} = \sum_{j=1}^H N_j^{d_z} \le \left(\sum_{j=1}^H N_j\right)^{d_z}$. By the same arguments as in deriving (A.12),

$$N_{\beta\sigma} \le \left(b_0 \frac{\overline{\beta}_n}{\underline{\sigma}_n} + b_1 \log \frac{\overline{\sigma}_n}{\underline{\sigma}_n} + 1 \right)^{d_z},$$

where b_0, b_1 depend on δ, but not on n. Thus, the claimed entropy bounds are obtained.

(iv) This part is implied by parts (i)–(iii) and the general strong posterior consistency result (Theorem 2.2).

■

APPENDIX B. Computation

B.1. MCMC Algorithm. This section describes an MCMC algorithm for a KSB model given in (6.1). Let us denote the data by $Y = \{y_i\}_{i=1}^N$ and $X = \{x_i\}_{i=1}^N$ and parameters by $\alpha = \{\alpha_j\}_{j=1}^\infty$, $Q = \{Q_j\}_{j=1}^\infty$, $q = \{q_j\}_{j=1}^\infty$, and $\theta = (\alpha, Q, q, \{\beta_j, \sigma_j^2\}_{j=1}^\infty)$. The prior is $(\beta_{j,0}, \beta_{j,1})' \sim N(\mu_\beta, H_\beta^{-1})$, $\sigma_j^2 \sim InvGamma(v, b_\sigma)$, $\alpha_j \sim Beta(a, b)$, $q_j \sim U(0,1)$, $Q_j \sim Exponential(\tau)$ i.i.d. for each j.

We introduce latent variables $Z = \{z_i\}_{i=1}^N$ and $U = \{u_i\}_{i=1}^N$, such that $p(y_i | x_i, \theta, u_i, z_i = j) = \phi\left(y_i; \beta_{j,0} + \beta_{j,1} x_i, \sigma_j^2\right)$ and $p(z_i = j | x_i, \theta) = \pi_j(x; \alpha,$

Q,q). As in slice sampling algorithms (Neal, 2003; Walker, 2007), the latent variables U are such that $p(u_i|z_i,x_i,\theta) = 1(u_i < \pi_{z_i}(x_i;\alpha,Q,q))/\pi_{z_i}(x_i;\alpha,Q,q)$. Then the posterior density of unobservables is

$$p(\theta,Z,U|Y,X) \propto \prod_{i=1}^{N} \left[p(y_i|x_i,u_i,\theta,z_i)p(u_i|z_i,x_i,\theta)p(z_i|x_i,\theta)\right]\cdot\Pi(\theta)$$
$$= \prod_{i=1}^{N} \phi\left(y_i;\beta_{z_i,0}+\beta_{z_i,1}x_i,\sigma_{z_i}^2\right)1\left(u_i < \pi_{z_i}(x_i;\alpha,\psi,q)\right)\cdot\Pi(\theta), \tag{B.1}$$

where $\Pi(\theta)$ is the prior density of the parameters.

The blocks of our Metropolis-within-Gibbs MCMC algorithm are as follows.

1. Blocks for $\{\beta_j,\sigma_j^2\}_{j=1}^{\infty}$. Following the retrospective sampling ideas from Papaspiliopoulos and Roberts (2008), we simulate only $\{\beta_j,\sigma_j^2\}_{j=1}^{M}$, where $M = \max\{Z\}$ is the maximum allocation number within any given iteration. The conditional posterior for $\{\beta_j,\sigma_j^2\}_{j=M+1}^{\infty}$ is independent of the rest of the variables and equal to the prior distribution. Thus, any finite part of $\{\beta_j,\sigma_j^2\}_{j=M+1}^{\infty}$ can be simulated in subsequent MCMC iterations from the prior if necessary (when M becomes larger, see Step 6 of the algorithm below).

   For $j \le M$, let $T_j = \sum_{i=1}^{N} 1\{z_i = j\}$. From posterior density (B.1) we find that

   $$p(\beta_j|Y,Z,X,U,\theta\backslash\beta_j) \propto \phi(\beta_j;\mu_\beta,H_\beta^{-1}) \prod_{i:z_i=j} \phi(y_i;\beta_{j,0}+\beta_{j,1}x_i,\sigma_j^2).$$

   This leads to the conditional posterior distribution for β_j,

   $$\beta_j|(Y,Z,X,U,\theta\backslash\beta_j) \sim N\left(\overline{\mu}_\beta,\overline{H}_\beta^{-1}\right), \quad \text{where}$$
   $$\overline{H}_\beta = H_\beta + \sigma_j^{-2}\sum_{i:z_i=j}\begin{pmatrix}1 & x_i'\\ x_i & x_i x_i'\end{pmatrix},$$
   $$\overline{\mu}_\beta = \overline{H}_\beta^{-1}\left(H_\beta\mu_\beta + \sigma_j^{-2}\sum_{i:z_i=j}\begin{pmatrix}x_i\\ x_i y_i\end{pmatrix}\right).$$

   Similarly,

   $$p(\sigma_j^2|Y,Z,X,U,\theta\backslash\sigma_j) \propto InvGamma\left(\sigma_j^2;\nu,b_\sigma\right)$$
   $$\prod_{i:z_i=j}\phi\left(y_i;\beta_{j,0}+\beta_{j,1}x_i,\sigma_j^2\right).$$

   Then the conditional posterior distribution of σ_j is

   $$\sigma_j^2|(Y,Z,X,U,\theta\backslash\sigma_j)$$
   $$\sim InvGamma\left(\nu+T_j/2,\left(b_\sigma^{-1}+0.5\sum_{i:z_i=j}(y_i-\beta_{j,0}-\beta_{j,1}x_i)^2\right)^{-1}\right).$$

2. Block $\{\alpha_j, u_i, i : z_i \geq j\}$ is updated separately for each $j = 1, \ldots, M$ by a Metropolis-within-Gibbs step with the following transition probability at MCMC iteration $m+1$:

   (a) Simulate proposal α_j^* from a Markov transition density

   $$R[\alpha_j^* | \alpha_j^m; Z^m, \theta^m \setminus \alpha_j^m],$$

   (which is parameterized by $(Z^m, \theta^m \setminus \alpha_j^m)$);

   (b) conditional on α_j^*, simulate u_i^* for i such that $z_i \geq j$ from the uniform density

   $$\frac{1\{u_i^* < \pi_{z_i}(x_i; \alpha_j^*, \alpha^m \setminus \alpha_j^m, q^m, Q^m)\}}{\pi_{z_i}(x_i; \alpha_j^*, \alpha^m \setminus \alpha_j^m, q^m, Q^m)}.$$

   Since u_i's are simulated from the conditional proposal equal to the conditional target, they do not affect the Metropolis-Hastings acceptance probability

   $$\min\left\{1, \frac{p(\alpha_j^* | X, Z^m, Y, \theta^m \setminus \alpha_j^*)/R[\alpha_j^* | \alpha_j^m; Z^m, \theta^m \setminus \alpha_j^m]}{p(\alpha_j^m | X, Z^m, Y, \theta^m \setminus \alpha_j^m)/R[\alpha_j^m | \alpha_j^*; Z^m, \theta^m \setminus \alpha_j^*]}\right\}, \tag{B.2}$$

   where

   $$p(\alpha_j^* | X, Z^m, Y, \theta^m \setminus \alpha_j^*) \propto \Pi(\alpha_j^*) \prod_{i: z_i \geq j}^{N} \pi_{z_i}(x_i; \alpha_j^*, \alpha^m \setminus \alpha_j^m, Q, q). \tag{B.3}$$

   We use the transition density for α_j,

   $$\alpha_j^* | \alpha_j^m; U^m, Z^m, \theta^m \setminus \alpha_j^m \sim Beta(a + T_j, b + b(\alpha_j^m)), \quad \text{where}$$

   $$b(\alpha_j^m) = \frac{\sum_{i: z_i > j} \log\left(1 - \alpha_j^m K(-Q_j ||x_i - q_j||^2)\right)}{\log(1 - \alpha_j^m)}.$$

   This transition density is constructed so that the kernels of the conditional posterior density in (B.3) and the proposal beta density are equal at α_j^m.

   The draws of u_i's obtained in this step are not used in the algorithm (they are resimulated in step 5 below). Thus, their role is only in the justification of a convenient update for α_j's, and they are not simulated in the algorithm implementation.

3. Updating block $\{q_j, u_i, i : z_i \geq j\}$ is analogous to updating $\{\alpha_j, u_i, i : z_i \geq j\}$, where instead of transition density R we use a Metropolis random walk density

   $$q_j^* | q_j^m; U^m, Z^m, \theta^m \setminus q_j^m \sim N(q_j^m, (2Q_j^m T_j + 4)^{-1}).$$

4. Updating block $\{Q_j, u_i, i : z_i \geq j\}$ is analogous to updating $\{\alpha_j, u_i, i : z_i \geq j\}$, where instead of transition density R we use $Q_j^* | \alpha_j^m; U^m, Z^m, \theta^m \setminus Q_j^m \sim N(Q_j^m, 0.5^2)$.

5. Updating U. For all $i = 1, \ldots, N$, $p(u_i | X, Y, Z, \theta) \propto 1(u_i < \pi_{z_i}(x_i))$. Therefore, simulate $u_i \sim U(0, \pi_{z_i}(x_i))$ for all i.

6. Updating $\{\alpha_j, q_j, \beta_j, \sigma_j^2, Q_j\}_{j=M+1}^{M^*}$, where M^* is such that for all $i = 1, \ldots, N$,

$$\sum_{j=1}^{M^*} \pi_j(x_i) > 1 - u_i. \tag{B.4}$$

As described below in Step 7 of the algorithm, this condition on M^* guarantees that draws of α_j, q_j, β_j, σ_j^2, and Q_j necessary for updating Z are available.

For all $j > M$ the density of $\{\alpha_j, q_j, \mu_j, \sigma_j^2, Q_j\}$ conditional on (X, Y, Z, U) and other parameters is equal to the prior density. Hence, we simulate $\{\alpha_j, q_j, \mu_j, \sigma_j^2, Q_j\}$ for $j = M+1, \ldots, M^*$ from the prior.

7. Updating Z. Note that

$$p(z_i = j|X, Y, U, \theta) \propto 1(u_i < \pi_j(x_i))\phi(y_i; \beta_{j,0} + \beta_{j,1}x_i, \sigma_j^2).$$

By construction, $\pi_j(x_i) < u_i$ for all $j > M^*$ (see equation (B.4)). Then $p(z_i = j|X, Y, U, \theta) = 0$ for all $j > M^*$ and hence updating z_i is a simple draw from a multinomial distribution for each i with

$$p\big(z_i = j|X, Y, U, \theta, j \le M^*\big) = \frac{1(u_i < \pi_j(x_i))\phi(y_i; \beta_{j,0} + \beta_{j,1}x_i, \sigma_j^2)}{\sum_{l=1}^{M^*} 1(u_i < \pi_l(x_i))\phi(y_i; \beta_{l,0} + \beta_{l,1}x_i, \sigma_l^2)}.$$

B.2. Posterior of Conditional Density. In the simulation exercise of Section 6, the estimator of conditional density at given (y, x) is the posterior mean of $p(y|x, \theta)$, $p(y|x, Y, X) = \int p(y|x, \theta)d\Pi(\theta|Y, X)$, which is also equal to the predictive density of y given x. To approximate $p(y|x, Y, X)$ and the 0.5% and 99.5% quantiles of the posterior for $p(y|x, \theta)$ also reported in Section 6, we can use MCMC draws $\{\theta^{(l)}, Z^{(l)}\}_{l=1}^{L}$. Specifically, for $J_l \ge \max Z^{(l)}$,

$$p(y|x, Y, X) \approx \frac{1}{L}\sum_{l=1}^{L}\left[\sum_{j=1}^{J_l} \pi_j(x; \theta^{(l)})\phi\left(\frac{y - \beta_{j,0}^{(l)} - \beta_{j,1}^{(l)\prime}x}{\sigma_j^{(l)}}\right)\right.$$
$$\left. + \left(1 - \sum_{j=1}^{J_l} \pi_j(x; \theta^{(l)})\right)\int \phi\left(\frac{y - \beta_0 - \beta_1{}'x}{\sigma}\right)d\Pi(\beta_0, \beta_1, \sigma)\right], \tag{B.5}$$

where the integral in the second line of the equation can be evaluated numerically. Note that when $J_l = \infty$, the r.h.s. of (B.5) is just a sample average for the population mean, $p(y|x, Y, X)$. The use of finite J_l not only makes the computation feasible but also reduces the variance of the approximation (see Sec. 4.4.1 in Geweke, 2005).

To approximate $p(y|x, \theta^{(l)}), l = 1, \ldots, L$, which is necessary for obtaining the posterior quantiles of $p(y|x, \theta)$, we use the expression in the square brackets of (B.5). The quality of the resulting quantile approximations improves as J_l increases; we choose J_l so that

$$1 - \sum_{j=1}^{J_l} \pi_j(x; \theta^{(l)}) < 10^{-4}, \forall l, x.$$

The results in Section 6 are obtained on a grid for pairs (y, x), where $y \in \{-1.5, -1.49, \ldots, 1.5\}$ and $x \in \{0.25, 05, 0.75\}$ unless specified otherwise.

APPENDIX C. Prior sensitivity analysis

The DGP for prior sensitivity analysis is given in (6.3). The results are presented for the sample size of $N = 500$. The prior defined in (6.2) is used as the benchmark for the analysis. We consider the following deviations from the benchmark prior:

(i) Benchmark prior;

(ii) Benchmark prior, but $H_\mu^{-1} = \text{diag}(100 \cdot \text{var}(Y), 1)$;

(iii) Benchmark prior, but $b_\sigma = 0.02(\text{var}(Y))^{-1}$;

(iv) Benchmark prior, but $\nu = 20$;

(v) Benchmark prior, but $b = 0.05$ and $\gamma = 1.05$;

(vi) Benchmark prior, but $b = 10$ and $\gamma = 11$.

Figures 3 and 4 show conditional density estimates with the rows representing $x = 0.25, 0.5, 0.75$ and the columns representing different priors. Results were obtained by running the MCMC algorithm from Appendix B for 400,000 iterations with a burn-in of 100,000 and using only every 20th iteration to construct the plots.

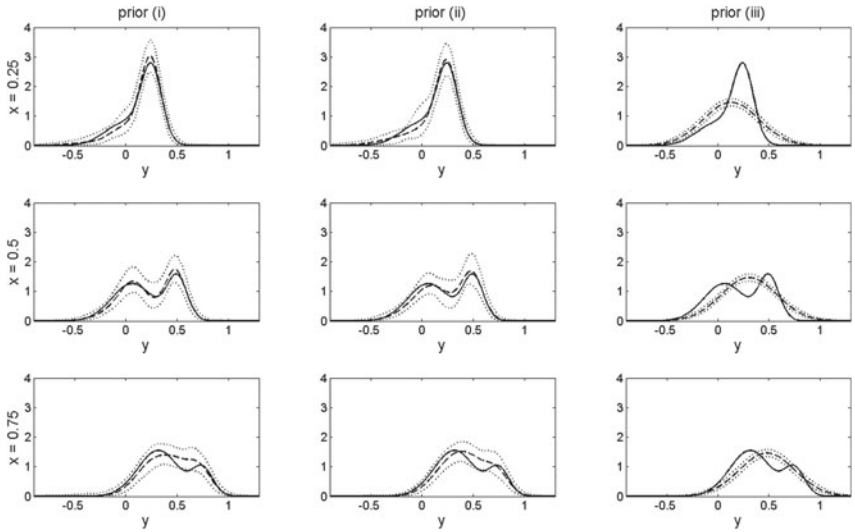

FIGURE 3. Estimated conditional response densities for different covariate values and different prior specifications. The solid lines are the true values, the dashed lines are the posterior means, and the dotted lines are pointwise 99% equal-tailed credible intervals.

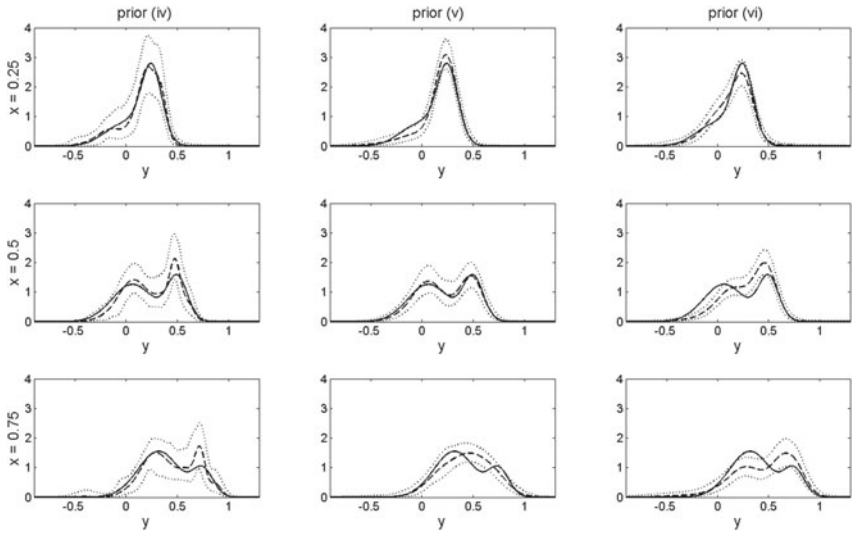

FIGURE 4. Estimated conditional response densities for different covariate values and different prior specifications. The solid lines are the true values, the dashed lines are the posterior means, and the dotted lines are pointwise 99% equal-tailed credible intervals.

Conditional density estimates for priors (i) and (ii) are very similar; the estimates for priors (iv) and (v) are comparable as well. Priors (vi) and, especially, (iii) lead to unreasonable results. Results for priors (vi) demonstrate that high values of γ limit the dependence on covariates and lead to oversmoothing. Prior (iii) performs poorly, as the choice of b_σ implies very high variance of responses within mixture components.