POSTERIOR CONSISTENCY IN CONDITIONAL DENSITY ESTIMATION
BY COVARIATE DEPENDENT MIXTURES *

BY ANDRIY NORETS AND JUSTINAS PELENIS†

Princeton University and Institute for Advanced Studies, Vienna

This paper considers Bayesian non-parametric 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 non-parametric 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, non-parametric estimation of conditional densities is useful for forecasting; see the literature survey in Fan (2005).

The use of Bayesian non-parametric models is less common, especially in methodological econometric research. However, Bayesian non-parametric methods have a number of attrac-
tive properties. First, they never result in logical 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 non-parametric methods have been proved to possess excellent frequentist properties in several important problems. For example, Rousseau (2010) and Kruijer et al. (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 non-parametric conditional density estimation includes papers by Geweke and Keane (2007), Villani et al. (2009), Li et al. (2010), Tran et al. (2012), and Villani et al. (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 non-parametric 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, 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 non-parametric 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 non-parametric 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 non-parametric 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 data-generating processes (DGP). 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 non-parametric literature. One approach to non-parametric 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 et al. (2009), and Tokdar et al. (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 non-parametric 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 et al. (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 et al. (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.


Muller et al. (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 et al. (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 et al. (1991), Jordan and Xu (1995), Peng et al. (1996), Wood et al. (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 non-parametric 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 suggest that these Kullback-Leibler approximation results imply posterior consistency in a weak topology norm. Pati et al. (2013) specify dependent Dirichlet processes that are similar to the 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 (2010) to finite mixture specifications in which mixing probabilities are proportional to a general kernel multiplied by a constant. We will call such mixture specifications kernel mixtures (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 $Y \times X$, $Y \subset \mathbb{R}$ and $X \subset \mathbb{R}^d$. Let $\mathcal{F} = \{ f : Y \times X \to [0, \infty), \int_Y f(y|x) dy = 1 \}$ 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 $\nu$ by $f_{x0}(x)$ and the data-generating conditional density of interest by $f_0 \in \mathcal{F}$. The joint probability measure implied by $f_0$ and $f_{x0}(x)$ 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 non-parametric 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 et al. (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 $f_{x0}(x)$. 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,

$$\int |f_1 f_{x0} - f_2 f_{x0}| = \int |f_1(y|x) f_{x0}(x) - f_2(y|x) f_{x0}(x)| dy d\nu(x).$$

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

$$U = \{ f \in \mathcal{F} : \left| \int g_i f f_{x0} - \int g_i f_0 f_{x0} \right| < \epsilon, i = 1, 2, \ldots, k \},$$

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

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

$$K_\epsilon(f_0) = \left\{ f \in \mathcal{F} : \int \log \frac{f_0(y|x)}{f(y|x)} dF_0(y, x) = \int \log \frac{f_0(y|x) f_{x0}(x)}{f(y|x) f_{x0}(x)} dF_0(y, x) < \epsilon \right\}.$$

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 \( \Pi \) on \( F \) 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 \( \Pi(K, f_0) > 0 \) for any \( \epsilon > 0 \), then the corresponding posterior is weakly consistent at \( f_0 \): for any weak neighborhood \( U \) of \( f_0 \),

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

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 \( A \subset F \). 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 \( F \) with the property \( A \subset \bigcup_{i=1}^k \{ f : \int |f - f_i|f_0^x < \delta \} \).

**Theorem 2.2.** Suppose \( \Pi(K, f_0) > 0 \) 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 \( F_n \subset F \) such that for all \( n \) large enough:

1. \( \Pi(F_n^\infty) < c_1 \exp\{-c_2 n\} \) and
2. \( J(\delta, F_n) < \beta n \),

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

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 following model for a conditional density,

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

where \( \phi(y, \mu, \sigma) \) is a fixed symmetric density with location \( \mu \) and scale \( \sigma \) 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 \(\theta = \{Q_j, \mu_j, \sigma_j, q_j, \alpha_j\}_{j=1}^{\infty} \in \Theta = (R_+ \times Y \times R_+ \times X \times (0,1))^{\infty}\), 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.**

1. \(X = [0, 1]^d\) (the arguments would go through for a bounded \(X).\)
2. \(f_0(y|x)\) is continuous in \((y, x)\) a.s. \(F_0.\)
3. There exists \(r > 0\) such that
   \[
   \int \log \frac{f_0(y|x)}{\inf_{\|z-y\| \leq r, \|t-x\| \leq r} f_0(z|t)} F_0(dy, dx) < \infty. \tag{3.2}
   \]

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 (2010) shows that Laplace and Student’s \(t\)-distributions with covariate dependent parameters as well as non-parametrically 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 \(\phi.\)

**Assumption 3.2.**

1. \(\phi(y, \mu, \sigma) = \sigma^{-1}\psi((y-\mu)/\sigma),\) where \(\psi(z)\) is a bounded, continuous, symmetric around zero, and monotone decreasing in \(|z|\) probability density.
2. For any \(\mu\) 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, non-decreasing, 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 $\Theta^m$ 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 $\theta_{1:m}$,
\[
\int \log \frac{f_0(y|x)}{p(y|x, \theta_{1:m}, m)} dF_0(y, x) = \int \log \frac{f_0(y|x)}{p(y|x, \theta_{1:m}, m)} dF_0(y, x) + \int \log \frac{p(y|x, \tilde{\theta}_{1:m}, m)}{p(y|x, \theta_{1:m}, m)} dF_0(y, x).
\]
The first part of the right-hand side of this equality is bounded above by $\epsilon/2$. It suffices to show that the second part is continuous in $\theta_{1:m}$. Let $\theta_{1:m}^n$ be a sequence of parameter values converging to some $\tilde{\theta}_{1:m}$ as $n \to \infty$. For every $y$, $p(y|x, \tilde{\theta}_{1:m}, m)/p(y|x, \theta_{1:m}^n, m) \to 1$. The result will follow from the dominated convergence theorem if there is an integrable (with respect to $F_\infty$) upper bound on $|\log p(y|x, \theta_{1:m}^n, m)|$. Since $\theta_{1:m}^n \to \tilde{\theta}_{1:m}$, $\mu_j^n \in (\mu, \bar{\mu})$ and $\sigma_j^n \in (\sigma, \bar{\sigma})$ for some finite $\mu, \bar{\mu}, \sigma > 0$, and $\sigma$ for all sufficiently large $n$. From Assumption 3.2,

$$\frac{\psi(0)}{\sigma} \geq p(y|x, \theta_{1:m}^n) \geq 1_{(-\infty, \mu]}(y)\psi\left(\frac{y-\mu}{\sigma}\right) + 1_{(\mu, \infty)}(y)\psi\left(\frac{y-\mu}{\sigma}\right) + 1_{[\mu, \infty)}(y)\psi\left(\frac{y-\mu}{\sigma}\right).$$

(3.3)

The upper bound in (3.3) is a constant and the logarithm of the lower bound is integrable by part 2 of Assumption 3.2.

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

**Theorem 3.2.** Suppose

1. Assumptions 3.1-3.3 hold.

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

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

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

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

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

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
\[ J(\delta, F_n) \leq m_n \left( \log \left[ b_0 \frac{\mu_n}{\sigma_n} + b_1 \log \frac{\sigma_n}{\sigma_n} + 1 \right] + b_2 + b_3 \log \overline{Q}_n + b_4 \log \left( -\overline{Q}_n d_x \right) \right) \tag{3.4} \]
where \( b_0, b_1, b_2, b_3, \) and \( b_4 \) depend on \( \delta \) but not on \( m_n, \overline{Q}_n, \mu_n, \sigma_n, \) and \( \sigma_n. \)

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

1. A priori \((\mu_j, \sigma_j, Q_j)\) are independently identically distributed (i.i.d.) across \( j \) and independent from other parameters of the model.
2. For any \( \epsilon > 0 \), there exist \( \delta < \epsilon/4, \beta < \epsilon^2/8 \), positive constants \( c_1 \) and \( c_2 \), and sequences \( m_n, \overline{Q}_n, \mu_n, \sigma_n \uparrow \infty \) and \( \sigma_n \downarrow 0 \) with \( \sigma_n > \sigma_n \) such that
\[ m_n[\Pi(|\mu_j| > \mu_n)] + \Pi(\sigma_n > \sigma_j) + \Pi(\sigma_j > \sigma_n) + \Pi(Q_j > \overline{Q}_n)] + \Pi(m > m_n) \leq c_1 e^{-c_2 n}, \tag{3.5} \]
\[ m_n \left( \log \left[ b_0 \frac{\mu_n}{\sigma_n} + b_1 \log \frac{\sigma_n}{\sigma_n} + 1 \right] + b_2 + b_3 \log \overline{Q}_n + b_4 \log \left( -\overline{Q}_n d_x \right) \right) < n\beta, \tag{3.6} \]
where constants \( (b_0, \ldots, b_4) \) are defined in Proposition 3.1.
3. 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 \( \mu_n = \sqrt{n}, \sigma_n = 1/\sqrt{n}, \overline{Q}_n = e^n, \) 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 \((\mu_j, \sigma_j, Q_j)\) so that condition (3.5) holds. For a normal prior on \( \mu_j, \Pi(|\mu_j| > \mu_n) < c_1 e^{-c_2 n} \) for some \( c_1 \) and \( c_2 \). For an inverse gamma prior on \( \sigma_j \) we will show that \( \Pi(\sigma_n > \sigma_j) + \Pi(\sigma_j > \sigma_n) < c_1 e^{-c_2 n} \) for \( n \) large enough and some \( c_1 \) and \( c_2 \). For
n large enough

\[ \Pi(\sigma^2_n > \sigma^2_j) + \Pi(\sigma^2_j > \sigma^2_n) = \text{const} \cdot \left( \int_0^{1/n} x^{-\alpha-1} e^{-\beta/x} dx + \int_{\sigma^2_n}^{\infty} x^{-\alpha-1} e^{-\beta/x} dx \right) \]
\[ \leq \text{const} \cdot \left( \int_0^{1/n} (1/n)^{-\alpha-1} e^{-\beta/(1/n)} dx + \int_{\sigma^2_n}^{\infty} x^{-\alpha-1} dx \right) \]
\[ = \text{const} \cdot \left( n^\alpha e^{-\beta n} + e^{-2\alpha n} / \alpha \right) < c_1 e^{-nc_2}, \]

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),

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

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

where \( \theta, K, \) and \( \phi \) were defined in Section 3. Even though mixing probabilities \( \pi_j(x) \) look quite different from the mixing probabilities of KMs 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

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

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),

\[ p(y|x, \theta^{KSB}_{1:n}) = \sum_{j=1}^{n} \pi_j(x) \phi \left( \frac{y - \mu_j}{\sigma_j} \right). \]
(ii) Under Assumptions 3.1-3.3, (4.2) holds on an open neighborhood of $\theta_{1:n}^{KSB}$.

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 $\theta_{1:n}$, $\Theta^n$, $\Pi(\tilde{\theta}_{1:n} \in \Theta^n) > 0$. 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

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

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

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

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

$$
\int \log \left( \frac{f_0(y|x)}{p(y|x, \theta_{KSB}^{1:n})} \right) dF_0(y, x) \leq \int \log \left( \frac{f_0(y|x)}{p(y|x, \theta_{KM}, m)} \right) dF_0(y, x) + \int \log \left( \frac{p(y|x, \theta_{KM}, m)}{p(y|x, \theta_{KSB}^{1:n})} \right) dF_0(y, x) < \epsilon.
$$

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

By Theorem 2.1 the Kullback-Leibler property implies the following weak posterior consistency theorem.

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

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

\[
F_n = \{ p(y|x, \theta) : |\mu_j| \leq \overline{\mu}_n, Q_j \leq \overline{Q}_n, \sigma_n < \sigma_j < \overline{\sigma}_n, j = 1, \ldots, m_n, \sup_{x \in X} \sum_{j=m_n+1}^{\infty} \pi_j(x) \leq \delta \}.
\]

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 \( F_n \) in the following proposition.

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

\[
J(4\delta, F_n) \leq m_n \left( \log \left[ b_0 \frac{\overline{\mu}_n}{\sigma_n} + b_1 \log \frac{\overline{\sigma}_n}{\sigma_n} + 1 \right] + b_2 + b_3 \log \overline{Q}_n + b_4 \log m_n \right),
\]

where \( b_0, b_1, b_2, b_3, \) and \( b_4 \) depend on \( \delta \) but not on \( n, m_n, \overline{Q}_n, \overline{\mu}_n, \overline{\sigma}_n, \) and \( \sigma_n \).

A proof is given in Appendix A.

The next theorem formulates sufficient conditions for strong consistency.

**Theorem 4.2.** Suppose

1. A priori \((\alpha_j, \mu_j, \sigma_j, Q_j)\) are i.i.d. across \( j \).
2. 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 \( \sigma_n \downarrow 0 \) with \( \overline{\sigma}_n > \sigma_n \) such that

\[
m_n \left[ \Pi(|\mu_j| > \overline{\mu}_n) + \Pi(\sigma_n > \sigma_j) + \Pi(\sigma_j > \overline{\sigma}_n) + \Pi(Q_j > \overline{Q}_n) \right]
\]

\[
+ \Pi \left( \sup_{x \in X} \sum_{j=m_n+1}^{\infty} \pi_j(x) > \delta \right) \leq c_1 e^{-c_2 n},
\]

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

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

3. 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

\[
\Pi \left( \sup_{x \in X} \sum_{j=m_n+1}^{\infty} \pi_j(x) > \delta \right) < c_1 e^{-nc_2}
\]
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 $\sum_{j=m_n+1}^{\infty} \pi_j(x) = \prod_{j=1}^{m_n} (1 - \alpha_j K(-Q_j ||x - q_j||^2))$ and thus

$$\Pi \left( \sup_{x \in X} \sum_{j=m_n+1}^{\infty} \pi_j(x) > \delta \right) \leq \Pi \left( \prod_{j=1}^{m_n} (1 - \alpha_j K_j) > \delta \right),$$

where $K_j = K(-Q_j d_x) \leq K(-Q_j ||x - q_j||^2))$. The following lemma describes priors for $\alpha_j$ and $Q_j$ that imply an exponential bound on the right-hand side of (4.6).

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

$$\Pi \left( \prod_{j=1}^{m_n} (1 - \alpha_j K_j) > \delta \right) < e^{-0.5m_n \log m_n}.$$

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 $\mu$ and $\sigma$ and sequences $\overline{\mu}_n$, $\overline{\sigma}_n$, 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 $\alpha_j$ first-order stochastically dominate $Beta(1, \gamma)$ and $K_j = K(-Q_j d_x)$ first-order stochastically dominate $Beta(\gamma + 1, 1)$ for any $\gamma > 0$ (note that for $m_n = cn/\log n$, $\exp(-0.5m_n \log m_n) \leq \exp(-0.25cn)$ for large enough $n$).

Explicit priors for $Q_j$ and $\alpha_j$ 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 $K_j = \exp(-Q_j d_x) \sim Beta(\gamma + 1, 1)$, satisfy the conditions of Lemma 4.1. Also, $\Pi(Q_j > n^r) \leq c_1 e^{-nc_2}$ for $r \geq 1$.

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 \( \mu_j \) replaced by \( \beta'_j z(x) \), where \( \beta_j \in \mathbb{R}^{d_z} \) for each \( j \). We make the following assumption on the space \( Z \) and the function \( z \) to extend the consistency results to this set-up.

**Assumption 5.1.**
1. \( Z = [0, 1]^{d_z} \),
2. \( z(x)_1 = 1 \) (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 \( \mu_j \) presented in Sections 3-4 hold for models with locations \( \beta'_j z(x) \) with minor modifications. Theorem 8.1 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 \( \mu_j \) is replaced by independent normal priors on components of \( \beta_j \), 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 datasets 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,

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

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

The prior distribution we use is based on Example 4.1: \( \beta_j \sim N(\mu_\beta, H_\beta^{-1}) \), \( \sigma^2_j \sim InvGamma(\nu, b_\sigma) \), \( \alpha_j \sim Beta(a, b) \), \( q_j \sim U(0, 1) \), \( Q_j \sim \text{Exponential}(\tau) \) 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 the following data dependent values:

\[
\mu_\beta = (\bar{y}, 0)', \quad H_\beta = \text{diag}(0.5/\hat{\sigma}^2_y, 1), \quad \nu = 3, \quad b_\sigma = 2/\hat{\sigma}^2_y, \quad a = 1, \quad b = 0.5, \quad \tau = 1.5,
\]

where \( \bar{y} \) and \( \hat{\sigma}^2_y \) 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 \( (\mu_\beta, H_\beta) \) are chosen
so that observed values of $y$ are plausible. Parameters $\{\sigma^2_j\}_{j=1}^\infty$ control the variance in each mixture component and, thus, should reflect the range of observable $y$ at possible covariate values. The values of hyperparameters $(\nu, b_\sigma)$ 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 $\alpha_j$ concentrating near 1 lead to mixtures with a smaller number of components. Higher values of $\tau$ imply the prior for $Q_j$ concentrating near 0. Smaller $Q_j$ in turn imply 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 $\tau$ lead to considerable oversmoothing. Thus, we suggest using low values of $\tau$.

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 et al. (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

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

(6.3)

According to Dunson and Park (2008), 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 et al. (2007) and Dunson and Park (2008).

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, $N \in \{200, 500, 2000\}$. 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 suggest that the MCMC chains converge. The numerical standard errors (NSE) 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.

Fig 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.

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 \((b, \tau)\) can have a considerable effect on the estimation results. Smaller values of \(\tau\) (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 non-parametric kernel smoothing method of Hall et al. (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.
Fig 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 and the mean absolute error

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

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

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.

Table 1. Monte Carlo study for DGP (6.3).

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>KSB</td>
<td>0.148(0.019)</td>
<td>0.085(0.011)</td>
</tr>
<tr>
<td>np</td>
<td>0.195(0.017)</td>
<td>0.109(0.010)</td>
</tr>
<tr>
<td>Ratio KSB/np</td>
<td>0.758(0.092)</td>
<td>0.784(0.101)</td>
</tr>
</tbody>
</table>
In this particular example, the KSB model outperforms the non-parametric 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 et al. (1999):

\[ y_i = 2 \sin(\pi x_i) + \epsilon_i, \]  

(6.4)

where \( x_i \) and \( \epsilon_i \) 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 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>KSB</td>
<td>0.149(0.010)</td>
<td>0.074(0.005)</td>
</tr>
<tr>
<td>np</td>
<td>0.119(0.017)</td>
<td>0.051(0.005)</td>
</tr>
<tr>
<td>Ratio KSB/np</td>
<td>1.270(0.176)</td>
<td>1.478(0.149)</td>
</tr>
</tbody>
</table>

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 data-generating process 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.


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 non-negative,

\[
0 \leq \int \log \frac{f_0(y|x)}{p(y|x, \theta_{1:m}, m)} F_0(dy, dx) \leq \int \log \max\{1, \frac{f_0(y|x)}{p(y|x, \theta_{1:m}, m)}\} F_0(dy, dx).
\]
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 point-wise 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

$$\forall m \geq M_0, \ C_{\delta_m}(y) \cap A_0^m = \emptyset,$$

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 $j \neq 0$, 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 $\lambda$ 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 follows,

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

where $\sigma_0$ is fixed, $\sigma_m$ converges to zero as $m$ increases, and $\mu_j^m$ is the center of $A_j^m$. One can always construct a partition $A_j^m$ so that

$$\delta_m \to 0, \quad \sigma_m/\delta_m \to 0, \quad h_m/\sigma_m \to 0,$$

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

$$\frac{K(-2Q_m s_m)}{K(-Q_m s_m)} d_x^{2/d_x} \to 0, \quad s_m = d_x \lambda(B_i^m)^{2/d_x} \to 0.$$

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 $I_1^m(x, s_m) = \{i : \|q_i^m - x\|^2 \leq 2s_m\}$ and $I_2^m(x, s_m) = \{i : \|q_i^m - x\|^2 > 2s_m\}$. Since $s_m$ is the squared diagonal of $B_1^m$, there exists $i \in I_1^m(x, s_m)$ such that,

$$K(-Q^m\|x - q_i^m\|^2) \geq K(-Q^m s_m).$$  \hfill (A.4)

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

$$K(-Q^m\|x - q_i^m\|^2) \leq K(-2Q^m s_m).$$  \hfill (A.5)

Note that

$$\sum_{i \in I_1^m(x, s_m)} \frac{K(-Q^m\|x - q_i^m\|^2)}{\sum_{i=1}^{I_1^m(x, s_m)} K(-Q^m\|x - q_i^m\|^2)} \geq 1 - \frac{\sum_{i \in I_2^m(x, s_m)} K(-Q^m\|x - q_i^m\|^2)}{\sum_{i \in I_1^m(x, s_m)} K(-Q^m\|x - q_i^m\|^2)} \geq 1 - \frac{\text{card}(I_2^m(x, s_m)) K(-2Q^m s_m)}{K(-Q^m s_m)} \geq 1 - d_x^2/2 \frac{K(-2Q^m s_m)}{K(-Q^m s_m) d_x/2},$$

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

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

$$F(A_j^m\|x_i^m\) \geq \lambda(A_j^m) \inf_{z \in C_{\delta_m}(y), \|t - x\|^2 \leq 2s_m} f(z|t).$$  \hfill (A.7)

Inequalities (A.6), (A.7), and Lemma 8.3 imply that $p(y|x, \theta, m)$ exceeds

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

By (A.2) and (A.3), given some $\epsilon_1 > 0$ there exists $M_1$ such that for $m \geq M_1$ 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 $m \geq M_2$, $[f_0(y|x)/\inf_{z \in C_{\delta_m}(y), \|t - x\|^2 \leq 2s_m} f_0(z|t)] \leq (1 + \epsilon_1)$ since $\delta_m, s_m \to 0$. For any $m \geq \max\{M_1, M_2\}$

$$1 \leq \max\{1, \frac{f(y|x)}{\inf_{z \in C_{\delta_m}(y), \|t - x\|^2 \leq 2s_m} f_0(z|t)}\} \leq \max\{1, \frac{f_0(y|x)}{\inf_{z \in C_{\delta_m}(y), \|t - x\|^2 \leq 2s_m} f_0(z|t)}\} \leq 1 + \frac{\epsilon_1}{1 - \epsilon_1}.$$

Thus, $\log \max\{1, f_0(y|x)/\inf_{z \in C_{\delta_m}(y), \|t - x\|^2 \leq 2s_m} f_0(z|t)\} \to 0$ 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:

\[
p(y|x, \theta, m) \geq \left[ 1 - d_x^{d_x/2} \frac{K(-2Q^m_s m)}{K(-Q^m_s s_m)^{d_x/2}} \right] \cdot \left[ 1 - 1_{A_0^n}(y) \cdot \inf_{\|z-y\| \leq r, \|t-x\| \leq r} f_0(z|t) \cdot \sum_{j:A^n_j \subset C_r(y) \cap (A^n_0)^c} \lambda(A^n_j) \phi(y, \mu_j^m, \sigma_m) \right]
\]

(A.9)

For any \( m \) larger than some \( M_3 \), the Riemann sum in (A.9) is bounded below by 1/4 (by Lemma 8.3) and

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

(by (A.3)).

Choose \( \sigma_0 \) so that for \( y \in A^n_0, 1 > 1/4 \geq \lambda(C_r(y) \cap A^n_0) \phi(y, 0, \sigma_0) \geq r \phi(y, 0, \sigma_0) \), for example, \( \sigma_0 = 8r \psi(0) \). Then

\[
\log \max \left\{ 1, \frac{f_0(y|x)}{p(y|x, \theta, m)} \right\} \leq \log \max \left\{ 1, \frac{f_0(y|x)}{\inf_{\|z-y\| \leq r, \|t-x\| \leq r} f_0(z|t) \cdot \phi(y, 0, \sigma_0) \cdot (r/2)} \right\}
\]

\[
= \log \frac{1}{\phi(y, 0, \sigma_0)(r/2)} \max \left\{ \phi(y, 0, \sigma_0)(r/2), \frac{f_0(y|x)}{\inf_{\|z-y\| \leq r, \|t-x\| \leq r} f_0(z|t)} \right\}
\]

\[
\leq - \log(\phi(y, 0, \sigma_0)(r/2)) + \log \frac{f_0(y|x)}{\inf_{\|z-y\| \leq r, \|t-x\| \leq r} f_0(z|t)}.
\]

(A.10)

The first expression in (A.10) is integrable by Assumption 3.2 part 2. The second expression in (A.10) is integrable by Assumption 3.1 part 3. Thus the proposition is proved.

\[\blacksquare\]

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 \( \alpha_j \)'s equal to zero). It is proved in Lemma 8.1 below that for \( p(y|x, \theta^m_{m_n}, m_n) \in \mathcal{F}_n, \theta^m_{m_n} = \{ Q^i_j, \mu^i_j, \sigma^i_j, q^i_j, \alpha^i_j \}_{j=1}^{m_n}, i = 1, 2, \) and any
\[ x \in X, \]
\[ \int |p(y|x, \theta^1_{m_n}, m_n) - p(y|x, \theta^2_{m_n}, m_n)| \, dy \]  
\[ \leq 2 \max_{j=1,...,m_n} \left( \psi(0) \frac{|\mu^1_j - \mu^2_j|}{\sigma_j^1} + \frac{|\sigma^1_j - \sigma^2_j|}{\min(\sigma^1_j, \sigma^2_j)} \right) \]
\[ + \frac{2}{K(-Q'_n d_x)} \sum_{j=1}^{m_n} |\alpha^1_j - \alpha^2_j|, \]
\[ + \frac{2K'_n}{K(-Q'_n d_x)} \max_{j=1,...,m_n} |Q^1_j - Q^2_j| \]
\[ + \frac{4K'_n}{K(-Q'_n d_x)} \max_{j=1,...,m_n} \max_{1 \leq i \leq d_x} |q^2_{j,i} - q^1_{j,i}|, \]

where \( \overline{K} \) is a finite fixed bound on the derivative of \( K \) (Assumption 3.3) and \( \tilde{\alpha}^i_j = \alpha^i_j / \sum_{i=1}^{m_n} \alpha^i_j \).

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

Consider \( G_{\mu \sigma} \) first. Let \( \zeta = \min(\delta/12, 1) \). Define \( \sigma_h = \sigma_n(1 + \zeta)^h, h \geq 0. \) Let \( H \) be the smallest integer such that \( \sigma_H = \sigma_n(1 + \zeta)^H \geq \overline{\sigma}_n. \) This implies that \( H \leq \log(\frac{\overline{\sigma}_n}{\sigma_n}) + 1 \) and for any \( h \geq 1, \frac{2^{h-1} \sigma_n}{\sigma_{n-1}} \leq \frac{\delta}{6}. \)

Define \( N_{ij} = \left\lfloor \frac{24\psi(0)}{\sigma_n} \right\rfloor \).

For any \( (\mu^1, \sigma^1) \) and \( (\mu^2, \sigma^2) \) in \( E^\mu_{ij}, \)
\[ \left( \frac{2}{\sigma^1} \right)^2 \left( \frac{\mu^1 - \mu^2}{\sigma^1} + \frac{1}{\sigma^1} \right) \leq \frac{\delta}{3}. \]

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

\[ N_{\mu \sigma} \leq \sum_{j=1}^{H} \left( \frac{24}{\delta} \frac{\overline{\sigma}_n}{\sigma_j} + 1 \right) = \frac{24}{\delta} \frac{\overline{\sigma}_n}{\sigma_j} \sum_{j=1}^{H} (1 + \zeta)^{-j} + H \]
\[ \leq \frac{24}{\delta} \frac{\overline{\sigma}_n}{\sigma_j} + \frac{1}{\log(1 + \zeta)} \log(\frac{\overline{\sigma}_n}{\sigma_j}) + 1 \]
\[ = c_0 \frac{\overline{\sigma}_n}{\sigma_j} + c_1 \log \frac{\overline{\sigma}_n}{\sigma_j} + 1, \quad (A.12) \]
where \( c_0, c_1 \) depend on \( \delta \), but not on \( n \).

Next, consider \( G_\alpha \). Since only the normalized values of \( \alpha_j \)'s appear on the right-hand side of (A.11), \( G_\alpha \) can include only points that belong to the \((m_n - 1)\)-dimensional simplex. Thus, we can take \( G_\alpha \) 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 \( N_\alpha \) 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 \( \delta \), but not on \( n \).

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

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

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

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

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

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

Since for any \( q_i^1, r_i \in [0, 1] \) there exists \( r_l \) such that \(|q_i^1 - r_i| \leq \delta K(-\overline{Q}_n d_x)/(12K' d_x \overline{Q}_n)\), 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 \( N_{\mu \sigma}, N_\alpha, N_Q, \text{ and } N_q \) imply

\[
J(\delta, \mathcal{F}_n) \leq m_n \log(N_{\mu \sigma} N_Q N_q) + \log N_\alpha \leq m_n \left( \log \left[ b_0 \tfrac{\overline{r}_n}{\overline{2}_n} + b_1 \log \tfrac{\overline{\sigma}_n}{\overline{2}_n} + 1 \right] + b_2 + b_3 \log \overline{Q}_n + b_4 \log K(-\overline{Q}_n d_x) \right),
\]

where \( b_0, b_1, b_2, b_3, b_4 \) do not depend on \( n \).

**Lemma 8.1.** 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 $x \in X$

$$
\int |f_1(y|x) - f_2(y|x)|\,dy 
= \int \left| \sum_{j=1}^{m_n} \pi_j^1(x) \frac{1}{\sigma_j^1} \left( \frac{y - \mu_j^1}{\sigma_j^1} \right) - \sum_{j=1}^{m_n} \pi_j^2(x) \frac{1}{\sigma_j^2} \left( \frac{y - \mu_j^2}{\sigma_j^2} \right) \right| \,dy 
\leq \int \left| \sum_{j=1}^{m_n} \pi_j^1(x) \psi_j^1(y) - \pi_j^2(x) \psi_j^2(y) \right| \,dy + \int \sum_{j=1}^{m_n} |\pi_j^1(x) - \pi_j^2(x)| \,dy
= \sum_{j=1}^{m_n} \pi_j^1(x) \int |\psi_j^1(y) - \psi_j^2(y)|\,dy + \sum_{j=1}^{m_n} |\pi_j^1(x) - \pi_j^2(x)|,
$$

(A.13)

where $\psi_j^2(y) = (\sigma_j^2)^{-1} \psi((y - \mu_j^2)/\sigma_j^2)$. 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

$$
\int |\psi_j^1(y) - \psi_j^2(y)|\,dy 
= \int \left| \frac{1}{\sigma_j^1} \psi \left( \frac{y - \mu_j^1}{\sigma_j^1} \right) - \frac{1}{\sigma_j^2} \psi \left( \frac{y - \mu_j^2}{\sigma_j^2} \right) \right| \,dy 
\leq \int \left| \frac{1}{\sigma_j^1} \psi \left( \frac{y - \mu_j^1}{\sigma_j^1} \right) \right| \,dy + \int \left| \frac{1}{\sigma_j^2} \psi \left( \frac{y - \mu_j^2}{\sigma_j^2} \right) \right| \,dy.
$$

Note that

$$
\int \left| \frac{1}{\sigma_j^1} \psi \left( \frac{y - \mu_j^1}{\sigma_j^1} \right) \right| \,dy = 2 \int_{\psi(0)}^{\psi(y_1)} \frac{1}{\sigma_j^1} \psi(\tau) \,d\tau = 2 \psi(0) \frac{\mu_j^1 - \mu_j^2}{\sigma_j^1},
$$

(A.14)

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

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

Combining the two pieces together we find that

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

(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 \(K_j^1(x) = K(-Q_j^1 ||x - q_j^1||^2)\) and \(A_i(x) = \sum_{j=1}^{m_n} \tilde{\alpha}_j K_j^1(x)\). Then for any \(x \in X\)

\[
\sum_{j=1}^{m_n} |\pi_j^1(x) - \pi_j^2(x)| = \sum_{j=1}^{m_n} \left| \frac{\tilde{\alpha}_j K_j^1(x)}{A_1(x)} - \frac{\tilde{\alpha}_j^2 K_j^2(x)}{A_1(x)} \right| \leq \frac{1}{A_1(x) A_2(x)} \sum_{j=1}^{m_n} \frac{|\tilde{\alpha}_j K_j^1(x) A_2(x) - \tilde{\alpha}_j^2 K_j^2(x) A_1(x) + \tilde{\alpha}_j K_j^1(x) A_2(x) - \tilde{\alpha}_j^2 K_j^2(x) A_2(x)|}{A_1(x) A_2(x)} \\
\leq \frac{\sum_{j=1}^{m_n} |\tilde{\alpha}_j K_j^1(x) - \tilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)} + \frac{\sum_{j=1}^{m_n} |\tilde{\alpha}_j^2 K_j^2(x)| A_2(x) - A_1(x)}{A_1(x)} \\
= \frac{\sum_{j=1}^{m_n} |\tilde{\alpha}_j K_j^1(x) - \tilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)} + \frac{|A_2(x) - A_1(x)|}{A_1(x)} \\
\leq 2 \frac{\sum_{j=1}^{m_n} |\tilde{\alpha}_j K_j^1(x) - \tilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)} = \frac{\sum_{j=1}^{m_n} |\tilde{\alpha}_j K_j^1(x) - \tilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)} + \frac{|\tilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)} \\
\leq 2 \frac{1}{K(-Q_n d_x)} \left[ \max_{j=1, \ldots, m_n} |K_j^1(x) - K_j^2(x)| + \sum_{j=1}^{m_n} |\tilde{\alpha}_j^1 - \tilde{\alpha}_j^2| \right].
\]  

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

\[
|K(-Q_j^1 ||x - q_j^1||^2) - K(-Q_j^2 ||x - q_j^2||^2)| \\
\leq |K(-Q_j^1 ||x - q_j^1||^2) - K(-Q_j^2 ||x - q_j^1||^2)| + |K(-Q_j^2 ||x - q_j^1||^2) - K(-Q_j^2 ||x - q_j^2||^2)| \\
\leq \overline{K}' (||x - q_j^1||^2) |Q_j^1 - Q_j^2| + \overline{K}' Q_n \sum_{l=1}^{d_x} 2(||q_{j,l} - q_{j,l}||^2) \\
\leq \overline{K}' d_x |Q_j^2 - Q_j^1| + 2\overline{K}' d_x \overline{Q}_n \max_{l=1, \ldots, d_x} ||q_{j,l} - q_{j,l}||
\]  

(A.17)

\[\Box\]

**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

\[
\theta^{KSB}_{1:m,M} = \{\alpha_j \delta, Q_j, q_j, \mu_j, \sigma_j\}_{j=1}^{m} \times \cdots \times \{\alpha_j \delta, Q_j, q_j, \mu_j, \sigma_j\}_{j=1}^{m},
\]

So that \(\theta^{KSB}_{1:m,M}\) is given by \(M\) repetitions of \(\theta^{KM}\) (except \(\alpha_j\)'s are multiplied by \(\delta\)). For brevity
let \( K_j(x) = K(-Q_j||x - q_j||^2) \). Then

\[
p(y|x, \theta_{1,m}^{KS}) = \sum_{j=1}^{m} \alpha_j \delta K_j(x) \prod_{l<j} (1 - \alpha_l \delta K_l(x)) \phi(y, \mu_j, \sigma_j)
\]

\[
= \sum_{h=1}^{M} \left( \sum_{j=1}^{m} \phi(y, \mu_j, \sigma_j) \alpha_j \delta K_j(x) \prod_{l<j} (1 - \alpha_l \delta K_l(x)) \right) \left[ \prod_{i=1}^{m} (1 - \alpha_i \delta K_i(x)) \right]^{h-1}
\]

\[
= \left( \sum_{j=1}^{m} \phi(y, \mu_j, \sigma_j) \alpha_j \delta K_j(x) \prod_{l<j} (1 - \alpha_l \delta K_l(x)) \right) \sum_{h=1}^{M} \left[ \prod_{i=1}^{m} (1 - \alpha_i \delta K_i(x)) \right]^{h-1}
\]

\[
> \sum_{j=1}^{m} \phi(y, \mu_j, \sigma_j) \alpha_j \delta K_j(x) \prod_{i=1}^{m} (1 - \alpha_i \delta K_i(x)) \left[ \prod_{i=1}^{m} (1 - \alpha_i \delta K_i(x)) \right]^{M}
\]

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

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

Let \( \delta < (1 - \exp(-\epsilon/(2m))) / \max_{j=1,...,m} \alpha_j \), then \( [1 - \delta \max_{j=1,...,m} \alpha_j]^{m} > \exp \{-\epsilon/2\} \).

There exists \( j \) such that \( \alpha_j > 1/m \) and by Assumption 3.3 \( K_j(x) > K(-Qd_x) \) for any \( x \in X \), where \( Q = \max_{j=1,...,m} Q_j \). Therefore,

\[
\prod_{i=1}^{m} (1 - \alpha_i \delta K_i(x)) < 1 - \frac{\delta K(-Qd_x)}{m}.
\]

For \( M > \frac{\log(1-e^{-\epsilon/2})}{\log(1-\frac{\delta K(-Qd_x)}{m})} \) the following is true

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

Thus, \( \log(p(y|x, \theta^{KM}, m)/p(y|x, \theta_{1,m}^{KS})) < \epsilon \) and the proposition claim (i) follows.

(ii) By part (i) of the proposition, (4.2) holds for \( \epsilon/2 \) and some \( \tilde{\theta}_{1,m}^{KS} \). 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 \( \theta_{1,m}^{KS} \) and \( \tilde{\theta}_{1,m}^{KS} \) 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 8.2 below that for \( p(y|x, \theta^i) \in \mathcal{F}_n, \ i = 1, 2 \), and any \( x \in X \),

\[
\int |p(y|x, \theta^1) - p(y|x, \theta^2)|dy \leq 2\delta \tag{A.18}
\]

\[
+ 2 \max_{j=1, \ldots, m_n} \left( \psi(0) \left| \frac{\mu_j^1 - \mu_j^2}{\sigma_j^1} + \frac{|\sigma_j^1 - \sigma_j^2|}{\min(\sigma_j^1, \sigma_j^2)} \right| \right)
\]

\[
+ m_n^2 \max_{j=1, \ldots, m_n} |\alpha_j^1 - \alpha_j^2|,
\]

\[
+ m_n^2 \overline{K}' \max_{j=1, \ldots, m_n} |Q_j^1 - Q_j^2|
\]

\[
+ m_n^2 2\overline{K}' \max_{j=1, \ldots, m_n} \max_{l=1, \ldots, d_x} |q_{j,l}^1 - q_{j,l}^2|,
\]

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 \( G_\alpha \) to be a uniform grid on \([0, 1]^{m_n} \),

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

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

Since for any \( \alpha_j^1 \in [0, 1] \) there exists \( \kappa_l \) such that \( |\alpha_j^1 - \kappa_l| \leq \delta/(3m_n^2) \), 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 \( Q_i = (2i - 1)\delta/(3\overline{K}'d_xm_n^2) \), \( i = 1, \ldots, N_Q \).

\[
N_Q = \left\lceil \frac{3\overline{K}'d_xm_n^2}{2\delta} \right\rceil.
\]

Since for any \( Q_j^1 \in [0, \overline{Q}_n] \) there exists \( Q_i \in G_Q \) such that \( |Q_j^1 - Q_i| \leq \delta/(3\overline{K}'d_xm_n^2) \), 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 \([0, 1]^{d_x} \),

\[
G_q = \left\{ r_l = (2l - 1) \frac{\delta}{6Kd_xQ_nm_n^2}, \ l = 1, \ldots, (N_q)^{1/d_x} \right\}^{d_x},
\]

\[
N_q = \left\lceil \frac{6\overline{K}'d_xQ_nm_n^2}{2\delta} \right\rceil.
\]

Since for any \( q_{j,l}^1 \in [0, 1] \) there exists \( r_l \) such that \( |q_{j,l}^1 - r_l| \leq \delta/(6\overline{K}'d_xQ_nm_n^2) \), 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 $N_{\mu\sigma}$, $N_{\alpha}$, $N_{Q}$, and $N_{q}$ imply

$$J(4\delta, \mathcal{F}_n) \leq m_n \log(N_{\mu\sigma} N_{Q} N_{q}) + \log N_{\alpha}$$

$$\leq m_n \left( \log \left[ b_0 \frac{\mu_n}{\sigma_n} + b_1 \log \frac{\sigma_n}{\mu_n} + 1 \right] + b_2 + b_3 \log Q_n + b_4 \log m_n \right).$$

\[\blacksquare\]

**Lemma 8.2.** Inequality (A.18) holds.

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

$$\int_X \sum_{j=1}^{\infty} |\pi_1^j(x) \phi(y; \mu_j^1, \sigma_j^1) - \pi_2^j(x) \phi(y; \mu_j^2, \sigma_j^2)| \, dy$$

$$\leq \int_X \sum_{j=1}^{m_n} |\pi_1^j(x) \phi(y; \mu_j^1, \sigma_j^1) - \phi(y; \mu_j^2, \sigma_j^2)| \, dy$$

$$+ \int_X \sum_{j=1}^{m_n} |\pi_1^j(x) - \pi_2^j(x)| \phi(y; \mu_j^2, \sigma_j^2) \, dy$$

$$+ \sum_{j=m_n+1}^{\infty} |\pi_1^j(x) - \pi_2^j(x)|$$

$$\leq \sum_{j=1}^{m_n} \pi_1^j(x) \int_X |\phi(y; \mu_j^1, \sigma_j^1) - \phi(y; \mu_j^2, \sigma_j^2)| \, dy$$

$$+ \sum_{j=1}^{m_n} ||\pi_1^j - \pi_2^j||_1 + \sup_{x \in X} \sum_{j=m_n+1}^{\infty} |\pi_1^j(x)| + |\pi_2^j(x)|$$

$$\leq \sum_{j=1}^{m_n} \pi_1^j(x) \int_X |\phi(y; \mu_j^1, \sigma_j^1) - \phi(y; \mu_j^2, \sigma_j^2)| \, dy$$

$$+ \sum_{j=1}^{m_n} ||\pi_1^j - \pi_2^j||_1 + 2\delta$$

where the last inequality is true by construction of $\mathcal{F}_n$ as $\sup_{x \in X} \sum_{j=m_n+1}^{\infty} |\pi_i^j(x)| \leq \delta$ for $i = 1, 2$. As shown in Lemma 8.1, the first expression on the r.h.s. of the last inequality is
bounded by the first bound in (A.18).

\[ |\pi_j(x) - \pi_j^m(x)| = \left| \alpha_j^1 K_j^1(x) \prod_{i<j} (1 - \alpha_i^1 K_i^1(x)) - \alpha_j^2 K_j^2(x) \prod_{i<j} (1 - \alpha_i^2 K_i^2(x)) \right| \]

\[ \leq \left| \alpha_j^1 K_j^1(x) - \alpha_j^2 K_j^2(x) \right| \prod_{i<j} (1 - \alpha_i^1 K_i^1(x)) + \prod_{i<j} (1 - \alpha_i^2 K_i^2(x)) \]

\[ \leq \left| \alpha_j^1 K_j^1(x) - \alpha_j^2 K_j^2(x) \right| \prod_{i<j} (1 - \alpha_i^1 K_i^1(x)) - \prod_{i<j} (1 - \alpha_i^2 K_i^2(x)) \]

\[ \leq \sum_{i=1}^j \left| \alpha_i^1 - \alpha_i^2 \right| + |K_j^1(x) - K_j^2(x)| \]

Using the bound on \(|K_j^1(x) - K_j^2(x)|\) from (A.17) in Lemma 8.1 and noting that \(\sum_{j=1}^m j \leq m^2\), complete the proof.

**Proof.** Lemma 4.1.

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

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

\[ Pr(c < x) = \frac{\int_0^x \gamma^m t^{m-1} e^{-\gamma t} dt}{(m-1)!} = \frac{\int_0^\gamma t^{m-1} e^{-t} dt}{(m-1)!} < (\gamma x)^m / m! \]

\[ = \frac{(\gamma x)^m}{\exp \{ m \log m - m + O(\log m) \}} \] (by Sterling formula)

\[ = \exp \{ -m \log m + m + m \log(\gamma x) - O(\log(m)) \} \]

\[ = \exp(-0.5m \log m) \frac{\exp(m \log(\gamma x) + m + O(\log(m)))}{\exp(0.5m \log m)} < \exp(-0.5m \log m) \]

when \(m\) is sufficiently large.

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

\[
\Pi \left( \prod_{j=1}^{m_n} (1 - \alpha_j K_j) > \delta \right)
\]

\[
= \int \Pi \left( \alpha_1 K_1 < 1 - \frac{\delta}{\prod_{j \neq 1} (1 - \alpha_j K_j)} |\alpha_j, K_j, j \neq 1| \right) d\Pi(\alpha_j, K_j, j \neq 1)
\]

\[
\leq \int \Pi \left( \alpha_1 \alpha_2 < 1 - \frac{\delta}{\prod_{j \neq 1} (1 - \alpha_j K_j)} |\alpha_j, K_j, j \neq 1| \right) d\Pi(\alpha_j, K_j, j \neq 1)
\]

\[
\leq \int \Pi \left( b_1 < 1 - \frac{\delta}{\prod_{j \neq 1} (1 - \alpha_j K_j)} |\alpha_j, K_j, j \neq 1| \right) d\Pi(\alpha_j, K_j, j \neq 1)
\]

\[
= \Pi \left( (1 - b_1) \prod_{j \neq 1} (1 - \alpha_j K_j) > \delta \right)
\]

(repeat for \( b_2, \ldots, b_{m_n} \))

\[
\leq \Pi \left( \prod_{j=1}^{m_n} (1 - b_j) > \delta \right) = \Pi \left( \sum_{j=1}^{m_n} - \log(1 - b_j) < -\log(\delta) \right)
\]

\[
= \Pi \left( c < -\log(\delta) \right) < e^{-0.5m_n \log m_n}.
\]

**Lemma 8.3.** 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 \( C_\delta(y) = [y - \delta, y + \delta] \subset A_j \) is an interval with center \( y \) and length \( \delta \). 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,

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

Note next that

\[
\int_{\bigcup_{j \in J} [A_j \cap C_\delta(y)]} \phi(y, \mu - h, \sigma)d\mu \geq \int_{y-h}^{y} \phi(y, \mu - h, \sigma)d\mu = \int_{y-h}^{y-2h} \phi(y, \mu, \sigma)d\mu
\]

\[
\geq \int_{y-\delta}^{y} \phi(y, \mu, \sigma)d\mu - \int_{y-2h}^{y} \phi(y, \mu, \sigma)d\mu
\]

\[
\geq \int_{y-\delta}^{y} \phi(y, \mu, \sigma)d\mu - \frac{2h\psi(0)}{\sigma}
\]
By symmetry the same results can be obtained for $J = \{ j : A_j \cap C_\delta(y) \subset [y, y + \delta] \}$. Thus
\[
\sum_{j=1}^{m} \lambda(A_j \cap C_\delta(y)) \phi(y, \mu_j, \sigma) \geq \int_{y-\delta}^{y+\delta} \phi(y, \mu, \sigma) d\mu - \frac{2h\psi(0)}{\sigma}. 
\]
A change of variables delivers the claim of the lemma.

**Theorem 8.1.** 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
\[
\theta = \{ Q_j, \mu_j, \sigma_j, q_j, \alpha_j \}_{j=1}^{\infty} \in \Theta = (R_+ \times Y \times R_+ \times X \times (0, 1))^{\infty}
\]
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 $\theta_1: m$, $\theta_{KM}$, $\theta_{KSB}$, and $\theta_{KSB}^{1:n}$).

(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 $|\mu_j| \leq \mu_n$ in the sieve definitions by $|\beta_{j,l}| \leq \beta_n$. Then, the entropy bounds in Propositions 3.1 and 4.3 are changed as follows: the term
\[
\log \left[ b_0 \frac{\mu_n}{\sigma_n} + b_1 \log \frac{\sigma_n}{\sigma_n} + 1 \right]
\]
in the bounds is replaced by
\[
d_z \log \left[ b_0 \frac{\beta_n}{\sigma_n} + b_1 \log \frac{\sigma_n}{\sigma_n} + 1 \right].
\]

(iv) Replace $\Pi(|\mu_j| > \mu_n)$ in (3.4) and (4.4) with $\sum_{l=1}^{d_z} \Pi(|\beta_{j,l} > \beta_n)$ 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 $|\beta_{j,l}^n| \leq \beta_j$ 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 $\mu = -\overline{\beta}$ and $\mu = \overline{\beta}$ 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 $|\mu_j^1 - \mu_j^2|$ with $\sum_{l=1}^{d_z} |\beta_{j,l}^1 - \beta_{j,l}^2|$. Thus,
Thus, when \(\delta/b\) where the definition of \(G\) consists of centers of sets \(\lfloor 2k\delta/n\rfloor\), 1 \(\leq\) \(i_k\) \(\leq\) \(N_j\), and 1 \(\leq\) \(j\) \(\leq\) \(H\). If \((\beta_1, \sigma_1), (\beta_2, \sigma_2)\) \(\in\) \(E_{\beta_\sigma}^{i_1, \ldots, i_d}\), then
\[
2\psi(0)\sum_{l=1}^d |\beta_{1,l} - \beta_{2,l}| + 2 \frac{|\sigma_1 - \sigma_2|}{\min(\sigma_1, \sigma_2)} \leq \frac{\delta}{3}.
\] (A.20)

Thus, when \(G_{\beta\sigma}\) consists of centers of sets \(E_{\beta_\sigma}^{i_1, \ldots, i_d}\), an analog of the first bound in (A.11) can be made no larger than \(\delta/3\) with \((\beta_1^2, \sigma_1^2, \ldots, \beta_m^2, \sigma_m^2)\) \(\in\) \([G_{\beta\sigma}]^{mn}\). The number of points in \(G_{\beta\sigma}\), \(N_{\beta\sigma}\) = \(\sum_{j=1}^H N_j^d\) \(\leq\) \((\sum_{j=1}^H N_j)^d\). By the same arguments as in deriving (A.12)
\[
N_{\beta\sigma} \leq \left( b_0 + b_1 \log \frac{\sigma_n}{\sigma_n} + 1 \right)^d,
\]
where \(b_0, b_1\) depend on \(\delta\), 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).


9.1. 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 \((\beta_j, \sigma_j^2) \sim \text{InvGamma}(\nu, b_0)\), \(\beta_j \sim \text{Beta}(a, b)\), \(q_j \sim U(0, 1)\), \(Q_j \sim \text{Exponential}(\tau)\) 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 \(p(y_i|x_i, \theta, u_i, z_i = j) = \phi(y_i; \beta_{j,0} + \beta_{j,1}x_i, \sigma_j^2)\) 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
\[
p(\theta, Z, U|Y, X) \propto \prod_{i=1}^N [p(y_i|x_i, \theta, z_i)p(u_i|z_i, x_i, \theta)p(z_i|x_i, \theta)] \cdot \Pi(\theta)
\]
\[
= \prod_{i=1}^N \phi(y_i; \beta_{z_i,0} + \beta_{z_i,1}x_i, \sigma_{z_i}^2)1(u_i < \pi_{z_i}(x_i; \alpha, \psi, q)) \cdot \Pi(\theta),
\] (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 $j \leq M$, let $T_j = \sum_{i=1}^{N} 1 \{ z_i = j \}$. From posterior density (B.1) we find that

$$p(\beta_j | Y, Z, X, U, \theta \setminus \beta_j) \propto \phi(\beta_j; \mu_\beta, H^{-1}_\beta) \prod_{i: z_i = j} \phi(y_i; \beta_{j0} + \beta_{j1} x_i, \sigma_j^2).$$

This leads to the conditional posterior distribution for $\beta_j$

$$\beta_j | (Y, Z, X, U, \theta \setminus \beta_j) \sim N\left( \mu_\beta, H^{-1}_\beta \right),$$

where

$$H_\beta = H_\beta + \sigma_j^{-2} \sum_{i: z_i = j} \begin{pmatrix} 1 \\ x_i \\ x_i' \end{pmatrix},$$

$$\mu_\beta = H^{-1}_\beta \left( H_\beta \mu_\beta + \sigma_j^{-2} \sum_{i: z_i = j} \begin{pmatrix} x_i \\ x_i y_i \end{pmatrix} \right).$$

Similarly,

$$p(\sigma_j^2 | Y, Z, X, U, \theta \setminus \sigma_j) \propto InvGamma(\sigma_j^2; \nu, b_\sigma) \prod_{i: z_i = j} \phi(y_i; \beta_{j0} + \beta_{j1} x_i, \sigma_j^2).$$

Then the conditional posterior distribution of $\sigma_j$ is

$$\sigma_j^2 | (Y, Z, X, U, \theta \setminus \sigma_j) \sim InvGamma\left( \nu + T_j/2, \left( b_\sigma^{-1} + 0.5 \sum_{i: z_i = j} (y_i - \beta_{j0} - \beta_{j1} x_i)^2 \right)^{-1} \right).$$

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 $\alpha_j^* \sim \text{Markov transition density}$

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

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

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

(B.2)

where

\[
p(\alpha_j^*|X, Z^m, Y, \theta^m \setminus \alpha_j^m) \propto \Pi(\alpha_j^*) \prod_{i: z_i \geq j} \pi_{z_i}(x_i; \alpha_j^*, \alpha_m \setminus \alpha_j^m, Q, q).
\]

We use the following transition density for \( \alpha_j \),

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

where

\[
b(\alpha_j^m) = \sum_{i: z_i \geq j} \log\left(1 - \frac{\alpha_j^m \pi \left(-Q_j ||x_i - q_j||^2\right)}{\log(1 - \alpha_j^m)}\right).
\]

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 \( \alpha_j^m \).

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

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

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 \( i = 1, \ldots, N \),

\[
\sum_{j=1}^{M^*} \pi_j(x_i) > 1 - u_i.
\]

(B.4)
As described below in Step 7 of the algorithm, this condition on $M^*$ guarantees that draws of $\alpha_j, q_j, \beta_j, \sigma^2_j$, and $Q_j$ necessary for updating $Z$ are available.

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

7. Updating $Z$. Note that

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

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

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

9.2. 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 $J_l \geq \max Z^{(l)}$,

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

where the integral in the second line of the equation can be evaluated numerically. Note that when $J_l = \infty$, the right-hand side 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 Section 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 $x \in \{0.25, 0.5, 0.75\}$ unless specified otherwise.
10. 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 $H^{-1} = \text{diag}(100 \cdot \text{var}(Y), 1)$;
(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.

Fig 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.
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 $\gamma$ limit the dependence on covariates and lead to oversmoothing. Prior (iii) performs poorly as the choice of $b_\sigma$ implies very high variance of responses within mixture components.

**Fig 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.

**Economics Department,**
**Princeton University, Princeton, NJ 08544**
E-mail: anorets@princeton.edu

**Institute for Advanced Studies,**
**Department of Economics and Finance,**
**Stumpergasse 56, 1060 Vienna, Austria**
E-mail: pelenis@ihs.ac.at