282

Reihe Ökonomie

Economics Series

Posterior Consistency in Conditional Density Estimation by Covariate Dependent Mixtures

Andriy Norets, Justinas Pelenis

Institut für Höhere Studien
Institute for Advanced Studies
Vienna

282

Reihe Ökonomie

Economics Series

Posterior Consistency in Conditional Density Estimation by Covariate Dependent Mixtures

Andriy Norets, Justinas Pelenis

December 2011

Institut für Höhere Studien (IHS), Wien
Institute for Advanced Studies, Vienna

Contact:

Andriy Norets
Department of Economics
313 Fisher Hall
Princeton University
Princeton, NJ 08544, USA
email: anorets@princeton.edu

Justinas Pelenis
Department of Economics and Finance
Institute for Advanced Studies
Stumpergasse 56
1060 Vienna, Austria
email: pelenis@ihs.ac.at

Founded in 1963 by two prominent Austrians living in exile – the sociologist Paul F. Lazarsfeld and the economist Oskar Morgenstern – with the financial support from the Ford Foundation, the Austrian Federal Ministry of Education and the City of Vienna, the Institute for Advanced Studies (IHS) is the first institution for postgraduate education and research in economics and the social sciences in Austria. The **Economics Series** presents research done at the Department of Economics and Finance and aims to share “work in progress” in a timely way before formal publication. As usual, authors bear full responsibility for the content of their contributions.

Das Institut für Höhere Studien (IHS) wurde im Jahr 1963 von zwei prominenten Exilösterreichern – dem Soziologen Paul F. Lazarsfeld und dem Ökonomen Oskar Morgenstern – mit Hilfe der Ford-Stiftung, des Österreichischen Bundesministeriums für Unterricht und der Stadt Wien gegründet und ist somit die erste nachuniversitäre Lehr- und Forschungsstätte für die Sozial- und Wirtschafts-wissenschaften in Österreich. Die **Reihe Ökonomie** bietet Einblick in die Forschungsarbeit der Abteilung für Ökonomie und Finanzwirtschaft und verfolgt das Ziel, abteilungsinterne Diskussionsbeiträge einer breiteren fachinternen Öffentlichkeit zugänglich zu machen. Die inhaltliche Verantwortung für die veröffentlichten Beiträge liegt bei den Autoren und Autorinnen.

Abstract

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

Keywords

Bayesian nonparametrics, posterior consistency, conditional density estimation, mixtures of normal distributions, location-scale mixtures, smoothly mixing regressions, mixtures of experts, dependent Dirichlet process, kernel stick-breaking process

JEL Classification

C11, C14

Comments
First version: November 2010, current version: December 1, 2011. We are grateful to Ulrich Müller for helpful discussions. All remaining errors are ours.

Contents

1. Introduction. Estimation of conditional distributions is an important problem in empirical research. There are two alternative approaches to modeling conditional densities in the Bayesian framework. First, the conditional distributions of interest can be obtained as a byproduct of the joint distribution estimation. Second, the conditional distribution can be modeled directly and the marginal distribution of the covariates can be left unspecified. Bayesian nonparametric modeling of densities involves specifying a flexible prior on the space of densities. Widely accepted minimal requirement for such priors is posterior consistency (see Ghosh & Ramamoorthi (2003) for a textbook treatment). The theory of posterior consistency for (unconditional) density estimation is well developed. However, if only conditional density is of interest modeling marginal distribution of covariates is an unnecessary complication. While there are many proposed methods for direct conditional density estimation, their consistency properties are largely unknown. We address this gap in the literature by demonstrating consistency for Bayesian nonparametric procedures based on countable mixtures of location-scale densities with covariate dependent mixing probabilities. The mixing probabilities are modeled in two ways. First, we consider finite covariate dependent mixture models, in which the mixing probabilities are proportional to a product of a constant and a kernel and a prior on the number of mixture components is specified. Second, we consider kernel stick-breaking processes of Dunson & 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. Below, we provide a more detailed overview of the literature and our contribution.

Practical Bayesian nonparametric approaches to density estimation are mostly based on mixtures of distributions.[1] 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 & West (1995) who used Polya urn representation of the Dirichlet process from Blackwell & MacQueen (1973) (see Dey et al. (1998) for a more extensive list of references and applications). An alternative approach to modeling mixing distribution 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 & 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 the-

[1]There is also mostly theoretical literature on Gaussian process priors for density estimation, see, for example, Tokdar & Ghosh (2007) and van der Vaart & van Zanten (2008).

ory 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 Dirichlet process mixture of normals context. An alternative approach to consistency was introduced by Walker (2004). Ghosal & Tang (2006) used this approach to obtain posterior consistency for Markov processes. Zeevi & Meir (1997), Genovese & Wasserman (2000), Roeder & Wasserman (1997), and Li & Barron (1999) also obtained approximation and classical and Bayesian consistency results for mixture models. Posterior convergence rates for mixture models were obtained by Ghosal et al. (2000) and Kruijer et al. (2009) among others. Wu & Ghosal (2010) and Norets & Pelenis (2009) considered consistency in estimation of multivariate densities.

Muller et al. (1996), Roeder & Wasserman (1997), Norets & Pelenis (2009), Taddy & Kottas (2010) suggested obtaining conditional densities of interest from joint distribution estimation. MacEachern (1999), De Iorio et al. (2004), Griffin & Steel (2006), Dunson & Park (2008), and Chung & 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 & Xu (1995), Peng et al. (1996), Wood et al. (2002), Geweke & Keane (2007), Villani et al. (2009), and Norets (2010)).

Posterior consistency results for direct conditional density estimation are scarce. Norets (2010) shows that large nonparametric classes of conditional densities can be approximated in the Kullback-Leibler distance by three different specifications of finite mixtures of normal densities: (i) only means of the mixed normals depend flexibly on covariates, (ii) only mixing probabilities depend flexibly on covariates, and (iii) only mixing probabilities modeled by multinomial logit model depend on covariates. Schwartz (1965) theory suggest that these Kullback-Leibler approximation results imply posterior consistency in weak topology norm. Pati et al. (2010) 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 these two papers 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 by 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 kernel stick breaking processes introduced by Dunson & 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. (2010) 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 data generating process 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 prior 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 the 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 prior for locations of the mixing probability kernels is not restricted under both types of mixing.

The organization of the paper is as follows. Section 2 defines weak and strong posterior consistency for conditional densities and present 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 concludes.

2. The notion of posterior consistency for conditional densities. Consider a product space $Y \times X$, $Y \subset R$ and $X \subset R^{d_x}$. Let $\mathcal{F} = \{f : Y \times X \to [0, \infty), \int_Y f(y|x)dy = 1\}$ be the set of all conditional densities on Y with respect to Lebesgue measure. Let us denote the data

generating density of covariates x with respect to some generic measure ν by $f_0^x(x)$ and the data generating conditional density of interest by $f_0 \in \mathcal{F}$. The joint probability measure implied by f_0 and $f_0^x(x)$ is denoted by F_0.

To define a notion of posterior consistency we need to define neighborhoods on the space of conditional densities. Previous literature on Bayesian nonparametric density estimation employed weak and strong topologies on spaces of densities with respect to some common dominating measure. Quite general weak and strong posterior consistency theorems were established (Schwartz (1965), Barron (1988), Barron et al. (1999), Ghosal 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 $f_0^x(x)$. For example, a distance between conditional densities $f_1, f_2 \in \mathcal{F}$ that generates strong neighborhoods is defined by the total variation distance between the joint distributions,

$$\int |f_1 f_0^x - f_2 f_0^x| = \int |f_1(y|x) f_0^x(x) - f_2(y|x) f_0^x(x)| dy d\nu(x).$$

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

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

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

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

Similarly defined integrated total variation and Kullback-Leibler distances for conditional densities were considered in Ghosal & Tang (2006) and Norets & Pelenis (2009).

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

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

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

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

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

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

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

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

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

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

3. Kernel mixtures with variable number of components. Consider the following model for a conditional density,

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

where $\phi(y, \mu, \sigma)$ is a fixed symmetric density with location μ and scale σ evaluated at y and $K(.)$ 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 on $\Theta \times \{1,2,\dots,\infty\}$.

3.1. *Weak consistency.* We impose the following assumption on the data generating process.

ASSUMPTION 3.1. *1. $X = [0,1]^{d_x}$ (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

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

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, 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 nonparametrically specified data generating processes satisfy this assumption.

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

ASSUMPTION 3.2. *1. $\phi(y,\mu,\sigma) = \sigma^{-d}\psi((y-\mu)/\sigma)$, where $\psi(z)$ is a bounded, continuous, symmetric around zero, and monotone decreasing in $|z|$ probability density.*

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

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

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

ASSUMPTION 3.3. *The kernel $K(\cdot)$ is positive, bounded above, continuous, non-decreasing, and has a bounded derivative on $(-\infty,0]$. The upper bound can be set to 1 and, thus, $1 \ge K(z) > 0$ 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 exists 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 the Appendix. 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 is m and an open neighborhood Θ^m such that for any $\theta_{1:m} \in \Theta^m$,*

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

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

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

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

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

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

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

The upper bound in (3.3) is a constant and the logarithm of the lower bound is integrable by part 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 $f_0(y|x)f_0^x(x)$ have positive prior probability, where the prior on the density of x puts probability one on f_0^x and the prior for conditional densities is defined by Π 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 $\theta_{1:m}$, Θ^m, $\Pi(\tilde{\theta}_{1:m} \in \Theta^m, m) > 0$.*

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

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

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

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

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

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

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

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

A proof is provided in the Appendix. 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* (μ_j, σ_j, Q_j) *are 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, $\overline{Q}_n$, $\overline{\mu}_n$, $\overline{\sigma}_n \uparrow \infty$ *and* $\underline{\sigma}_n \downarrow 0$ *with* $\overline{\sigma}_n > \underline{\sigma}_n$ *such that*

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

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

where b_i *are defined in Proposition 3.1.*

3. *Conditions of Theorem 3.2 hold.*

Then the posterior is strongly consistent at f_0.

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

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

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

as desired. Let $m = \lfloor \tilde{m} \rfloor$ *and choose a Weibull prior with shape parameter* $k \ge 2$ *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 infinite and at least countable 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 kernel stick-breaking (KSB) mixture introduced by Dunson & Park (2008),

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

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

where θ, K, and ψ 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. *For any m, $\theta^{KM} \in \Theta$, and $\epsilon > 0$ there exists $\theta^{KSB} \in \Theta$ and n such that*

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

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

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

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

4.1. *Weak consistency.* To show that a KSB mixture is weakly consistent we will prove that the KL property holds.

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

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

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

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

By Proposition 4.1 there exists n and $\theta^{KSB}_{1:n}$ such that the left hand side in (4.2) is smaller than $\epsilon/4$. From the arguments in Corollary 3.1 it follows that the left hand side in (4.2) is

continuous in $\theta_{1:n}^{KSB}$. Therefore, there exists an open neighborhood of $\theta_{1:n}^{KSB}$, Θ^n, such that for any $\tilde{\theta}_{1:n}^{KSB} \in \Theta^n$

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

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

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

By the proposition assumption $\Pi(\tilde{\theta}_{1:n}^{KSB} \in \Theta^n) > 0$ 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 $f_0(y|x)$,*

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

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

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

The restriction on the mixing probabilities in the sieve definition is similar to the one used by Pati et al. (2010). We calculate a bound on the metric entropy of $\mathcal{F}_n$ in the following proposition.

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

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

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

A proof is given in the Appendix.

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* $\underline{\sigma}_n \downarrow 0$ *with* $\overline{\sigma}_n > \underline{\sigma}_n$ *such that*

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

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

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

3. *Conditions of Theorem 4.1 hold.*

Then the posterior is strongly consistent at f_0.

Theorem 3.3 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

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

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

LEMMA 4.1. *If prior distributions of* α_j *and* $K_j = K(-Q_j d_x)$ *first order stochastically dominate* $Beta(\gamma, 1)$ *for some* $\gamma > 2$, *then*

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

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

EXAMPLE 4.1. *Suppose priors for* μ *and* σ *and sequences* $\overline{\mu}_n$, $\underline{\sigma}_n$, *and* $\overline{\sigma}_n$ *are the same as in Example 3.1 (normal and inverse gamma priors). Then for* $m_n = cn/\log n$ *and* $\overline{Q}_n = n^r$, *where* c *and* r *are constants, condition* (4.5) *is satisfied for* c *sufficiently small.*

Condition (4.4) *is satisfied if the prior distributions for* $K(-Q_j d_x)$ *and* α_j *first order stochastically dominate* $Beta(\gamma, 1)$ *for some* $\gamma \geq 2$ *by Lemma 4.1 (note that for* $m_n = cn/\log n$, $\exp(-0.5 m_n \log m_n) \leq \exp(-0.25cn)$ *for large enough* n*).*

Explicit priors for Q_j *and* α_j *satisfying the sufficient conditions can be constructed for particular choices of* $K(\cdot)$*. For example, for* $K(\cdot) = \exp(\cdot)$, $\alpha_j \sim Beta(\gamma, 1)$ *and* $Q_j \sim Exponential(\gamma d_x)$, *which is equivalent to* $K_j = \exp(-Q_j d_x) \sim Beta(\gamma, 1)$*, satisfy conditions of Lemma 4.1. Also,* $\Pi(Q_j > n^r) \leq c_1 e^{-nc_2}$ *for* $r \geq 1$.

5. Discussion. The regularity conditions on the data generating process assumed in proving weak and strong posterior consistency are very mild. The conditions require the tails of the mixed location-scale density not to 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 estimation of unconditional densities. Examples of priors that satisfy the assumptions include normal prior 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 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. However, the techniques for deriving convergence rates are rather different from the ones used in this paper. Thus, we leave this problem for future research.

6. Appendix.

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 part of Y: for any $y \in Y$ there exists M_0 such that

$$\forall m \geq M_0, \;\; C_{\delta_m}(y) \cap A_0^m = \emptyset, \tag{6.1}$$

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

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

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

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

$$\delta_m \to 0, \;\; \sigma_m/\delta_m \to 0, \;\; h_m/\sigma_m \to 0, \tag{6.2}$$

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

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

$$\frac{K(-2Q^m s_m)}{K(-Q^m s_m) s_m^{d_x/2}} \to 0, \quad s_m = d_x \lambda(B_i^m)^{2/d_x} \to 0. \tag{6.3}$$

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

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

$$K(-Q^m ||x - q_i^m||^2) \ge K(-Q^m s_m). \tag{6.4}$$

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

$$K(-Q^m ||x - q_i^m||^2) \le K(-2Q^m s_m). \tag{6.5}$$

Note that

$$\begin{aligned} &\frac{\sum_{i \in I_1^m(x,s_m)} K(-Q^m ||x - q_i^m||^2)}{\sum_{i=1}^{m_x} K(-Q^m ||x - q_i^m||^2)} \\ &\ge 1 - \frac{\sum_{i \in I_2^m(x,s_m)} K(-Q^m ||x - q_i^m||^2)}{\sum_{i \in I_1^m(x,s_m)} K(-Q^m ||x - q_i^m||^2)} \\ &\ge 1 - \frac{\text{card}(I_2^m(x, s_m)) K(-2Q^m s_m)}{K(-Q^m s_m)} \ge 1 - d_x^{d_x/2} \frac{K(-2Q^m s_m)}{K(-Q^m s_m) s_m^{d_x/2}}, \end{aligned} \tag{6.6}$$

where the second inequality follows from (6.4) and (6.5). The last inequality follows from $\text{card}(I_2^m(x, s_m)) \le m_x = d_x^{d_x/2} s_m^{-d_x/2}$.

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

$$F(A_j^m | x_i^m) \ge \lambda(A_j^m) \inf_{z \in C_{\delta_m}(y),\ ||t-x||^2 \le 2s_m} f(z|t). \tag{6.7}$$

Inequalities (6.6), (6.7), and Lemma 6.1 imply that $p(y|x, \theta, m)$ exceeds

$$\begin{aligned} &\sum_{j: A_j^m \subset C_{\delta_m}(y)} \sum_{i \in I_1^m(x,s_m)} F(A_j^m | q_i^m) \frac{K(-Q^m ||x - q_i^m||^2)}{\sum_l K(-Q^m ||x - q_l^m||^2)} \phi(y, \mu_j^m, \sigma_m) \\ &\ge \inf_{z \in C_{\delta_m}(y),\ ||t-x||^2 \le 2s_m} f(z|t) \\ &\cdot \left[1 - \frac{6\psi(0) h_m}{\sigma_m} - 2 \int_{\delta_m/\sigma_m}^{\infty} \psi(\mu) d\mu\right] \cdot \left[1 - d_x^{d_x/2} \frac{K(-Q^m s_m)}{K(-Q^m s_m/2^2) s_m^{d_x/2}}\right]. \end{aligned} \tag{6.8}$$

By (6.2) and (6.3), given some $\epsilon_1 > 0$ there exists M_1 such that for $m \ge M_1$ the product in the last line of (6.8) 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 $m \geq \max\{M_1, M_2\}$

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

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

Let us derive an integrable upper bound for the DCT:

$$\begin{aligned} p(y|x,\theta,m) \geq & \left[1 - d_x^{d_x/2}\frac{K(-2Q^m s_m)}{K(-Q^m s_m)s_m^{d_x/2}}\right] \\ & \cdot \Bigg([1 - 1_{A_0^m}(y)] \cdot \inf_{||z-y||\leq r,\, ||t-x||\leq r} f_0(z|t) \cdot \sum_{j: A_j^m \subset C_r(y)\cap (A_0^m)^c} \lambda(A_j^m)\phi(y,\mu_j^m,\sigma_m) \\ & + 1_{A_0^m}(y) \cdot \inf_{||z-y||\leq r,\, ||t-x||\leq r} f_0(z|t) \cdot \lambda(C_r(y)\cap A_0^m)\phi(y,0,\sigma_0)\Bigg) \end{aligned} \tag{6.9}$$

For any m larger then some M_3, the Riemann sum in (6.9) is bounded below by $1/4$ (by Lemma 6.1) and

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

(by (6.3)).

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

$$\begin{aligned} & \log\max\{1, \frac{f_0(y|x)}{p(y|x,\theta,m)}\} \leq \log\max\{1, \frac{f_0(y|x)}{\inf_{||z-y||\leq r,\, ||t-x||\leq r} f_0(z|t)\cdot\phi(y,0,\sigma_0)\cdot(r/2)}\} \\ & = \log\frac{1}{\phi(y,0,\sigma_0)(r/2)}\max\{\phi(y,0,\sigma_0)(r/2), \frac{f_0(y|x)}{\inf_{||z-y||\leq r,\, ||t-x||\leq r} f_0(z|t)}\} \\ & \leq -\log(\phi(y,0,\sigma_0)(r/2)) + \log\frac{f_0(y|x)}{\inf_{||z-y||\leq r,\, ||t-x||\leq r} f_0(z|t)}. \end{aligned} \tag{6.10}$$

The first expression in (6.10) is integrable by Assumption 3.2 part 2. The second expression in (6.10) is integrable by Assumption 3.1 part 3. Thus the proposition is proved. □

Proof. Proposition 3.1.

The proof generalizes the ideas from Ghosal et al. (1999), Theorem 6 and Tokdar (2006) Lemma 4.1 to general location scale densities and covariate dependent mixing weights.

Suppose $f_1, f_2 \in \mathcal{F}_n$. We would like to find the restrictions on the parameters $\theta_m^i = \{Q_j^i, \mu_j^i, \sigma_j^i, q_j^i, \alpha_j^i\}_{j=1}^m$ for $i = 1, 2$ such that $\int |f_1(y|x) - f_2(y|x)| dy f_0^x(x) dx < \delta$. For notational simplicity let

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

Then for any given $x \in X$

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

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

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

Note that

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

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

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

Combining the two pieces together we find that

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

Next step is to find an upper bound for $\sum_{j=1}^m |\pi_j^1(x) - \pi_j^2(x)|$. We introduce additional notation, where $\widetilde{\alpha}^i$ is a vector of normalized weights α^i, i.e. $\widetilde{\alpha}_j^i = \alpha_j^i / \sum_{l=1}^m \alpha_l^i$, $K_j^i(x) = K(-Q_j^i||x - q_j^i||^2)$ and $A_i(x) = \sum_{j=1}^m \widetilde{\alpha}_j^i K_j^i(x)$. Then for any $x \in X$

$$\begin{aligned}
&\sum_{j=1}^m |\pi_j^1(x) - \pi_j^2(x)| = \sum_{j=1}^m \left| \frac{\widetilde{\alpha}_j^1 K_j^1(x)}{\sum_{i=1}^m \widetilde{\alpha}_i^1 K_i^1(x)} - \frac{\widetilde{\alpha}_j^2 K_j^2(x)}{\sum_{i=1}^m \widetilde{\alpha}_i^2 K_i^2(x)} \right| = \sum_{j=1}^m \left| \frac{\widetilde{\alpha}_j^1 K_j^1(x)}{A_1(x)} - \frac{\widetilde{\alpha}_j^2 K_j^2(x)}{A_2(x)} \right| \\
&= \frac{1}{A_1(x)A_2(x)} \sum_{j=1}^m \left| \widetilde{\alpha}_j^1 K_j^1(x) A_2(x) - \widetilde{\alpha}_j^2 K_j^2(x) A_1(x) + \widetilde{\alpha}_j^2 K_j^2(x) A_2(x) - \widetilde{\alpha}_j^2 K_j^2(x) A_2(x) \right| \\
&\leq \frac{\sum_{j=1}^m |\widetilde{\alpha}_j^1 K_j^1(x) - \widetilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)} + \frac{\sum_{j=1}^m \widetilde{\alpha}_j^2 K_j^2(x) |A_2(x) - A_1(x)|}{A_1(x)A_2(x)} \\
&= \frac{\sum_{j=1}^m |\widetilde{\alpha}_j^1 K_j^1(x) - \widetilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)} + \frac{|A_2(x) - A_1(x)|}{A_1(x)} \\
&= \frac{\sum_{j=1}^m |\widetilde{\alpha}_j^1 K_j^1(x) - \widetilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)} + \frac{|\sum_{j=1}^m \widetilde{\alpha}_j^1 K_j^1(x) - \widetilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)} \\
&\leq 2\frac{\sum_{j=1}^m |\widetilde{\alpha}_j^1 K_j^1(x) - \widetilde{\alpha}_j^2 K_j^2(x)|}{A_1(x)} = 2\frac{\sum_{j=1}^m |\widetilde{\alpha}_j^1 K_j^1(x) - \widetilde{\alpha}_j^2 K_j^2(x) + \widetilde{\alpha}_j^1 K_j^2(x) - \widetilde{\alpha}_j^1 K_j^2(x)|}{\sum_{j=1}^m \widetilde{\alpha}_j^1 K_j^1(x)} \\
&\leq 2 \left[\frac{\sum_{j=1}^m \widetilde{\alpha}_j^1 |K_j^1(x) - K_j^2(x)}{\sum_{j=1}^m \widetilde{\alpha}_j^1 K_j^1(x)} + \frac{\sum_{j=1}^m |\widetilde{\alpha}_j^1 - \widetilde{\alpha}_j^2| K_j^2(x)}{\sum_{j=1}^m \widetilde{\alpha}_j^1 K_j^1(x)} \right] \\
&\leq 2\frac{1}{K(-\overline{Q}_n d_x)} \left[\max_{j=1,\ldots,m} |K_j^1(x) - K_j^2(x)| + \sum_{j=1}^m |\widetilde{\alpha}_j^1 - \widetilde{\alpha}_j^2| \right].
\end{aligned}$$

Given any $\delta > 0$ and any $f^* \in \mathcal{F}_n$ we want to ensure that there exists an i such that f^*, f_i satisfy

$$\begin{aligned}
&\int |f^*(y|x) - f_i(y|x)|dy \\
&\leq \sum_{j=1}^m \pi_j^1(x) \left(2\psi(0) \frac{|\mu_j^* - \mu_j^i|}{\sigma_j^*} + 2\frac{\sigma_j^* - \sigma_j^i}{\sigma_j^i} \right) + 2\frac{1}{K(-\overline{Q}_n d_x)} \left[\max_{j=1,\ldots,m} |K_j^*(x) - K_j^i(x)| + \sum_{j=1}^m |\widetilde{\alpha}_j^* - \widetilde{\alpha}_j^i| \right] \\
&\leq \frac{\delta}{3} + 2\frac{1}{K(-\overline{Q}_n d_x)} \left[\frac{\delta K(-\overline{Q}_n d_x)}{6} + \frac{\delta K(-\overline{Q}_n d_x)}{6} \right] = \delta.
\end{aligned}$$

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

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

Then if $(\mu^1,\sigma^1),(\mu^2,\sigma^2) \in E_{ij}$, then $\left(2\psi(0)\frac{|\mu^1-\mu^2|}{\sigma^1}+2\frac{\sigma^1-\sigma^2}{\sigma^2}\right) \leq \frac{\delta}{3}$ as desired. Take $N = \sum_{j=1}^{H} N_j$, then

$$\begin{aligned} N &\leq \sum_{j=1}^{H}\left(\frac{24\psi(0)}{\delta}\frac{\overline{\mu}_n}{\sigma_j}+1\right) = \frac{24\psi(0)}{\delta}\frac{\overline{\mu}_n}{\underline{\sigma}_n}\sum_{j=1}^{H}(1+\zeta)^{-j}+H \\ &\leq \frac{24\psi(0)}{\delta}\frac{\overline{\mu}_n}{\underline{\sigma}_n}\frac{1}{\zeta}+\frac{1}{\log(1+\zeta)}\log(\frac{\overline{\sigma}_n}{\underline{\sigma}_n})+1 \\ &= c_0\frac{\overline{\mu}_n}{\underline{\sigma}_n}+c_1\log\frac{\overline{\sigma}_n}{\underline{\sigma}_n}+1 \end{aligned}$$

where c_0, c_1 depend on δ, but not on $\overline{\mu}_n, \overline{\sigma}_n, \underline{\sigma}_n$. Hence the logarithm of the number of grid points to bound $\sum_{j=1}^{m}\pi_j^1(x)\left(2\psi(0)\frac{|\mu_j^*-\mu_j^i|}{\sigma_j^*}+2\frac{\sigma_j^*-\sigma_j^i}{\sigma_j^i}\right) < \frac{\delta}{3}$ is given by $m_n\log\left(c_0\frac{\overline{\mu}_n}{\underline{\sigma}_n}+c_1\log\frac{\overline{\sigma}_n}{\underline{\sigma}_n}+1\right)$.

As shown by Ghosal et al. (1999), Lemma 1, the logarithm of the number N of vectors $(\widetilde{\alpha}^1,\ldots,\widetilde{\alpha}^N)$ needed to make $\sum_{j=1}^{m}|\widetilde{\alpha}_j^*-\widetilde{\alpha}_j^i| < \frac{\delta K(-\overline{Q}_n d_x)}{6}$ for some $i \in \{1,\ldots,N\}$ is bounded above by $m_n\left(1+\log\frac{1+\delta K(-\overline{Q}_n d_x)/6}{\delta K(-\overline{Q}_n d_x)/6}\right)$. This bound can be expressed as $m_n(c_2+c_3\log(K(-\overline{Q}_n d_x)))$ where c_2, c_3 depend on δ, but not $\overline{Q}_n$.

Finally, we need to construct a bound on the logarithm of the number of grid points for $\left\{Q_j^i, q_j^i\right\}_{j=1}^{m}$ so that there exists an i such that $|K(-Q_j^i||x-q_j^i||^2)-K(-Q_j^*||x-q_j^*||^2)| < \frac{\delta K(-\overline{Q}_n d_x)}{6}$. By Assumption 3.3, K' is bounded above, $K' < \overline{K}'$, then

$$\begin{aligned} &|K(-Q_j^i||x-q_j^i||^2)-K(-Q_j^*||x-q_j^*||^2)| \leq \overline{K}'(||x-q_j^i||^2)|Q_j^i-Q_j^*|+\overline{K}'\overline{Q}_n\sum_{l=1}^{d_x}2(|x_l-q_{j,l}^i|) \\ &\leq 2\overline{K}' d_x\overline{Q}_n\max_{l=1,\ldots,d_x}|x_l-q_{j,l}^i|+\overline{K}' d_x|Q_j^*-Q_j^i| \leq \frac{\delta K(-\overline{Q}_n d_x)}{12}+\frac{\delta K(-\overline{Q}_n d_x)}{12} = \frac{\delta K(-\overline{Q}_n d_x)}{6}. \end{aligned}$$

Hence the number of grid points for $\{Q_j\}_{j=1}^{m_n}$ is determined by ensuring that there exists an i and Q_j^i such that $|Q_j^*-Q_j^i| \leq \frac{\delta K(-\overline{Q}_n d_x)}{12\overline{K}' d_x}$. Since $Q_j^i \in (0,\overline{Q}_n)$, therefore the logarithm of the number of grid points is bounded above by $m_n(\log(\overline{Q}_n)-\log(K(-\overline{Q}_n d_x))+\log(12\overline{K}' d_x/\delta))$. Similarly, we want to ensure that there exists an i such that $2\overline{K}' d_x\overline{Q}_n\max_{l=1,\ldots,d_x}|x_l-q_{j,l}^i| < \delta K(-\overline{Q}_n d_x)/12$. Since q_j belongs to the unit cube $[0,1]^{d_x}$ the number of grid points for q_j is bounded above by $\left(\frac{24\overline{K}' d_x\overline{Q}_n}{\delta K(-\overline{Q}_n d_x)}\right)^{d_x}$. Then the bound on logarithm of the number grid points for $\{q_j\}_{j=1}^{m_n}$ is $m_n d_x\left(\log(24K' d_x/\delta)+\log(\overline{Q}_n)-\log(K(-\overline{Q}_n d_x))\right)$. The joint bound on possible grid points for Q and q is given by $m_n(c_4+c_5\log(\overline{Q}_n)+c_6\log(K(-\overline{Q}_n d_x))$ where c_4, c_5, c_6 depend on δ, but not on $\overline{Q}_n$.

Combining all the pieces together we get that

$$
\begin{aligned}
J(\delta, \mathcal{F}_n) &\leq m_n \log\left(c_0 \frac{\overline{\mu}_n}{\underline{\sigma}_n} + c_1 \log \frac{\overline{\sigma}_n}{\underline{\sigma}_n} + 1\right) \\
&\quad + m_n(c_2 + c_3 \log(K(-\overline{Q}_n d_x))) \\
&\quad + m_n(c_4 + c_5 \log \overline{Q}_n + c_6 \log K(-\overline{Q}_n d_x)) \\
&\leq m_n \left(\log\left[b_0 \frac{\overline{\mu}_n}{\underline{\sigma}_n} + b_1 \log \frac{\overline{\sigma}_n}{\underline{\sigma}_n} + 1\right] + b_2 + b_3 \log(\overline{Q}_n) + b_4 \log(K(-\overline{Q}_n d_x))\right)
\end{aligned}
$$

where b_0, b_1, b_2, b_3, b_4 do not depend on the parameter θ values. □

Proof. Proposition 4.1.

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 KSB mixture be

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

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

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

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

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

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

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

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

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

□

Proof. Proposition 4.3.

For $f_1, f_2 \in \mathcal{F}_n$ the following is true

$$\begin{aligned}
||f_1-f_2||_1 &\le \int_X\int_Y\sum_{j=1}^{\infty}\left|\pi_j^1(x)\phi(y;\mu_j^1,\sigma_j^1)-\pi_j^2(x)\phi(y;\mu_j^2,\sigma_j^2)\right|dyf_0^x(x)dx\\
&\le \int_X\int_Y\sum_{j=1}^{m_n}\pi_j^1(x)\left|\phi(y;\mu_j^1,\sigma_j^1)-\phi(y;\mu_j^2,\sigma_j^2)\right|dyf_0^x(x)dx\\
&+\int_X\int_Y\sum_{j=1}^{m_n}\left|\pi_j^1(x)-\pi_j^2(x)\right|\phi(y;\mu_j^2,\sigma_j^2)dyf_0^x(x)dx\\
&+\int_X\sum_{j=m_n+1}^{\infty}|\pi_j^1(x)-\pi_j^2(x)|f_0^x(x)dx\\
&\le \int_X\sum_{j=1}^{m_n}\pi_j^1(x)\int_Y\left|\phi(y;\mu_j^1,\sigma_j^1)-\phi(y;\mu_j^2,\sigma_j^2)\right|dyf_0^x(x)dx\\
&+\sum_{j=1}^{m_n}||\pi_j^1-\pi_j^2||_1+\sup_{x\in X}\sum_{j=m_n+1}^{\infty}|\pi_j^1(x)|+|\pi_j^2(x)|\\
&\le \int_X\sum_{j=1}^{m_n}\pi_j^1(x)\int_Y\left|\phi(y;\mu_j^1,\sigma_j^1)-\phi(y;\mu_j^2,\sigma_j^2)\right|dyf_0^x(x)dx\\
&+\sum_{j=1}^{m_n}||\pi_j^1-\pi_j^2||_1+2\delta
\end{aligned}$$

where last inequality is true by construction of $\mathcal{F}_n$ as $\sup_{x\in X}\sum_{j=m_n+1}^{\infty}|\pi_j^1(x)\le\delta$.

Then, given any $\delta > 0$ and any $f^* \in \mathcal{F}_n$ we want to define a grid in a such way that there would exist an i and $f_i \in \mathcal{F}_n$ such that f^* and f_i satisfy $||f*-f_i||_1 < 4\delta$. For the first part $\int_X\sum_{j=1}^{m_n}\pi_j^i(x)\int_Y\left|\phi(y;\mu_j^i,\sigma_j^i)-\phi(y;\mu_j^*,\sigma_j^*)\right|dyf_0^x(x)dx$ the logarithm of the number of grid points on μ,σ is bounded by $m_n\log\left[b_0\frac{\overline{\mu}_n}{\underline{\sigma}_n}+b_1\log\frac{\overline{\sigma}_n}{\underline{\sigma}_n}+1\right]$ as shown in Proposition 3.1.

Use the notation that $\pi_j(x) = \alpha_j K_j(x) \prod_{l<j}(1-\alpha_l K_l(x))$. Assume that $|\alpha_j^1 K_j^1(x) - \alpha_j^2 K_j^2(x)| < \delta/m_n^2$, then

$$
\begin{aligned}
&|\pi_j^1(x) - \pi_j^2(x)| = |\alpha_j^1 K_j^1(x) \prod_{i<j}(1-\alpha_i^1 K_i^1(x)) - \alpha_j^2 K_j^2(x) \prod_{i<j}(1-\alpha_i^2 K_i^2(x))| \\
&\leq |\alpha_j^1 K_j^1(x) - \alpha_j^2 K_j^2(x)| \prod_{i<j}(1-\alpha_i^1 K_i^1(x)) + \alpha_j^2 K_j^2(x)| \prod_{i<j}(1-\alpha_i^1 K_i^1(x)) - \prod_{i<j}(1-\alpha_i^2 K_i^2(x))| \\
&\leq |\alpha_j^1 K_j^1(x) - \alpha_j^2 K_j^2(x)| + |\prod_{i<j}(1-\alpha_i^1 K_i^1(x)) - \prod_{i<j}(1-\alpha_i^2 K_i^2(x))| \\
&\leq \sum_{i\leq j} |\alpha_i^1 K_i^1(x) - \alpha_i^2 K_i^2(x)| = j\frac{\delta}{m_n^2} \leq \frac{\delta}{m_n}.
\end{aligned}
$$

We need to find a bound on a logarithm of grid points on α, Q, q so that $\sum_{j=1}^{m_n} ||\pi_j^i - \pi_j^*||_1 < \delta$. From the inequality above $\sum_{j=1}^{m_n} ||\pi_j^i - \pi_j^*||_1 < \delta$ if $|\alpha_j^i K_j^i(x) - \alpha_j^* K_j^*(x)| < \delta/m_n^2$. Note that $|\alpha_j^1 K_j^1(x) - \alpha_j^* K_j^*(x)| < |\alpha_j^1 - \alpha_j^*| + |K_j^1(x) - K_j^*(x)|$, therefore we consider bounding $|\alpha_j^1 - \alpha_j^*| < \frac{\delta}{2m_n^2}$ and $|K_j^1(x) - K_j^*(x)| < \frac{\delta}{2m_n^2}$. Hence, the number of grid points for $\{\alpha_j\}_{j=1}^{m_n}$ is determined by ensuring that there exists an i and α_j^i such that $|\alpha_j^i - \alpha_j^*| < \frac{\delta}{2m_n^2}$. As $\alpha_j \in (0,1)$ therefore the logarithm on the number of grid points is bounded above by $m_n \log(2m_n^2/\delta)$. Finally, we need to construct a bound on the logarithm of the number of grid points for $\left\{Q_j^i, q_j^i\right\}_{j=1}^{m}$ so that there exists an i such that $|K(-Q_j^i||x - q_j^i||^2) - K(-Q_j^*||x - q_j^*||^2)| < \frac{\delta}{2m_n^2}$. Following the proof of Proposition 3.1 the logarithm of the grid points for $\{Q_j\}_{j=1}^{m_n}$ is bounded above by $m_n(\log(\overline{Q}_n) + 2\log(m_n)) + \log(4\overline{K}' d_x/\delta))$ and the logarithm of grid points for $\{q_j\}_{j=1}^{m_n}$ is bounded above by $m_n d_x \left(\log(8K' d_x/\delta) + \log(\overline{Q}_n) + 2\log(m_n)\right)$. The joint bound on possible grid points for Q and q is given by $m_n(c_4 + c_5 \log(\overline{Q}_n) + c_6 \log(m_n))$ where c_4, c_5, c_6 depend on δ, but not on $\overline{Q}_n$ or m_n.

Combining all the pieces together we find that

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

□

Proof. Lemma 4.1.

First, we will prove a secondary result that will be used later. Suppose that $a, b \overset{i.i.d.}{\sim} Beta(\gamma, 1)$ for $\gamma > 2$, then $a \cdot b$ first order stochastically dominate $Beta(1, \alpha)$ distribution for $\gamma \geq \alpha \geq 2$. To prove this we need to show that $Pr(a \cdot b \leq z) \leq 1 - (1-z)^\alpha$. Since $a, b \overset{i.i.d.}{\sim} Beta(\gamma, 1)$, therefore

$-\log a, -\log b \overset{i.i.d.}{\sim} Exponential(\gamma)$ and $-\log a - \log b \sim Gamma(2, 1/\gamma)$.

$$\begin{aligned} Pr(a \cdot b \leq z) &= 1 - Pr(-\log a - \log b \leq -\log z) \\ &= 1 - \int_0^{-\gamma \log z} te^{-t} dt = z^\gamma (1 - \gamma \log z). \end{aligned}$$

Then the desired result follows from

$$A(z) = (1-z)^\alpha + z^\gamma (1 - \gamma \log z) \leq 1 \tag{6.11}$$

for all $z \in [0,1]$ and $\gamma \geq \alpha \geq 2$. To check the inequality (6.11) first verify $A(0+) = A(1) = 1$. Second, $A'(z) = 0$ gives $\log z = \alpha(1-z)^{\alpha-1}/(-\gamma^2 z^{\gamma-1})$ and after plugging in this value for $\log z$,

$$A(z) \leq \max\{1, z^\gamma + z(1-z)^{\alpha-1})\alpha/\gamma + (1-z)^\alpha\} \leq \max\{1, 1 + z^2 - z\} \leq 1.$$

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

$$\begin{aligned} Pr(c < x) &= \frac{\int_0^x \alpha^m t^{m-1} e^{-\alpha t} dt}{(m-1)!} = \frac{\int_0^{\alpha x} t^{m-1} e^{-t} dt}{(m-1)!} < (\alpha x)^m / m! \\ &= \frac{(\alpha x)^m}{\exp\{m \log m - m + O(\log(m))\}} \qquad \text{(by Sterling formula)} \\ &= \exp\{-m \log m + m + m \log(\alpha x) - O(\log(m))\} \\ &= \exp(-0.5m \log m) \frac{\exp(m \log(\alpha x) + m + O(\log(m)))}{\exp(0.5m \log m)} < \exp(-0.5m \log m) \end{aligned}$$

when m is sufficiently large.

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

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

□

LEMMA 6.1. *Let $A_1, \ldots, A_m$ be a partition of an interval on R such that $\lambda(A_j) \leq h$ and $\mu_j \in A_j$. Assume $C_\delta(y) = [y-\delta, y+\delta] \subset \cup A_j$ is an interval with center y and length δ. Then*

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

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

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

$$\sum_{j \in J} \lambda(A_j \cap C_\delta(y)) \phi(y, \mu_j, \sigma) \geq \int_{\cup_{j \in J}[A_j \cap C_\delta(y)]} \phi(y, \mu - h, \sigma) d\mu. \tag{6.12}$$

Note next that

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

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

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

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

References.

Barron, A. (1988). The exponential convergence of posterior probabilities with implications for bayes estimators of density functions.

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

Billingsley, P. (1999). *Convergence of Probability Measures (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

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

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

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

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

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

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

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

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

Genovese, C. R., & Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.*, *28*, 1105–1127.

Geweke, J., & Keane, M. (2007). Smoothly mixing regressions. *Journal of Econometrics*, *138*.

Ghosal, S., Ghosh, J. K., & Ramamoorthi, R. V. (1999). Posterior consistency of dirichlet mixtures in density estimation. *The Annals of Statistics*, *27*, pp. 143–158.

Ghosal, S., Ghosh, J. K., & Vaart, A. W. v. d. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, *28*, pp. 500–531.

Ghosal, S., & Tang, Y. (2006). Bayesian consistency for markov processes. *Sankhya*, *68*, 227–239.

Ghosh, J., & Ramamoorthi, R. (2003). *Bayesian Nonparametrics*. Springer; 1 edition.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Griffin, J. E., & Steel, M. F. J. (2006). Order-based dependent dirichlet processes. *Journal of the American Statistical Association*, *101*, 179–194.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput.*, *3*, 79–87.

Jordan, M., & Xu, L. (1995). Convergence results for the em approach to mixtures of experts architectures. *Neural Networks*, *8*, 1409 – 1431.

Kruijer, W., Rousseau, J., & Van Der Vaart, A. (2009). Adaptive bayesian density estimation with location-scale mixtures, .

Li, J. Q., & Barron, A. R. (1999). Mixture density estimation. In *In Advances in Neural Information Processing Systems 12* (pp. 279–285). MIT Press.

MacEachern, S. N. (1999). Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*, .

McLachlan, G., & Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, Inc.

Muller, P., Erkanli, A., & West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, *83*, pp. 67–79.

Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *Annals of statistics*, *38*, 1733–1766.

Norets, A., & Pelenis, J. (2009). Bayesian modeling of joint and conditional distributions.

Pati, D., Dunson, D., & Tokdar, S. (2010). Posterior consistency in conditional distribution estimation.

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

Roeder, K., & Wasserman, L. (1997). Practical bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.*, *92*, 894–902.

Schwartz, L. (1965). On bayes procedures. *Z. Wahrsch. Verw. Gebiete*, (pp. 10–26).

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

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

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

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

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

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

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

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

Wu, Y., & Ghosal, S. (2010). The l1-consistency of dirichlet mixtures in multivariate bayesian density estimation. *Journal of Multivariate Analysis*, *101*, 2411 – 2419.

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

Andriy Norets
Fisher Hall, Economics Department,
Princeton University,
Princeton, NJ 08544, USA
E-mail: anorets@princeton.edu

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

Authors: Andriy Norets, Justinas Pelenis

Title: Posterior Consistency in Conditional Density Estimation by Covariate Dependent Mixtures

Reihe Ökonomie / Economics Series 282

Editor: Robert M. Kunst (Econometrics)
Associate Editors: Walter Fisher (Macroeconomics), Klaus Ritzberger (Microeconomics)

ISSN: 1605-7996

Stumpergasse 56, A-1060 Vienna • ☎ +43 1 59991-0 • Fax +43 1 59991-555 • http://www.ihs.ac.at

ISSN: 1605-7996